The Capitalization Effect of Designating Scenic Rivers, the Indemnification Effects of Successive Droughts, and Using Machine Learning to Price Specialty Crop Insurance

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Jonathon Siegle

Graduate Program in Agricultural, Environmental & Developmental Economics

The Ohio State University

2023

Dissertation Committee

H. A. Klaiber

Z. Plakias

M. J. Miranda

B. Sohngen

Abstract

Ohio's Scenic Rivers Program (OSRP) designates high-quality streams to protect pristine riparian corridors. My first essay uses a large data set of home sales across Ohio to identify the effect of designation on local housing values using temporal and spatial regressions. I find an average capitalization effect of approximately 6-8% for homes within 1km of designated streams that is robust to numerous sample definitions. This finding provides a unique ex-ante – ex-post hedonic analysis of scenic river designation to assist local governments' decisions on future program expansions. More fundamentally, this essay demonstrates how the value of environmental amenities may be dependent on a promise of future protection.

The United States' Federal Crop Insurance Program (FCIP) is a critical support structure for the country's farm financial security, but its size and centrally controlled prices allow for large efficiency losses through moral hazard and adverse selection. My second essay investigates successive droughts as one potential source of these issues. I hypothesize that successive droughts represent a right-tail risk that is predictable enough to incentivize adverse selection into the FCIP. My results show that successive droughts substantially and significantly lower irrigation's ability to mitigate losses from drought. This relationship carries over from lost acreage to loss ratios, a signal that the FCIP pricing cycles allow for this temporal adverse selection.

The Federal Crop Insurance Program (FCIP) has made significant progress in achieving universal coverage; however, expansion to specialty crop products remains a challenge. My third essay demonstrates a Machine Learning process with which I use historical data on weather and crop losses from ‘training’ counties to predict indemnification from a specialty crop in counties out of sample. I show that this method produces an average loss ratio comparable to USDA RMA performance without any geographic weighting, making it an appealing and low-cost baseline for considering a crop’s risk in a new area. In addition to assisting in pricing new crop insurance products, this algorithm can generally model catastrophic crop risk where data on the crop in question is sparse, a key addition to future models of food system resiliency.

Vita

2013-2017 .... B.S. Environmental Economics and Policy,

Michigan State University

2019-2021 .... Teaching Assistant,

The Ohio State University

2019 to present .... Research Assistant,

The Ohio State University

Fields of Study

Major Field: Agricultural, Environmental & Developmental Economics

Table of Contents

List of Tables

List of Figures

Chapter 1. The Capitalization Effect of Scenic River Designation

1. Introduction

Healthy riparian corridors provide numerous economic and environmental benefits, including riparian amenities, recreation opportunities, and protection for downstream environmental quality (Bowker and Bergstrom 2017). However, local development poses risks to stream quality (Hupp et al. 2013). 'Urban stream syndrome,' characterized by erosive flow events and reduced biodiversity, is a threat to riparian health and its associated amenities. This degradation may even undermine the value of amenities that initially drove nearby development (Walsh et al. 2005). Addressing the negative environmental externalities associated with development has become a billion-dollar enterprise (Bernhardt and Palmer 2007), often with limited restoration capacity (Smucker and Detenbeck 2014). Ohio's Scenic River Program (OSRP) was established in 1968 to "protect the aquatic resources and terrestrial communities dependent on healthy riparian habitats."[1] The OSRP can be viewed in the context of Urban Stream Syndrome as a preventative regulation for Ohio's highest quality river segments as well as an eco-tourism signal.

The OSRP is distinct from the National Scenic and Wild Rivers program, although both programs have similar goals of protecting high value riparian corridors. The OSRP designates rivers that meet environmental benchmarks, are nominated by the

[1] https://ohiodnr.gov/wps/portal/gov/odnr-core/divisions/division-d-dnap/related-resource/scenic-rivers

Ohio Department of Natural Resources, and receive consenting resolutions from a majority of local governments. A designated river segment is designated as either Recreational, Scenic, or Wild, with increasing environmental and regulatory standards. Once a stream segment is designated, public development within 1000 ft of the stream, or other modifications of the water channel, require approval from the acting director of Ohio's Department of Natural Resources (DNR). By increasing the oversight on these projects, the DNR hopes to guide development into designs with a negligible impact on the river's natural flow. While in conversations with the DNR they have stated that they do not aim to completely prohibit local development, designation could still delay, deter, or distort projects. Given that there are potentially both protection values and regulatory costs from designation for the same geographic area, we do not know a-priori the capitalization effect of designation to nearby homes. Despite the longstanding history of the OSRP, as well as similar systems such as the federal program, the literature contains little revealed preference analysis of the impacts of these programs on demand for local properties (Keith et al. 2008).

Hedonic property models, which trace back to the seminal work by Rosen (1974), are a revealed preference method that estimates the net willingness to pay for an amenity by parameterizing a function estimating equilibrium housing prices. In the cases of long timeframes and/or changes in underlying amenities, changes in housing prices are best described as 'capitalization effects' as changes in the underlying hedonic price function no longer capture welfare changes in this setting (Kuminoff and Pope 2014). These capitalization effects may be of particular importance to local governments, due to local voter pressure and/or a desire to protect property tax revenue; because their resolutions of support control the final decision if a river is designated, these concerns of local governments are of particular importance to the future of the program. Therefore, this hedonic property model investigates a measure of the OSRP's value that is understudied and bears considerable social importance.

This essay adds to a significant body of work on how riverine protection imparts value to users and residents. Using travel cost models based on interviews and mailed surveys of users of the Chatooga River, designated under the federal scenic rivers program, Moore and Siderelis (2003) estimate the economic impact to recreational visitors due to the designation at $2.608 million; they also find that over 80% of respondents said the river's designation was "Very Important" to them. The same authors conducted similar surveys for the west branch of the Farmington River, another segment designated under the federal program (Moore and Siderelis 2002). They estimated demand models for trips to the river under current conditions as well as under a hypothetical impairment to river quality and used a parsimonious regression of residential land value against distance from the river. They found drastic decreases to consumer surplus under a hypothetical decrease in river quality and found that distance was significantly correlated with lower per-acre land values. Using a travel cost and contingent valuation method to estimate the potential recreation benefits of designating the Mad River[2] as part of the OSRP, Zhang and Wishart (2019) estimated an average willingness to pay per year for designation of $47.98 and a travel cost surplus estimate of $178 per person per trip.

[2] At the time of writing, the Mad River remains undesignated

In contrast to the literature on use values, the existing hedonic literature on river designations is thin and shows mixed results. White and Leefers (2007) found no significant effect on housing values from proximity to rivers designated under the federal program. Other papers find significant capitalization effects for scenic rivers (Moore and Siderelis 2002); however, as Keith et al. (2008) point out, most of these hedonic studies of properties near designated rivers suffer from a lack of carefully constructed control groups or quasi-experimental settings. This challenge in forming an appropriate control group is particularly daunting as the treatment groups of designated rivers are defined by both preexisting environmental conditions and local communities self-selecting into the program. Constructing an appropriate control group and controlling for potentially omitted variables that impact scenic river designation are critical components to avoiding biased estimates of the 'treatment effect', as demonstrated recently by Towe et al. (2021) with a repeat sales analysis of stream restoration projects.

Using the OSRP for my empirical analysis, I add to the literature on the value of designating environmental amenities using a large and varied dataset of sales near multiple rivers with similar environmental qualities and under the same regulatory framework. Using detailed spatial data, I identify capitalization effects using comparisons against sales prior to designation as well as sales just outside of the designations' boundaries. Previewing my results, I find that designation had a significant and positive capitalization effect of between 6 and 8% on nearby homes with no systematic evidence of regulatory takings.

2. Ohio's Scenic River Program

In 1968, The Ohio Wild, Scenic and Recreational River Act established Ohio's scenic river program, in which qualifying stream segments may be designated as Wild, Scenic, or Recreational to earn additional state protections. The stated goal of the program is to "protect the aquatic resources and terrestrial communities dependent on healthy riparian habitats."[3] Designation[4] includes the river itself as well as adjacent lands proposed by the director located no further than 1000 feet from the stream. Once a stream segment has been designated public constructions such as highways, structures within designated land, or other modifications of a designated water channel require approval from the acting director. However, this special approval is only required for projects originating from state departments, state agencies, or political subdivisions: private construction is not regulated by designation. This regulation may curtail public infrastructure investment but allows homeowners to modify and expand development on private lots so long as new public infrastructure is not required to support this development.

To be recommended for designation, a stream segment must pass various environmental quality criteria for its riparian corridor, though historical cultural significance is also considered in designation studies. The riparian corridor requirements for each type of designation are detailed in Table 1, and include metrics of habitat, greenery, and a lack of local urban development. In general, these criteria set maximums for how developed the surrounding area of a prospective scenic river may be; the number of bridge crossings, parallel roads, commercial developments, and other components of

[3] https://ohiodnr.gov/discover-and-learn/land-water/rivers-streams-wetlands/scenic-rivers-program

[4] Full text of the administrative codes resulting from Ohio's Scenic River laws can be found in Ohio Revised Codes 1547.81 to 1547.85 and Ohio's Administrative Code 1501:47-4-01 to 47-4-04. The process is summarized here.

development are considered simultaneously with habitat standards and a minimum requirement of native forest or wetland surrounding the stream. The section to be designated must have a length of at least 15 continuous river miles, and Scenic or Recreational type designations require at least 20 continuous river miles. The DNR and designation studies also consider the historical and cultural importance of the river.

Finally, a river designation requires resolutions of support from at least 50% of local political subdivisions within 1000 ft of the area to be designated. From a discussion with a representative of Ohio's DNR, the regulation can cause contention with local governments and landowners. The prospective designation of segments of the Mad River, for example, was stopped due to a lack of support from local governments.

At present, 15 river systems have designated segments. Separating these by designation type and the date of designation results in 25 distinct designations. For this essay, I use distinct designations to distinguish sub-samples as well as fixed-effect controls. In Figure 1, I present a graphic summarizing these distinct designations locations and type[5]. I also include in Figure 1 river segments that have been proposed for designation but are either not designated at the time of writing or otherwise designated after the time of my data.

Of the 25 distinct designations, 14 distinct designations were completed by 1980. Within the time frame of my cleaned dataset, 5 distinct designations were completed, including designations along the Chagrin, Conneaut, Mohican, and Ashtabula rivers. These do not include any Recreational type designations.

3. Data

[5] The graphic in Figures 1 and 2 were generated using ArcMap GIS software

The hedonic dataset I use is based on a set of residential sales in Ohio, including home and lot characteristics, purchased from the private data vendor CoreLogic. The dataset includes over 900,000 sales of single-family homes from 1990 to 2016 across Ohio; however, the preponderance of the data is from sales from the years 2000 or later, and my analysis does not include sales prior to 1999 due to the dataset's gaps in coverage for those years. I normalize sale prices to 2000 dollars using the S&P/Case-Shiller OH-Cleveland Home Price Index prior to estimation.

I supplement the sales data with geographic information including shapefiles of Scenic Rivers and polygons describing 100-year flood hazard areas, both from the Ohio DNR, as well as polylines of Ohio's major rivers and the proposed designation extents of the Mad River, the Paint Creek and Vermillion River. Each housing transaction is assigned the minimum distance from its geographic coordinates to each designation and proposed segment as well as 'buffer' regions just upstream or downstream from a designation and the nearest major river not associated with the OSRP. Each housing sale is assigned to the nearest of that list of rivers. Additionally, each transaction is assigned a binary variable indicating if the sale occurred in an area currently designated as a 100-year flood hazard area.

The environmental standards for designation into the OSRP, by type of designation, are detailed in Table 1. Designating a river into the OSRP does not mean that the entire river system, nor even the entire length of an individual channel, becomes designated. A typical designation will specify segments of the river system, which may be composed of multiple stream channels or even be disconnected. Furthermore, a river system may undergo multiple designations, with segments having separate dates of designation, differing designation types, or both. For the purposes of my analysis, a 'designation' is defined by date of designation, river system, and type of designation: separate streams will share the same 'distinct designation' only if they share all those characteristics.

Table 2 briefly describes the treatment and control samples used in this essay's analysis. The table separates these samples by designation and includes their year of designation, type of designation, and which analysis they will be used in. To limit the role of unobserved variables, I restrict samples to sales less than 4km away from their assigned river segment. Treatment groups for Temporal regressions correspond to sales occurring after the designation date; the control group includes sales closest to that same designation but sold prior to the designation date. The treatment groups for the spatial regressions include sales within 4km of a within-Ohio endpoint of their designation, while being closest to the designated portion of the river; the control group is composed of sales within 4km of a within-Ohio endpoint of the designation while being closest to the 'buffered' section past the designation's endpoint. Figure 2 illustrates one example of this divide: the downstream endpoint of the Maumee designation. Some permutations of these regressions will restrict the samples further, and the full criteria for sample assignment are explained further in section 4. Distinct designations not listed in Table 2 are excluded due to lacking at least 100 qualifying sales in their control and treatment groups.

Table 3 reports summary statistics for key variables used in the Temporal regressions on modern designations; the table displays the mean and standard deviation for each variable for the control and treatment groups of each designation. The table shows that these homes are generally several decades old, are infrequently within a 100-year flood area, and usually feature garages. Asterisks imply statistically significant differences (* p<0.1, ** p<0.05, *** p<0.01) between the control and treatment groups of that designation. The table shows some significant differences between treatment and control groups. For example, Conneaut homes sold post designation are significantly older than those sold prior to designation. In general, however, these variables show close means and substantial overlap across variables for the treatment and control groups around each designation.

Table 4 shows the same summary statistics as Table 3, but for the historical designations used in the Spatial regressions. Like the designations used in the Temporal regressions, these summary statistics show old homes, frequently with garages and infrequently in a flood area. However, statistically significant differences between treatment and control groups are omnipresent in these Spatial regressions, while they were rarer in the Temporal regression. These differences generally point to homes on the designated side of the endpoint being larger and on bigger lots than the non-designated side.

4. Methodology

My identification strategy aims to estimate the capitalization effect of designation per se, controlling for pre-existing environmental quality as well as structural home characteristics. In each case, there are strong arguments a-priori that the capitalization of designation may vary by distance to the river: the further away the home, the less relevant the riparian corridor and/or the regulations surrounding the program. In the proceeding subsection, I detail a set of Temporal regressions using designations that occurred during my study period that estimate capitalization effects comparing sales from before and after the dates of designation. These regressions use a combination of spatial and temporal fixed effects as well as quasi-experimental sample specifications to rigorously identify designation's capitalization effect. These regressions have an immediate interpretation and utilize a controlled ex-ante – ex-post identification strategy: a unique contribution to the topic of Scenic Rivers programs. However, they are inherently limited to a fraction of the designated segments within the OSRP because most designations occurred decades prior to the available sales data. I supplement those Temporal Regressions with Spatial regressions at the endpoints of historical designations. These Spatial Regressions check for significant price differences between riparian properties just within versus just beyond a designation's limit.

Temporal Regressions on Modern Designations

Nine designations occurred within the time frame of my data, five of which occurred between 1999 and 2016. Of these designations, Ashtabula in 2008, Chagrin in 2002, and the Scenic and Wild type designations on the Conneaut Creek in 2005 had at least 100 sales in both control and treatment groups. I estimate a series of pooled regressions using the following format:

$$\begin{aligned}(1)\ \text{Ln (Sale Price)} &= a + \beta{*}X + \theta_1{*}\text{Near} + \theta_2{*}\text{Post-Designation}\\ &\quad + \theta_3{*}\text{Near}{*}\text{Post-Designation}\\ &\quad + \text{Designation Specific Yearly FE} + \text{error}\end{aligned}$$

where 'Near' is a binary variable that equals 1 if the sale occurred within 1km of its assigned designation and X constitutes all of the structural controls in Table 2, as well as quadratic terms for square footage and acreage. I selected 1km to represent a distance cutoff after exploratory testing showed it to yield the best model fit. I also include yearly fixed effects interacted with each designation to allow for designation-specific baselines and price trends.

I estimate several regressions following (1) with increasingly strict time-bounds for which sales are admitted into the sample; by using tighter time windows, I avoid time varying omitted variables and ambiguous transitory periods at the expense of sample size. These additional regressions include considering only sales that occurred within 2 years of the designation day, those sales that occurred at least 60 days away from the designation day, and sales that met both criteria. Table 5 presents the results of these regressions.

Spatial Regressions using historical designations

Designations do not typically cover the entire length of a river. Rather, the Ohio DNR nominates a continuous and particularly pristine segment of a river channel. In some cases, this designation is cut-off at the borders of Ohio, e.g. the Wild section of Conneaut Creek. In other cases, the designation ends within Ohio, with the rest of the river remaining undesignated. This creates a treated side within the designation and a 'control' side without. Figure 2 gives an illustration of one such point and the home sales assigned to either side of the designation divide. As I use historical designations for this analysis, the capitalization measurements represent a long-term effect of designation in contrast to the more immediate effects of the Temporal regressions above.

I consider housing sales within 4km of a point where the designated portion of a historical designation ends within Ohio. In addition to a binary variable representing being assigned to the treated river section, I include the same Near variable and structural controls as in the previous section. The following details the form of these pooled regressions:

$$\text{(2) Ln (Sale Price)} = a + \beta X + \theta_1\text{*Near} + \theta_2\text{*Designated Area}$$
$$+ \theta_3\text{*Designated Area*Near} + \text{Designation Specific Yearly FE} + \text{error}$$

In some cases, homes were nearly equidistant between the designated and non-designated portions of the river. This may raise concerns that these sales with 'Ambiguous' assignments may bias my estimates. For robustness, I also estimate models excluding sales with small differences in distance (<25m) between the designated and non-designated sides of the endpoint. For additional robustness I also estimate (2) using only sales within 2km of the river both with and without those sales with 'Ambiguous' assignment. Table 6 presents these results.

In the regressions described above, I rely on designation-year fixed effects to de-mean sale prices by river segment. While designated segments by definition share many similarities, the non-designated segments may be uniquely different as the designation may end for any of a number of reasons. Furthermore, treatment effects may be heterogeneous across river segments.

To address these concerns, I reconduct some of the above analysis while disaggregating designations. For sample size reasons, this is done only for the Spatial regressions around the down-streams ends of historical designations. I then re-estimate (2), with the widest sample specification of those described above, for each designation's downstream boundary.

5. Results

Table 5 shows the results of my Temporal regressions, which compose my top-line outputs for this essay. As I alter temporal thresholds, I reduce the number of observations from 4300 for all sales within 4km of a recent designation to 1300 sales in the most restrictive model. However, all four regressions provide the same qualitative result: a statistically significant capitalization effect from designation between 6% and 8% for homes within 1km. None of the regressions found significant benefits for being near these rivers prior to designation, nor did they find a significant capitalization effect from designation for homes more than 1km away.

The results of my Spatial regressions, shown in Table 6, are also qualitatively resistant to sample specification. In all four variations, we see a significant capitalization effect for being within 1km of the river in question, but we see no significantly different effect from river proximity for being on the designated or undesignated side of the boundary. The slight negative effect for being on the designated side is insubstantial and only weakly significant in the regression with the entire applicable sample.

Table 7 separates the results from the previous paragraph by river and focuses only on the downstream boundaries of each designation: otherwise, these regressions use the widest sample specification used in Table 6. These results show significant capitalization effects that are highly heterogeneous by river, including in sign. Stillwater's subsample shows penalties for both river proximity and being on the designated side of the boundary, that are mostly counteracted by a beneficial interaction effect. In contrast, while proximity to the Olentangy increases sale price by almost 9% ceteris paribus, homes sold nearby and on the designated portion see that riparian benefit wiped out. Sandusky shows a riparian penalty on the non-designated side of the boundary, but a capitalization effect of almost 10% for being on the designated side, while the regressions along Chagrin and Cuyahoga fail to result in any coefficients significant at the p<0.05 level due to thin data.

6. Discussion

The requirements and process for designation into Ohio's Scenic Rivers program are quite detailed, as overviewed earlier in this paper. However, on a basic level the tradeoff between protection and development is well trod in Environmental Economics; applying that field's hedonic analysis framework, I derive quantitative results on the tradeoff between promised protection as a Scenic River and further regulating public development.

My primary results show that designation increases local sale prices for homes by 6-8%. This increase is statistically significant, practically substantial, and controls for a number of endogenous factors including the river's pre-existing environmental quality.

The results from my Spatial regressions are attempting to answer a very different question. Rather than isolating the impact of designation within the moment, the hedonic analyses in Tables 6 and 7 offer potential residents a choice between different 'finished products'; these contrasted neighborhoods are more than 20 years past their designation and diverge not only by pre-exiting environmental quality but also ensuing development, possibly lopsided due to the designation itself.

At these endpoints we do not see a collective significant difference in sale price between the designated and undesignated sides of the boundary. However, disaggregating the regressions shows a large degree of heterogeneity in estimated treatment effects. This is not prima facie surprising. For example, the downstream end of

Olentangy's designation is in the outskirts of Columbus, so the negative coefficient associated with the designated end could in fact be due to the high value of proximity to a major metropolitan area. In contrast, the regression on the edge of Sandusky's downstream designation boundary showed results much more in line with the temporal regressions with a designation boundary relatively far from Fremont's metropolitan area. The location of these designations relative to metropolitan areas may not only affect background geographic patterns in home value, but the opportunity cost of regulating public development.

Regardless of why heterogeneity exists between river systems, my topline results are well controlled and tell a simple story. Designation as a Scenic River advertises the quality of these riparian corridors and promises them future protection; the home sales by modern designations show that potential residents value these protections by a substantial 6-8% premium. Neither these modern designation regressions nor our regressions by designation boundaries show strong evidence for any 'regulatory takings' from the additional oversight on development. Finally, we note that the structure of the OSRP accommodates heterogeneity: by requiring the consent of local governments, the program self-filters to those communities for whom designation doesn't unduly interfere with their own development plans.

7. Alternative Methodology: Matching

In this section, I consider and present an alternative method of identifying the average capitalization effect of Scenic River designation. I conduct nearest neighbor matching between the treatment and control groups from the Temporal and Spatial regressions above, as well as between sales near Scenic type designations and sales near two proposed but never completed Scenic type designations: the Vermillion River and Paint Creek. I review the principles behind Nearest Neighbor Matching, present my results and diagnostics from these algorithms, then briefly discuss their implications given the principal results above.

Nearest Neighbor Matching

Abadie and Imbens (2006) famous work details the fundamental properties of matching estimators used in this essay, which interprets treatment effects using the potential-outcomes framework attributed to Rubin (1973). For each sale *i,* I observe treatment status W_i and the outcome sale price.

3) $Y_i = \begin{cases} Y_i(0), if\ W_i = 0 \\ Y_i(1), if\ W_i = 1 \end{cases}$

I wish to estimate the average treatment effect on the population of the treated, the capitalization effect designation had on nearby properties:

4) $\tau^t \ = \ E[Y_i(1) - \ Y_i(0) \mid W_i = 1]$

However, I do not observe $Y_i(0 \mid W_i = 1)$. In its place, I must use data from control groups to construct proxies for the treated sample's alternative reality. This analysis fundamentally relies on several assumptions. The assumption of Unconfoundedness requires that, after conditioning on available covariates, status as treated group or control group is independent of that subject's true treatment effect. The assumption of Overlap requires that the treatment and control group's have enough overlap in their distributions of covariate values.

5) Unconfoundedness: W is independent of (Y(0), Y(1)) conditional on X=x

6) Overlap: $\eta < \Pr(W=1 \mid X=x) < 1- \eta$ for some $\eta>0$

Together, these allow for estimating the average treatment effect on the treated, the average capitalization effect for homes near designated rivers, by the difference in conditional averages between the treatment and control groups:

7) $\tau^t = E[Y\,|W = 1, X = x] - E[Y|W = 0, X = x]$

Nearest neighbor matching constructs statistical proxies using a fixed number of observations' 'nearest neighbors', defined by the distance across multi-dimensional space between the observation's covariates X. In the setting of hedonic analysis, the algorithm is essentially looking for 'twin' homes, with a similar size, matching values for binary variables such as 1(Has a Garage) or 1(Is in a Floodplain), etc. In this essay I conduct matching using the three nearest neighbors, so each member of a treatment group is matched to the three most similar sales in the control group and vice-versa.

Directly using (Abadie and Imbens 2006), the sale price in the unobserved reality for each actual observation would be imputed by the average sale price of the three nearest neighbors, and then average treatment effects on the treated would be the average difference from the observed treated value and the mean of the matched valued values from the control sample.

8) $\widehat{Y_i}(0) = \begin{cases} Y_i\,, if\ W_i = 0 \\ \frac{1}{M}\sum_{j \in J_M(i)} Y_j\,, if\ W_i = 1 \end{cases}$

9) $\widehat{Y_i}(1) = \begin{cases} \frac{1}{M}\sum_{j \in J_M(i)} Y_j\,, if\ W_i = 0 \\ Y_i\,, if\ W_i = 1 \end{cases}$

where $J_M(i)$ is the set of M matches to observation i.

However, on its own this methodology would not be $\mathrm{N}^{1/2}$ consistent; intuitively this is caused because the matching procedure will not likely find exact matches along covariates, and does not per se account for those small remaining distances between matched observations in X. To address this, Abadie and Imbens (2011) developed bias-corrected matching estimators, which I utilize in this essay. In effect, this method combines the matching framework described above, with the ability of regression equations to project imputed values across differences in covariates. Consider that regression equations use means of Y conditional on covariates X to impute $\widehat{\mu_1}$(X$_i$) as the counterfactual estimation for control observations, and $\widehat{\mu_0}$(X$_i$) as the counterfactual for treated observations. Abadie and Imbens (2011) then use these to construct bias-corrected counterfactuals:

$$10)\ \widetilde{Y_\iota}(0) = \begin{cases} Y_i\,, if\ W_i = 0 \\ \frac{1}{M}\sum_{j\in J_M(i)}(Y_j + \widehat{\mu_0}(\text{Xi}) - \widehat{\mu_0}(\text{Xj}))\,, if\ W_i = 1 \end{cases}$$

$$11)\ \widetilde{Y_\iota}(1) = \begin{cases} \frac{1}{M}\sum_{j\in J_M(i)}(Y_j + \widehat{\mu_1}(\text{Xi}) - \widehat{\mu_1}(\text{Xj}))\,, if\ W_i = 0 \\ Y_i\,, if\ W_i = 1 \end{cases}$$

with the corresponding estimator

$$12)\ \hat{\tau}_M^{bcm} = \frac{1}{N}\sum_{i=1}^{N}(\widetilde{Y_\iota}(1) - \widetilde{Y_\iota}(0))$$

Failed Designations

As described in previous sections, a stream segment nominated for designation by Ohio’s DNR does not become designated automatically: it must accrue resolutions of support by a majority of local governments. Some streams have gotten to the point of nomination or designation studies without being designated at the time of writing. These include segments of the Mad River, the Vermillion River, and Paint Creek.

Using sales near these failed designations gives my analysis an intriguing control group: river systems that came as close as possible to being designated, but aren’t designated due to a lack of support at the final step of the process. As a control, they introduce two potential biases. First, whether a segment gains the support of local governments may signal other critical characteristics of the neighborhoods and the local real estate market that I cannot control for using my set of structural characteristics. Moreover, using these sales requires matching across rivers, which assumes that the hedonic value of both treatment and structural controls are constant across distinct housing markets.

Regardless, for the interest of supplemental investigation as well as a robustness check against previous analyses, I conduct nearest neighbor matching using sales near these failed designations as the control group. Specifically, the set of sales closest to either the Paint Creek or the Vermillion River, within 4km of these proposed segments, and sold after the public release of the river's designation study. The treatment group is composed of sales near scenic-type designations, within 4km of these designations, and sold after the date of designation: this treatment group includes both modern and historical designations. I control for local environmental quality by only using designations within the same designation type as defined by the OSRP and demean sales prices to control for housing market baseline price differences.

I conduct nearest neighbor matching to the three nearest neighbors, using bias adjustment from (Abadie and Imbens 2011) where X includes the set of structural controls as well as the Near binary variable signifying 1(Within 1km). I use the inverse diagonal of the sample covariate covariance matrix as the distance metric (Abadie and Imbens 2006). I conduct this matching algorithm for the pooled treatment group, as well as individually for each distinct designation with a sufficient sample size. The results of the former are reported in Table 8, while the results of the latter are reported in Table 9.

Repeating Temporal and Spatial Regressions using the Matching Algorithm

I conduct the matching algorithm described above using treatment and control group definitions from the Temporal and Spatial regressions described in previous sections. I use the broadest sample definition from those regressions. I conduct a pooled matching algorithm, while forcing the algorithm to match within distinct designations, as well as designation specific matching algorithms; because these matching algorithms force matching within designation, I do not demean observed sales prices. The results of the former are reported in Table 8, while the results of the latter are reported in Tables 10 and 11.

Matching Results

Repeating the Temporal and Spatial regressions in matching form reached the same qualitative results. The pooled ATET from matching across the date of designation shows a capitalization effect of 8% for modern designations, while matching across designation endpoints does not yield a significant capitalization effect in either direction. Matching across failed designations to actual designations yielded a very large average capitalization effect of 20%. While this is possible, this matching algorithm has a higher average balancing error than the other two despite having the largest sample of the three.

Tables 10 and 11 detail the Temporal and Spatial matching results by designation. Some of these matching runs, such as the Spatial matching for the Cuyahoga river, resulted in large balancing errors implying insufficient overlap for an accurate result;

thus, the stupendous implied treatment effect for this designation can be discarded. The individual designation Spatial matching runs with a mean balancing error less than 2 generally adhere to the topline result of no significant capitalization effect: the notable exception is the Stillwater run, which implies a disastrous decrease in housing value for the designated side despite a relatively reasonable balancing error and large sample size. For the Temporal matching results, the Ashtabula run has a large and significant ATET capital loss of approximately 15%, although this too resulted in a larger than normal balancing error. The other runs did not yield any significant treatment effects.

Table 9 shows the results for the Failed Designation matching runs for each individual qualifying designation. These results do not show the extreme outliers present in the Spatial matching results, but generally have large balancing errors. The treatment effects are frequently significant and spread over a large range of magnitudes, including an 11% penalty for the Maumee river and up to a 61% increase in sale value, for the segment of the Chagrin river designated scenic-type in 2002. This large spread of heterogenous treatment effects does not show a clear pattern across year of designation nor balancing error to justify discarding either end of results. Therefore, I conclude that the topline average of 21% is obscuring an uninformatively large range of estimated treatment effects.

Matching Discussion

In general, these matching results suffer from a large heterogeneity in estimated capitalization effects between designations, making it difficult to confidently narrativize the pooled average results. The problems of heterogeneity and fit are if anything worse here than in the regression results presented in previous sections, justifying the pick of those regressions for my topline result of this essay's analysis.

Taking the results of these matching algorithms as they are, I show average treatment effects on the treated qualitatively similar to those implied from the regressions earlier in this essay. The Temporal matching shows the same approximate pooled capitalization effect of ~8%, while the Spatial matching shows no consistent significant effect. This could be because the benefits of designation spill over along the river more easily than news of the designation spread before it was made official. However, more analysis would be required to test this hypothesis.

On average, matching to sales near Failed Designations shows substantial value for succeeding in a designation push. However, despite controlling for structural covariates and normalizing sale prices using the S&P/Case-Shiller OH-Cleveland Home Price Index, the analysis still resulted in a spread of average treatment effects wider than both my priors and what would garner useful advice for stakeholders considering designation. For these reasons, these matching results are noted for transparency, but regretfully discarded.

Figure 1: Distinct Designations by Sub-Group, with Year of Designation / Designation Study

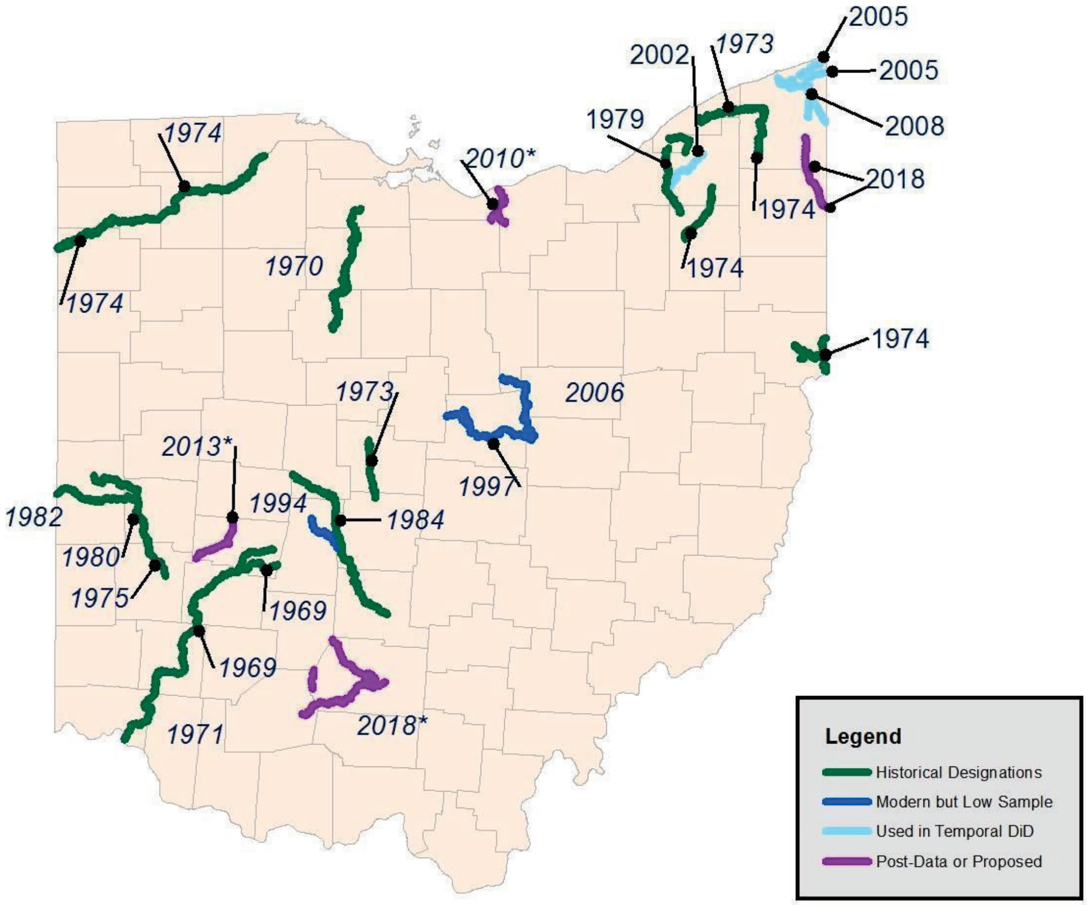


Figure 2: Treatment and Control Groups for the Spatial Regression Around the Downstream End-point of the Maumee Designation

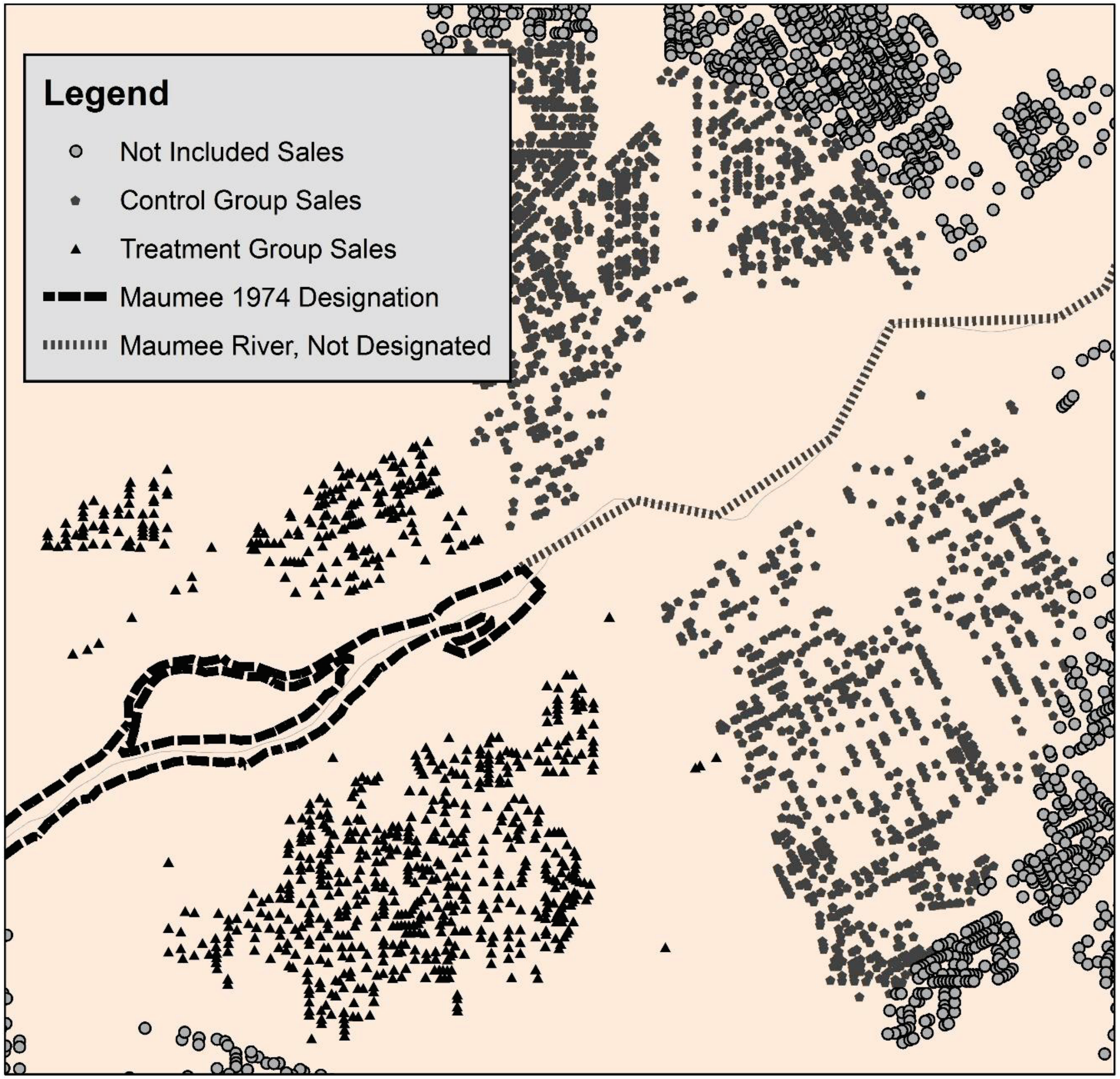


Table 1: Environmental Quality Requirements for Designated Rivers

Requirement Type	Wild	Scenic	Recreational
Free Flowing	100%	75%, connectivity to its natural floodplain for majority of length. Where not, recovered to point of supporting habitat community	60%, connectivity to its natural floodplain for majority of length. Where not, recovered to point of supporting habitat community
Roads	<10% of length within 300 ft	<= 25% of length within 300 ft	<= 50% of length within 300 ft
Bridge Crossings	Highway crossings <= 1 per 15 miles of river. Else <=2 per 5 miles of river		
Residential Dwellings	<= 2 per mile within 300ft		
Minimum Designation Length	15 continuous river miles	20 river miles, continuous or connected by other designations	20 river miles, continuous or connected by other designations
Commercial/Industrial Development	None within 300ft or visual corridor.	Some, may not impact habitat or quality of stream and floodplain	Some, may not impact habitat or quality of stream and floodplain
Impervious Surfaces	<= 5% of watershed may be covered	<= 10% of watershed. If at 10% and contained within urbanizing area then not designated	<= 10% of watershed. If at 10% and contained within urbanizing area then not designated
Native Forest/Wetland	>= 75% of length, >= 300ft. of remaining >=50%, >= 120ft	>= 25% of length, >= 300ft. of remaining >=50%, >= 120ft	>=50%, >= 120ft
Warm/Coldwater Habitat Standards	All meet exceptional standards unless limited by natural conditions. Pollution abatement must be developed if falls below.	All meet standards unless limited by natural conditions. Pollution abatement must be developed if falls below.	

Table 2: Distinct Designations and the Samples of Home Sales Each Contribute

River	Designated River Miles	Designation Year	Designation Type	N	
				Control	Treatment
Historical Designations : Discontinuity Across Designation Ends					
Sandusky	65	1970	Scenic	459	441
Olentangy	22	1973	Scenic	3404	2214
Grand	23	1973	Wild	971	830
Cuyahoga	25	1974	Scenic	79	216
Maumee	53	1974	Recreational	2418	1229
Stillwater	10	1975	Recreational	301	2320
Chagrin	49	1979	Scenic	1780	849
Modern Designations : Discontinuity Across Designation Day					
Chagrin	22	2002	Scenic	630	1927
Conneaut	5	2005	Scenic	459	472
Conneaut	16	2005	Wild	222	248
Ashtabula	46	2008	Scenic	265	143

Table 3: Summary Statistics, Temporal Regressions by Designation, Mean and (Std. Deviation)

Variable	Ashtabula		Chagrin		Conneaut		Conneaut	
	Control	Treatment	Control	Treatment	Control	Treatment	Control	Treatment
Ln(Sale Price)	11.6	11.6	12.5	12.6	11.38	11.44**	11.8	11.7
	(0.47)	(0.53)	(0.46)	(0.52)	(0.47)	(0.48)	(0.44)	(0.44)
Baths	1.6	1.8**	2.6	2.7	1.6	1.6	1.9	1.9
	(0.74)	(0.77)	(1.1)	(1.1)	(0.80)	(0.74)	(0.77)	(0.72)
Sq. Ft. 100s	15.8	16.4	35.0	36.4**	16.3	15.73	17.3	16.8
	(5.9)	(6.1)	(14.8)	(16.4)	(7.9)	(7.2)	(8.0)	(7.3)
Acres	2.1	2.7**	1.6	1.5	0.54	0.7*	1.4	1.5
	(2.1)	(2.4)	(1.5)	(1.5)	(0.89)	(0.99)	(1.7)	(1.7)
Age	46.0	44.2	38.2	43.7***	64.8	62.7	34.5	41.4***
	(23.1)	(25.3)	(24.6)	(23.3)	(24.0)	(23.0)	(26.8)	(24.3)
Garage	0.88	0.89	0.98	0.98	0.83	0.83	0.90	0.93
	(0.32)	(0.32)	(0.14)	(0.14)	(0.38)	(0.38)	(0.30)	(0.25)
100 year Flood Area	0.02	0.02	0.01	0.01	0	0.00	0	0
	(0.12)	(0.12)	(0.10)	(0.10)	0	(0.05)	0	0
N	265	143	630	1927	459	472	224	246
Designation Type	Scenic		Scenic		Scenic		Wild	

Table 4: Summary Statistics, Spatial Regressions by Designation, Mean and (Std. Deviation)

	Chagrin		Cuyahoga		Grand		Maumee	
Variable	Control	Treatment	Control	Treatment	Control	Treatment	Control	Treatment
Ln(Sale Price)	12.13	12.2***	12.2	12.2	11.6	12.0***	11.9	12.3***
	(0.49)	(0.52)	(0.4)	(0.46)	(0.43)	(0.4)	(0.38)	(0.44)
Baths	2.1	2.3***	2.5	2.3**	1.6	2.3***	1.8	2.6***
	(0.97)	(1)	(0.75)	(0.89)	(0.72)	(0.81)	(0.77)	-0.85)
Sq. Ft. 100s	20.5	22.7***	32.3	29.1**	14.8	19.0***	22.6	30.6***
	(11.8)	(13.8)	(11.6)	(12.8)	(5.6)	(7)	(8.4)	(10.9)
Acres	1	1.1	0.7	1.3***	0.28	0.56***	0.26	0.39***
	(1.4)	(1.4)	(1)	(1.6)	(0.26)	(0.81)	(0.19)	(0.32)
Age	49.9	43.2***	32.9	28.6*	58.8	32.9***	49.4	28.2***
	(22.2)	(20.9)	(14.2)	(23.2)	(28.2)	(24.3)	(20.9)	(24.7)
Garage	0.95	0.98***	0.97	0.98	0.81	0.97***	0.96	0.98***
	(0.22)	(0.16)	(0.16)	(0.14)	(0.39)	(0.17)	(0.2)	(0.14)
100 Year Flood Area	0.04	0.01***	0	0	0.02	0.01***	0.01	0.00***
	(0.19)	(0.09)	(0)	(0.07)	(0.16)	(0.08)	(0.07)	-0.03)
N	1780	849	79	216	942	811	2418	1229
Designation Type	Scenic		Scenic		Wild		Recreational	

Continued

Table 4: Continued

	Olentangy		Sandusky		Stillwater	
Variable	Control	Treatment	Control	Treatment	Control	Treatment
Ln(Sale Price)	12.4	12.4**	11.4	11.8***	11.3	11.2**
	(0.39	(0.44	(0.45	(0.44	(0.51	(0.35
Baths	2.8	2.9***	1.7	2.1***	1.8	1.4***
	(0.75)	(0.88)	(0.7)	(0.84)	(0.87)	(0.63)
Sq. Ft. 100s	21.5	23.1***	21.3	25.5***	17.5	14.8***
	(7.2)	(9.5)	(8.6)	(9.7)	(6.4)	(4.6)
Acres	0.32	0.36***	0.35	0.48***	0.11	0.20***
	(0.29)	(0.48)	(0.62)	(0.55)	(0.07)	(0.57)
Age	38.1	26.8***	60.4	40.1***	73.2	71.5
	(16.1)	(13.1)	(23.7)	(22.8)	(28.4)	(13.5)
Garage	0.98	0.99***	0.83	0.78**	0.75	0.93***
	(0.16)	(0.11)	(0.37)	(0.41)	(0.43)	(0.25)
100 Year Flood Area	0	0.01***	0.01	0.02	0	0.00**
	(0.04)	(0.08)	(0.08)	(0.13)	(0)	(0.05)
N	3404	2214	459	441	301	2320
Designation Type	Scenic		Scenic		Recreational	

Table 5: Temporal Regression Results

Variable	Entire Sample	|<2 Years |	| > 2 Months |	Both
Post Designation	-0.042	-0.037	-0.104	-0.094
	(0.021)	(0.031)	(0.054)	(0.067)
< 1km	-0.005	0.026	-0.005	0.025
	(0.054)	(0.057)	(0.044)	(0.037)
Post Designation x < 1km	0.076**	0.061**	0.075***	0.065**
	(0.017)	(0.018)	(0.008)	(0.017)
N	4366	1439	4234	1307
Within R^2	0.58	0.59	0.58	0.60

Table 6: Spatial Regression Results

Variable	Within 4km	<4km and Unambiguous	<2km	<2km and Unambiguous
Designated	-0.016*	-0.014	-0.019	-0.017
	(0.032)	(0.034)	(0.060)	(0.063)
< 1km	0.055**	0.055**	0.028*	0.029*
	(0.015)	(0.014)	(0.011)	(0.012)
Designated x < 1km	0.001	-0.003	0.09	0.005
	(0.018)	(0.017)	(0.025)	(0.026)
N	13692	13219	9308	9141
Within R^2	0.59	0.58	0.55	0.55

Table 7: Spatial Regression Results by Designation, Downstream Only

Variable	Chagrin	Cuyahoga	Grand	Maumee	Olentangy	Sandusky	Stillwater
Near	0.050***	-0.278	0.031	0.107***	0.089***	-0.077*	0.036
	(0.017)	(0.239)	(0.022)	(0.019)	(0.009)	(0.040)	(0.048)
Designated	-0.011	0.011	0.099***	0.000	-0.034***	0.075***	-0.053
	(0.013)	(0.050)	(0.020)	(0.012)	(0.007)	(0.028)	(0.036)
Designated x Near	-0.027	0.346	-0.041	-0.079***	-0.036**	0.112**	-0.010
	(0.032)	(0.243)	(0.030)	(0.023)	(0.014)	(0.049)	(0.050)
N	2247	269	1795	3647	5551	765	2606
R^2	0.65	0.71	0.66	0.69	0.73	0.66	0.31
Designation Type	Scenic	Scenic	Wild	Recreational	Scenic	Scenic	Recreational

Table 8: Matching Results, Pooled Results

Pooled Summary	Temporal Matching	Buffer Matching	Failed Designation Matching
ATET	0.08***	0.02	0.21***
SE	(0.01)	(0.01)	(0.03)
Z-Score	7.04	1.50	6.73
Mean Balancing Error	1.40	1.32	2.09
SD Balancing Error	1.48	1.57	1.54
N	4366	17511	51638

Table 9: Failed Designation Matching, Results by Designation

River	Designation Type	Year of Designation	Failed Designation Matching					
			ATET	Std Dev	Z-Score	N	Mean Balancing Error	SD Balancing Error
Little Miami River April	Scenic	1969	0.15***	(0.03)	5.13	1785	2.36	1.25
Little Miami River Sept	Scenic	1969	0.01	(0.04)	0.32	1384	3.25	2.07
Sandusky	Scenic	1970	0.06*	(0.03)	1.78	2697	2.05	1.37
Little Miami River	Scenic	1971	0.32***	(0.04)	7.74	18651	2.19	1.52
Olentangy	Scenic	1973	0.05	(0.06)	0.84	13262	2.14	1.34
Cuyahoga	Scenic	1974	0.21	(0.03)	6.26	2392	2.33	1.35
Maumee	Scenic	1974	-0.11**	(0.05)	-2.29	1449	2.99	1.60
Chagrin	Scenic	1979	0.46***	(0.05)	9.48	6625	2.54	1.78
Stillwater	Scenic	1980	-0.11***	(0.03)	-3.44	5590	2.03	1.43
Stillwater	Scenic	1982	0.08***	(0.03)	3.05	2704	2.04	1.46
Big & Little Darby Creeks	Scenic	1984	0.13**	(0.05)	2.50	3305	2.32	1.30
Darby	Scenic	1994	0.28***	(0.04)	7.49	1617	2.79	1.85
Kokosing	Scenic	1997	-0.07**	(0.03)	-2.07	1909	2.38	1.54
Chagrin	Scenic	2002	0.61***	(0.04)	17.36	3073	2.66	1.53
Conneaut	Scenic	2005	0.14***	(0.04)	3.85	1618	2.92	2.14
Ashtabula	Scenic	2008	0.03	(0.06)	0.55	1289	3.73	1.99

Table 10: Temporal Matching, Results by Designation

River	Designation Type	Year of Designation	Temporal Matching					
			ATET	Std Dev	Z-Score	N	Mean Balancing Error	SD Balancing Error
Chagrin	2002	Scenic	-0.02	(0.03)	-0.67	470	1.57	1.57
Conneaut	2005	Scenic	0.03	(0.03)	1.04	931	1.11	1.37
Conneaut	2005	Wild	-0.02	(0.03)	-0.67	470	1.57	1.56
Ashtabula	2008	Scenic	-0.15***	-0.04	-3.82	408	2.68	2.19

Table 11: Spatial Matching, Results by Designation

River	Designation Type	Year of Designation	Spatial Matching					
			ATET	Std Dev	Z-Score	N	Mean Balancing Error	SD Balancing Error
Sandusky	1970	Scenic	0.19***	(0.03)	5.58	900	2.06	1.84
Olentangy	1973	Scenic	-0.02	(0.01)	-1.62	5618	0.95	1.32
Grand	1973	Wild	0.04	(0.02)	1.56	1801	1.54	1.88
Cuyahoga	1974	Scenic	8.83***	(1.01)	8.74	295	4.90	3.28
Maumee	1974	Recreational	-0.00	(0.02)	-0.01	3647	1.34	1.12
Stillwater	1975	Recreational	-3.95***	(1.10)	-3.59	2621	1.66	1.92
Chagrin	1979	Scenic	0.03**	(0.01)	2.37	2629	1.68	1.72

Chapter 2. Stress Testing Under Successive Droughts: Crop Insurance Indemnification from Already Vulnerable Acres

1. Introduction

Drought is a leading risk factor for agriculture in the United States, and next to flooding drought is the largest driver of crop risk for Corn and Soybeans in the Midwest. In the disastrous drought of 2012, total drought related indemnities for insured Corn and Soybeans exceeded $12 billion.[6] While that drought is particularly notable in modern history, correlated losses frequently create ‘disaster years’ in agriculture. Duncan and Myers (1997) show that these systemic losses prohibit private sphere solutions to crop insurance demand, creating both a mission and challenge for the USDA’s Federal Crop Insurance Program (FCIP) (Glauber 2004). These problems may get substantially worse in the future. Caparas et al. (2021) predict that droughts could become 50% worse by 2050 due to climate change. The possibility of systemic Midwestern droughts like those seen in the Western United States poses a serious risk to economically critical commodities.

Through the FCIP the private and public spheres share this drought risk. The FCIP’s premiums are heavily subsidized as part of the USDA’s efforts to achieve 100% uptake and completely replace ad-hoc disaster insurance (Glauber 2004). Crop insurance products offered by the FCIP exist among a variety of imperfect substitutes to reduce farm risk; these options include financial means such as the futures market and

[6] USDA Cause of Loss Data

agronomic means such as investment in irrigation infrastructure. Farmers choose among these according to both their needs and potentially perverse incentives. Their choices determine the financial resiliency of their farms as well as the real resiliency of the food system to weather shocks.

Evidence exists that by adding to the availability of financial tools, the FCIP may dissuade U.S. farms from taking agronomic precautions: for example, Annan and Schlenker (2015) show that insured acres were more sensitive to extreme heat than non-insured acres. Adverse selection problems in the FCIP may intersect with and worsen these moral hazards. If farmers can predict their changing risk levels faster than USDA pricing is scheduled to adapt, then farmers may 'pile-in' to premium insurance contracts for high-risk years, resulting in large indemnifications, large amounts of lost crops for consumers, and a large delivery of federal subsidies to farmers and private underwriters. Ker and McGowan (2000) show potential for the private sphere to use weather 'warning signs,' specifically El Niño phenomena, to signal insurance companies to increase written contracts in vulnerable areas and thus generate additional rents.

Successive droughts offer an intuitive opportunity for weather adverse selection that is not yet covered by the literature on the FCIP. Successive droughts, when drought in one year is followed by further dry conditions in the next, pose particular danger to crop yield because soil water levels and irrigation water sources are already depleted from the previous year (Scanlon et al. 2012). While farmers may have limited ability to predict drought in the coming growing season, they can easily observe the prior droughts and their remaining water sources when deciding to purchase crop insurance. While crop insurance prices may update in response to particularly dramatic events, the normal pricing cycle of updating every 3 crop years leaves large windows for adverse selection.

If such strategies prove successful, they discourage farmers from adjusting practices to successive droughts as a right-tail risk. By neglecting dynamic, conditional risks such as successive droughts, the USDA and the literature risk poorly understanding farm's right-tail risk, leading to mispricing crop insurance and misspecifying models of farm risk management.

This essay considers the impact of successive droughts on crop insurance indemnification for irrigated and non-irrigated acres of staple Midwestern crops. By utilizing insurance indemnification information, this essay identifies the impact of drought and successive drought on the severe and widespread harvest failures that endanger food supply security and bedevil private attempts to insure agriculture (Duncan and Myers 1997). To preview my results, I find that successive droughts result in a large and significant increase in crop indemnification and that USDA pricing is not rapid enough to balance the resulting loss ratios. I find this effect across crop insurance plan categories, and for both irrigated and non-irrigated acres. These results indicate that irrigation adoption to protect against singular droughts may result in well-protected farmers. However, successive droughts exhaust the protective capacities of irrigation, resulting in large losses for subsidized government insurance offerings and widespread yield failure.

This essay continues with a brief review of literature on crop insurance and drought experience in the United States. Following, I describe my data sources, including the United States Drought Monitor, my derivation of binary drought variables from it and a summary of alternative transformations. In section 4, I describe the beta and linear regressions I use to estimate the effect of successive droughts on crop indemnification, and I present these results in section 5. I conclude with a brief discussion of plausible interpretations and their implications for the FCIP.

2. Literature Review

Glauber (2004) provides an overview of perennial issues facing the FCIP and crop insurance generally. In addition to moral hazard and adverse selection crop insurance faces the challenge of correlated, systemic losses. These correlated losses mean that the insurer cannot use the Law of Large Numbers to avoid disastrous years; Duncan and Myers (1997) formally describe this phenomenon and diagnose it as a terminal problem for privately sourced multi-peril crop insurance.

The USDA FCIP attempts to resolve this issue with offerings that are re-insured by the federal government, with the pricing and details of the policies centrally controlled by the USDA's Risk Management Agency (RMA). The program has grown extensively over the years since its inception, due to a central push to replace ad-hoc disaster payments as well as generous subsidies for the insurance premium. While the USDA RMA nominally sets pre-subsidy premium rates to be actuarially fair, the premiums farmers actually pay are far lower. At the extreme, 'catastrophic' yield loss (=>50%) can be insured against for only a processing fee. Farmers may 'buy up' to reduce their deductible and/or purchase more extensive 'revenue insurance', both at premium rates far lower than the USDA RMA's actuarial estimates. The result is a program that insured approximately 87% of Corn and Soybean acres planted in the United States during the 21st century.[7]

Drought is a classic risk to crop yields and of key interest in the USA's 'Corn Belt'. In the 21st century, drought related causes alone were responsible for over 37% of all indemnified Corn and Soybean acres in the FCIP, a total indemnity value of over $29 billion. Mafoua and Turvey (2004) provide an overview of how weather events, including drought, affect crop insurance loss ratios. While Bundy et al. (2022) found that losses to drought for Corn have been decreasing in recent years, current climate models predict that losses to drought could sharply rise in the near future. Caparas et al (2021) estimate the crop failures for these commodities could become 50% more common, or worse, by 2050. The potential for drastic increases in drought risk, as well as the economic and social importance of these commodities, incentivize us to understand this crop risk and its intersection with crop insurance in detail.

A healthy crop requires water to be absorbed from the water table into the root system of the plant. This basic description includes a great deal of complexity, where water uptake can depend on crop demand, root development and distribution, soil properties and the current water distribution in the soil (Wang and Smith 2004). The United States Drought Monitor (USDM) combines a multitude of factors to describe soil dryness and drought conditions in a single index. This index has been previously used by Kuwayama et al (2018) to parsimoniously regress farm losses on drought. This essay uses the USDM as the basis for determining local drought conditions.

3. Data

[7] USDA Summary Of Business Files, and USDA NASS Surveys

The United States Drought Monitor

The U.S. Drought Monitor[8] tracks the presence and severity of drought across the United States as part of a partnership between the University of Nebraska-Lincoln and government agencies. The USDM aggregates a range of measures of drought severity into a single ordinal variable that classifies land into one of 6 categories: the lack of abnormal dryness, and 5 categories of increasing drought severity from D0 through D4. These categories are described in Figure 3, taken from the USDM website.[9] These categories encompass a very wide range of potential losses. For example, D0 describes Abnormally Dry, and typically appears as a transition state between normal conditions and true drought. D2 and above capture more severe droughts, where economic damages are likely. D3 and D4 label drought conditions substantially worse than normal expectations: D3 and D4 are associated with multiple components of the USDM, e.g. the CPC Soil Moisture Model, entering the 5th percentile or lower. At this severity of drought, major crop losses are expected.

As the available data on agriculture is predominantly published at the county-year level, each crop year in my data is associated with 52 weeks of data describing the % of that county in each drought monitor category during that week. To operationalize this data, I explore several transformations. One possible transformation, also used by Kuwayama et. al (2018), is to take the weighted average of the proportion of a county within each category of drought over the crop year from previous harvest to the next. The resulting 5 variables, for county i and crop year t, are defined as:

[8] https://droughtmonitor.unl.edu/

[9] https://droughtmonitor.unl.edu/About/AbouttheData/DroughtClassification.aspx

1) $D\#_{i,t} = \sum_{w}(proportion\ of\ county\ i\ in\ D\#\ drought\ during\ week\ w)/52$

While these variables are intuitive, they are also somewhat limiting. They do not account for the timing of droughts, nor do they allow for non-linear effects from drought. For an example of these complications, Cakir (2004) details experiments with Corn showing a relatively high tolerance for irrigation omission in the milk and vegetative stages.

These variables are also, by construction, correlated as mutually exclusive ordinal categories. Keeping these categories as separate variables complicates the process of interacting one year's drought with the previous year's: the full dot product would result in many highly correlated variables and would make coefficient interpretation challenging.

To overcome these challenges, the Drought Severity and Coverage Index (DSCI) has been proposed as a means of aggregating and weighing these ordinal categories. The DSCI is calculated as follows:

2) $DSCI_{i,t} = D0_{i,t} + 2*D1_{i,t} + 3*D2_{i,t} + 4*D3_{i,t} + 5*D4_{i,t}$

However, as section 7 elaborates, the weights proposed in the DSCI do not match well with the coefficients observed from regressions of loss experience on the full 5 USDM categories.

This essay instead collapses the USDM drought index into a pair of binary variables. If D2, or any more severe category, is present during the standard growing season the year is labeled as featuring a Drought. If D3 or D4 are present, the year is also labeled as being in Extreme Drought. D2 and D3 were chosen as cut-off categories due to their definitions as droughts severe enough to harm crop production while still having a much thicker presence in the data than D4. While theoretically binary variables may miss classify a year if D2+ droughts were highly transitory in a region, this is extremely rare in practice. Droughts are generally persistent and widespread, and the spatial correlation of droughts is currently increasing in the United States (Ganguli and Ganguly 2016). Furthermore, because soil takes time to dry, severe drought generally develops after an area transitions through the less severe categories of drought. Successive Drought is then defined as a county-year that was in Drought the previous year also being in Drought during the current year.

Figure 4 displays the number of recorded Droughts, using my criteria, between 2000 and 2021 by county within the States used in my analysis. This figure shows that Droughts are highly correlated with geography: Pennsylvania at the eastern edge of this dataset has a noticeably smaller frequency of Droughts than states such as South Dakota or Nebraska. Furthermore, this figure shows how common Droughts are so far in the 21st century. Broad regions of the Midwest experienced Drought during the majority of years between 2000 and 2021.

Crop Insurance Data

I consider the effects of drought on Grain Corn and Soybeans within the 'Corn Belt' states in the Midwest. The full list of included states is given in Table 12 and comprises the 13 states with the highest estimated total acres planted to Grain Corn in the 21st century.[10] Figure 5 shows the states included, as well as the density of insured acreage by county. Data on the quantity of acres insured and the quantity of insurance losses by crop, plan, location, year and cause were obtained from the USDA RMA

[10] That list is not one-to-one with the list of states that produce the most Soybeans, but the two statistics are highly correlated; Soybeans and grain Corn commonly feature as part of the same crop rotation.

Summary of Business and Cause of Loss public files. While this data breaks down records of insurance purchases by irrigation, it does not break down loss experience by irrigation. For this reason, I control for irrigation using the percent of an observation of insured acres that were irrigated. Data on total acres planted and annual yields were obtained from the USDA NASS quick-stats system. Both of these data sources include years prior to the 2000 cut off used here. However, because crop insurance policies and general rates of participation have changed drastically over the past decades, I restrict analysis to 21st century crop years.

Weather Data

Henckel (1964) explains a fundamental interaction between drought and a crops inability to control their temperature during extreme heat. Cohen et. al. (2021) show this interaction effect between heat and drought empirically, and in general extreme heat has shown to be an important covariate for predicting commodity crop losses (Lobell et al 2013). Therefore, I include weather variables weighted from a 2.5 x 2.5 mile grid to the farming areas of counties over a March-November growing period, using data and methodology from Schlenker and Roberts (2009). This data is based on the PRISM weather data set, with missing values filled in by distance weighted averages from surrounding stations. Degree-day calculations are based on the integral of temperature over some baseline, where temperature over the course of the day-night cycle is assumed to follow a sine-wave hour by hour. Following both Schlenker and Roberts (2009) and Annan and Schlenker (2015), I use 10°C as the baseline for moderate heat for both crops, 29°C as the baseline for extreme heat for Corn and 30°C as the baseline for extreme heat for Soybeans. Additionally, these regressions include total precipitation and total precipitation squared, calculated in the same manner.

Summary statistics for all these variables are shown in Table 13. These statistics indicate that roughly 1/5 of these crops' insured acres were indemnified, and that almost half of those indemnifications were attributed to drought. The high variances on those proportions are due to the highly correlated nature of systemic crop failures. The table also gives a top-line summary of the drought frequency shown in Figure 4. By my criteria and within this sample of county-years, Drought occurred in ~26% of county-crop-years and Extreme Drought occurred in ~12% of crop years. Finally, the average rate of crop insurance saturation for both crops is slightly above 75% of planted acres as estimated by NASS surveys. These statistics are in line with typical modern insurance saturation statistics for commodities and show that this essay's population of insured acres represents most of these commodities' planted acres in the same county-years.

Fixed Effects

In addition to these control variables, the regressions in this essay also include fixed effects. All regressions include temporal fixed effects for every crop year, and most include spatial fixed effects at the county level. The regressions for loss ratios within a single category of insurance plan use state-level fixed effects.

4. Methodology

My principle set of regressions will investigate how Drought in the previous year affects indemnification to Drought. After establishing the link between Successive Drought and acreage loss, I consider similar regressions using Loss Ratios as the dependent variable, to control for background expected risk. Significant coefficients here imply that not only are successive droughts a particular danger to crop loss, but also that the USDA’s pricing regime fails to keep up with this inter-temporal process.

As a dependent variable, the percent of insured acres determined for loss is naturally bounded between 0 and 1. Ordinary least squares regression is poorly suited to such data, so I use beta regressions.

Beta Regressions

With a transformation of its traditional α and β parameters, the probability density function of the beta distribution may be defined as follows:

3) $$f(x \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi) * \Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}$$

with μ= α/(α+β) as the mean parameter and ϕ= (α+β) as its precision parameter. The beta distribution is defined for *x* on the interval (0,1) and may take a variety of shapes, depending on its parameters. Beta regression (Kieschnick and McCullough 2003; Ferrari and Cribari-Neto 2004; Smithson and Verkiulen 2006; summarized by Douma and Weedon 2019) assumes that the data generating process follows some beta distribution and that the mean parameter may be modeled as follows

4) $$g(\mu)= g(E[Y|X]) = \mathbf{X} \cdot \boldsymbol{\beta}$$

where, μ is the mean parameter of the data generating process’ beta distribution, and g() is a link function used to force the linear prediction into the acceptable bounds. **X·β** is the dot product of my regression variables and their coefficients: note that the *x* and β from (3) are not the same as **X** or **β** from (4).

In the analysis that follows, I use the logit link function:

5) $$g(\mu) = \log(\mu / 1-\mu) = \mathbf{X} \cdot \boldsymbol{\beta}$$

The coefficients $\boldsymbol{\beta}$ and the precision parameter ϕ are determined by maximum likelihood.[11] These coefficients do not have the same interpretations as in a linear regression. Instead, I consider the odds ratio that they imply for each variable, which gives the multiplicative change in the odds of indemnification for local acreage. For any $\boldsymbol{\Delta X}$, the ratio in odds ex-post - ex-ante equals:

6) Odds Ratio ($\boldsymbol{\Delta X}$) = exp($\boldsymbol{\Delta X} \cdot \boldsymbol{\beta}$)

Beta regressions can only process data inclusively within 0 and 1, not results where none nor all the insured acres were indemnified. In the case of 100% indemnification, these events are rare and discarded from this analysis. County-crop-years without any indemnified acres are also uncommon, albeit less so, representing approximately 4.6% of the data set. These observations are also dropped from this essay's beta regressions.

Loss Ratio Regressions: Can the USDA Pricing Cycle Keep Pace?

Acreage indemnified measures the scope of losses and is critical to understanding the impact of successive droughts on the resiliency of crop production. However, it does not control for expected risk levels, which influences how much the USDA charges for buy-up insurance plans. If successive droughts increase losses beyond the USDA's compensating pricing scheme, then they present an opportunity for temporal adverse selection as well as a threat to food production.

The metric to judge indemnification against expected risk is an insurance plans' loss ratio, which is equal to their indemnities paid out divided by their premiums

[11] The beta regressions used in this essay were estimated using the betareg R package by Cribari-Neto and Zerileis (2010).

assigned. To supplement the beta regressions described above, I conduct a series of least squares regressions on reported loss ratios. These regressions use the same covariates as the beta regressions.

Regressions Within Insurance Plan Categories

Insurance plans offered by the FCIP differ by both their deductibles and whether they insure against drops in futures prices. Therefore, while aggregating to the county-level, as above, gives a simple overview of the situation facing the FCIP, "% Indemnified" as a dependent variable is fundamentally defined differently for different insurance plans. Furthermore, farms' choices from an insurance menu may signal farm or risk characteristics through several vectors, some of which Makki and Somwaru (2001) identify. For these reasons, I break down the dataset by category of insurance plan and repeat the above analyses while only including contracts that share similar characteristics.

Plans insuring crop Yields are broken down into Catastrophic insurance, the most heavily subsidized option with a deductible of 50% of expected yields, and Buy-Up plans which include any Yield plan with a higher coverage level. Alternatively, a farm may purchase Revenue Insurance contracts, which also insure against drops in price as measured by the commodity's futures market. I split Revenue Insurance plans into groups that were sold pre or post 2011; the USDA changed which specific kinds of Revenue Insurance plans were available between those crop years. Other types of plans, such as Group insurance, are also available for Corn and Soybeans. However, they represent an insubstantial portion of the plans purchased and I drop them from this part of the analysis.

5. Results

I begin by reminding the reader that beta regression results should not be interpreted as is typical for linear regression. Coefficients and their associated variables in a beta regression add linearly before entering the link function, but no final coefficient directly compares to a linear change in the dependent variable, % Acres Indemnified. Instead, the odds-ratio for a combination of changes in variables is given by taking the product of all the odds-ratios associated with those coefficients. For example, consider a comparison in total indemnification rates for Corn between a county facing Successive Drought and one not, neither irrigated. The odds-ratios for Drought and Successive Drought for this, as seen in Table 14, are 1.51 and 0.95 respectively for a multiplied odds ratio of ~1.43, implying that the county with the successive drought will have slightly more than a 40% greater 'odds', or ratio of indemnified acres to non-indemnified acres.

% Acres Indemnified - County Level

Table 14 presents the coefficients on g(μ) from the county-level beta regression results for both Corn and Soybeans, for both total indemnification and indemnification to drought. In all four of these regressions, Drought posed a major risk to these crops, resulting in 30-50% greater odds of acreage lost. The interaction term between Drought and irrigation is negative, implying a decreased risk to crops, and the magnitude of this interaction term is substantially larger than even the sum of its components, Drought and Proportion Irrigated. These results show that better irrigated counties facing Drought have lower crop risk than even non-irrigated counties without Drought, or other years within the same county. This is within the priors given in this essay's introduction. Irrigated acres are resistant to Drought and years in Drought are resistant to excessive wet stress, the other main risk factor in this region.

Successive Drought significantly increases the average severity of losses from drought for both crops, but the magnitudes of these odds ratios are insubstantial for all four regressions. Based on these regressions, Successive Droughts are only associated with additional crop risk with irrigated acres, and this intersectional effect can be large. Controlling for other factors, including the Drought occurring, a Drought preceded by another is associated with more than doubling the odds of Corn acres indemnified for a fully irrigated county relative to a non-irrigated one.

Insurance Saturation significantly and substantially raises the total proportion of crops indemnified, even after controlling for weather, irrigation, and fixed effects. Controlling for other factors, including Drought, Extreme Drought is associated with further crop losses, but only for Corn.

% Acres Lost by Insurance Plan Category

Tables 15 through 18 show results from the beta regressions of % Acres Indemnified within each insurance plan category. The results of these 'within plan' regressions broadly show the same qualitative results as each other and the results from the pooled regressions in Table 14. There are, however, some marked differences.

A general pattern is that these within-plan results show less extreme impacts from the interaction of irrigation and Drought. For example, in the Corn county-level results, Irrigated*Drought yielded an odds-ratio of 0.13 and Irrigated*Successive Drought an odds-ratio of 2.64. In contrast, only post-2011 Revenue Insurance shows a Irrigated*Drought odds-ratio less than 0.2 and all Irrigated*Successive Drought odds-ratios but the one for Catastrophic Yield insurance are 2.3 or less. Despite this, the general pattern of Successive Drought exhausting the protective qualities of irrigation holds for most of these within-plan regressions. Only in the case of Catastrophic Yield coverage for Soybeans did the protective interaction between irrigation and Drought carry over through Successive Drought.

Effects on Loss Ratios

Tables 19 and 20 detail the results from my Loss Ratio regressions on Corn and Soybeans, respectively. The results on the key variables of interest show a pattern similar to the regressions on acres lost. Successive Droughts per se are linked only with either insignificant or relatively insubstantial effects on loss ratios. However, for well irrigated observations growing Corn Successive Droughts ‘exhaust’ some of the protective capacity of irrigation. Furthermore, irrigated observations undergoing Drought incur lowered loss ratios than non-irrigated observations not experiencing drought. For Corn counties, better irrigated years are associated with higher loss ratios. Notably, that association does not hold up after I break down county-level data by insurance plan. This may imply that the effect is due to sample composition changes associated with irrigation. For both Corn and Soybeans, irrigated county-years per se are associated with lower loss ratios despite USDA premiums adjusting for irrigation practice. For all regressions in this section, Extreme Drought leads to higher loss ratios.

6. Discussion

Drought not only damages crops, but also lowers local water tables and exhausts irrigation sources. This basic biophysical hypothesis predicts that successive droughts may be a significant and often overlooked factor in assessing crop risk. Because USDA premium rates do not update every year, this risk potential to USDA FCIP loss ratios can be compounded by adverse selection: farmers that know their capacity to resist drought is exhausted could use that asymmetric information to buy into the insurance program for the most dangerous, and therefore profitable, years.

This essay investigates the impact of successive droughts, irrigation, and their interaction on crop indemnification in the Midwest using a series of regressions that consider various sub-samples and risk metrics. After showing that irrigation struggles to lower the percent of acres indemnified during a successive drought, I show that this effect carries over into the program's loss ratios: successive droughts are a substantial concern for irrigation investments and other farm practices, and the USDA premium setting procedure is not fully controlling for this impact. These patterns in the pooled sample exist across insurance plan categories.

Irrigated acreage under drought suffers lower losses than non-irrigated acreage under no drought, and this result may seem odd at first glance. However, there is a simple and plausible explanation for this pattern. A lower water table from previous and/or current drought can enable faster absorption into and drainage from soil that would otherwise risk flooding. Losses from flooding compete with losses due to drought for the most serious crop risks for commodities in the Midwest, and these two sources of losses are strongly negatively correlated within years. This would also explain why the magnitude of the effect of drought on total acreage losses is weaker than its effect on drought losses: a drought 'protects' from losses due to flooding in the same way irrigation protects from losses to drought. An observation under drought while completely irrigated is an observation where farmers have control of the quantity and scheduling of water supplied to their crops. This leads to safe yields, so long as the supply of water for irrigation does not run out.

In both irrigated and rain-fed systems, successive droughts may deplete local water resources. However, these systems use dissimilar sources of water. Rain-fed systems depend on local water tables where long-lasting effects would be highly dependent on soil type and the root depth of the relevant crop compared to irrigated systems. For these reasons, it is within my priors that irrigated and non-irrigated systems have significantly different reactions to successive droughts, as shown in this essay’s results.

These regressions robustly display a predictable and important phenomenon for crop risk and for the USDA’s FCIP. The results immediately imply a source of misratings and adverse selection in the USDA FCIP that is understudied in the current literature. Furthermore, the results of the loss ratio regressions imply that the current pricing updating cycle is not capable of keeping up with successive droughts as a dynamic risk factor. More broadly, this essay serves as an example of how agricultural data from a given year is informed by farmers' previous experiences.

7. Considering the USDCI to Represent Drought

The interacting variables framework used in this essay’s regressions requires drought to be represented in a single variable. This essay used binary indicator variables to achieve this from the USDM, which categorizes drought experience into 6 ordinal severities. However, an alternative strategy may be to construct an index from those ordinal categories, and the USDCI statistic has been proposed by the USDM itself as a method of doing so. This section investigates and rejects that proposal for this essay’s analysis.

Table 21 presents the results of linear regressions of loss ratios for Corn and Soybeans. These regressions include the same covariate structure as in the rest of this essay, but with the 1(Drought Occurred) binary variable replaced by either the USDCI statistic or the average % of land in each drought category. Interaction terms instead utilize the USDCI statistic in both versions of each regression. This is only one example of tests that could be run contrasting these as covariates, but the results are representative of the ones I conducted.

Generally, more severe categories of drought are associated with increases in loss ratio of greater magnitude. However, this trend is not linear, nor does it follow the proportions outlined by the USDCI. For example, D2 has a smaller coefficient in both commodities' regression than D1, despite the USDCI assuming that D2 would have a 50% greater impact. In the case of Soybeans, the exponential increase in coefficient magnitude from D2 to D4 is far greater than what the USDCI weights would expect; the coefficient on D4 is 9 times as great as that on D2 and 3 times as great as that on D3.

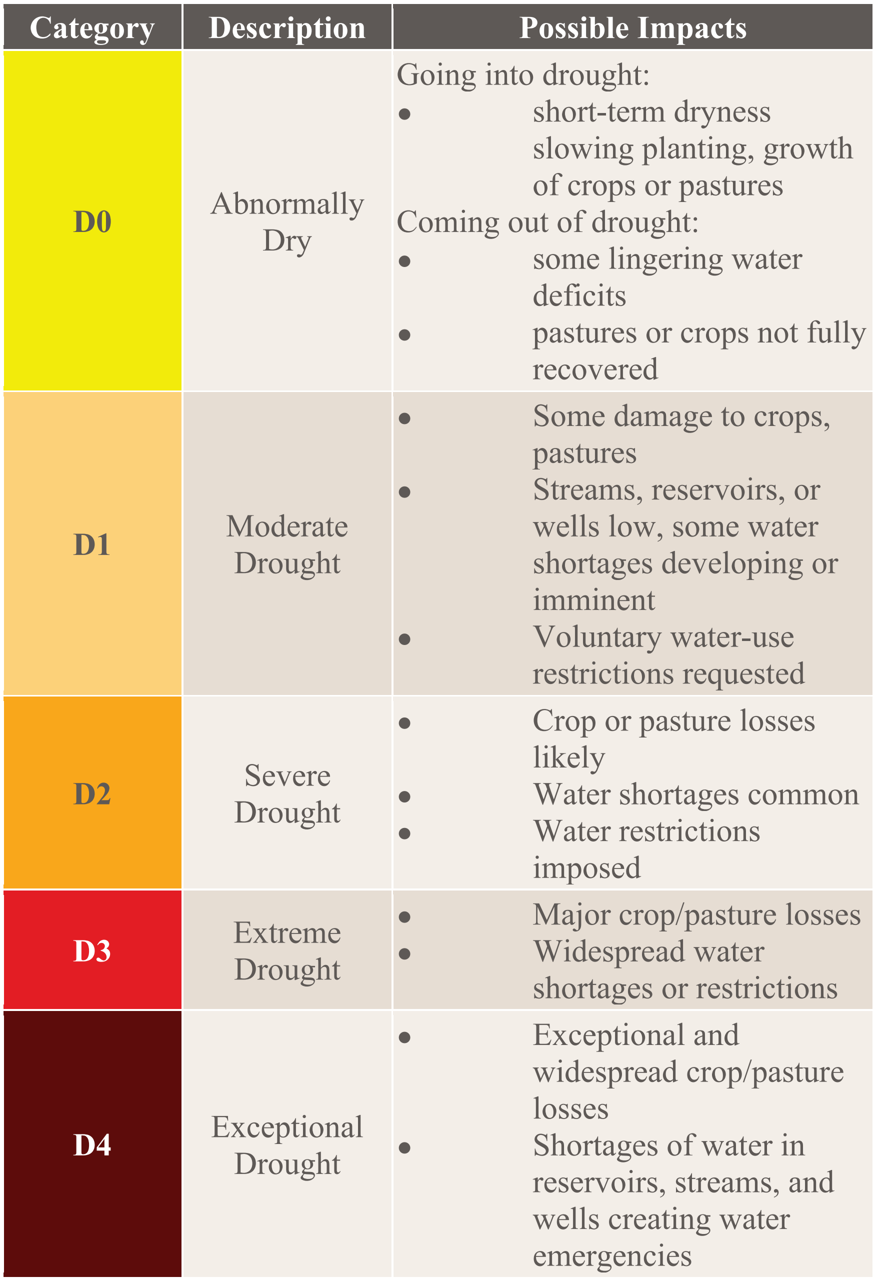
When we contrast the coefficients on these USDM categories with the one on the USDCI itself, the coefficient on the USDCI do not generally match what we would expect by taking the category coefficient and multiplying it by its USDCI weight. In other words, an increase in a USDM drought category at the expense of No Drought does not have the same estimated effect as an increase in the USDCI of the commensurate amount. Of the 10 examples here, only D4 for Corn has a close equivalency with the USDCI coefficient (5*1.73 = 8.65 v.s. 7.63) and several are off by a factor of 2 or more.

The regression results also imply that D0, 'Abnormally Dry', receives a negative coefficient implying that it lowers losses relative to normal conditions. This is sensible if

D0 is associated with lowered chances of flooding or wet soil stunting root growth. However, it is the opposite sign of what the USDCI imputes, or indeed any weight something like the USDCI would be expected to choose. The reader should note that these results exist despite precipitation and its square being included as control variables.

The poor fit of the generated coefficients implies that the USDCI does not appropriately weigh USDM categories for either of this essay's example commodities. The issue of the insuring property of mild dryness further implies that anything equivalent to the USDCI, a single linear index of drought, may be incompatible with modelling drought experience as a regressor for crop loss experience in reduced form studies.

Figure 3:United States Drought Monitor Categories



Category	Description	Possible Impacts
D0	Abnormally Dry	Going into drought: • short-term dryness slowing planting, growth of crops or pastures Coming out of drought: • some lingering water deficits • pastures or crops not fully recovered
D1	Moderate Drought	• Some damage to crops, pastures • Streams, reservoirs, or wells low, some water shortages developing or imminent • Voluntary water-use restrictions requested
D2	Severe Drought	• Crop or pasture losses likely • Water shortages common • Water restrictions imposed
D3	Extreme Drought	• Major crop/pasture losses • Widespread water shortages or restrictions
D4	Exceptional Drought	• Exceptional and widespread crop/pasture losses • Shortages of water in reservoirs, streams, and wells creating water emergencies

Table 12: Included States, and the Average Insured Acreage Per Year for Corn and Soybeans[12]

State	Corn			Soybeans		
	Planted	Insured	Ratio	Planted	Insured	Ratio
Illinois	11,200,000	9,400,000	84%	9,800,000	7,600,000	78%
Indiana	5,400,000	4,200,000	77%	5,500,000	4,000,000	73%
Iowa	12,600,000	11,900,000	94%	9,700,000	8,900,000	92%
Kansas	4,100,000	3,900,000	95%	3,700,000	3,000,000	80%
Michigan	2,200,000	1,600,000	73%	2,000,000	1,400,000	70%
Minnesota	7,400,000	7,200,000	97%	7,200,000	6,900,000	95%
Missouri	3,000,000	2,800,000	91%	5,200,000	4,300,000	84%
Nebraska	8,700,000	8,300,000	95%	4,900,000	4,500,000	92%
North Dakota	2,200,000	2,600,000	114%	4,500,000	4,600,000	101%
Ohio	3,300,000	2,600,000	77%	4,600,000	3,300,000	71%
Pennsylvania	1,300,000	700,000	50%	500,000	300,000	60%
South Dakota	4,700,000	5,000,000	107%	4,400,000	4,500,000	101%
Wisconsin	3,700,000	2,600,000	71%	1,700,000	1,300,000	73%

[12] The Dakotas here show more insured acres than planted acres, within the same Years and Crop. This table is directly quoting data from the USDA, which may include survey sampling errors when estimating acreage planted. For analyses in this paper, if surveyed planted acreage is lower than known insured acreage, planted acreage is set to the level of known insured acreage

Table 13: Covariates and their Summary Statistics, County-Level of all Corn or Soybeans

Variable	Corn	Soy
	County	County
Year	2,011.0	2,011.1
	(6.0)	(6.1)
Acres Planted	80,454	69,121
	(62,939)	(50,986)
Acres Insured	63,753	53,110
	(57,553)	(46,920)
% Indemnified	0.22	0.20
	(0.24)	(0.21)
% Indemnified to Drought	0.09	0.08
	(0.18)	(0.16)
Drought Occurred	0.26	0.26
	(0.44)	(0.44)
Extreme Drought Occurred	0.12	0.11
	(0.32)	(0.31)
% Irrigated	0.12	0.11
	(0.24)	(0.25)
Precipitation	776	785
	(205)	(200)
Moderate Degree Days	1,796	1,829
	(343)	(345)
Extreme Degree Days	34	22
	(37)	(29)

Figure 4: Frequency of Droughts, By County-Years, Over States Included in this Analysis

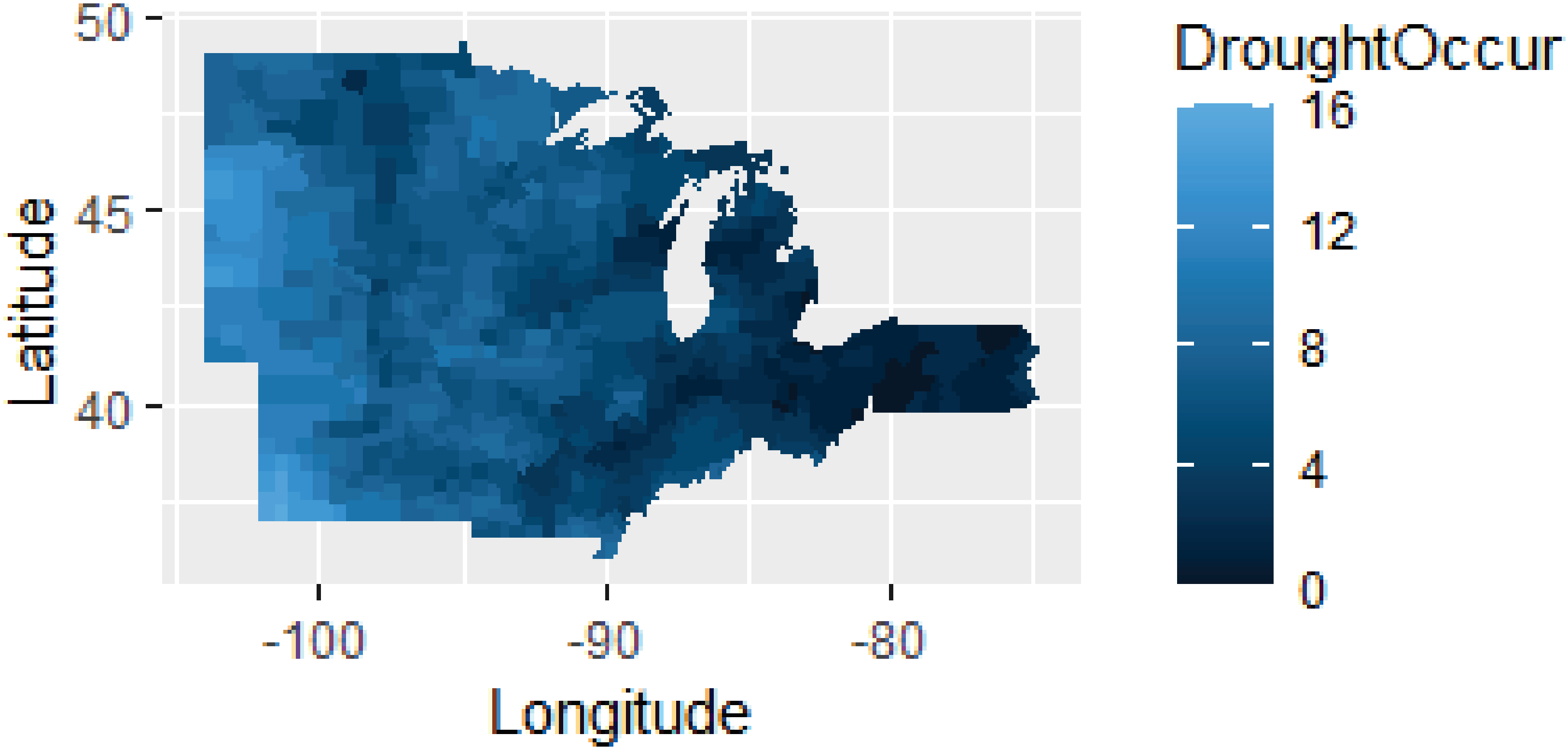


Figure 5: Gross Insured Acres for Corn and Soybeans, By County, from 2000-2021

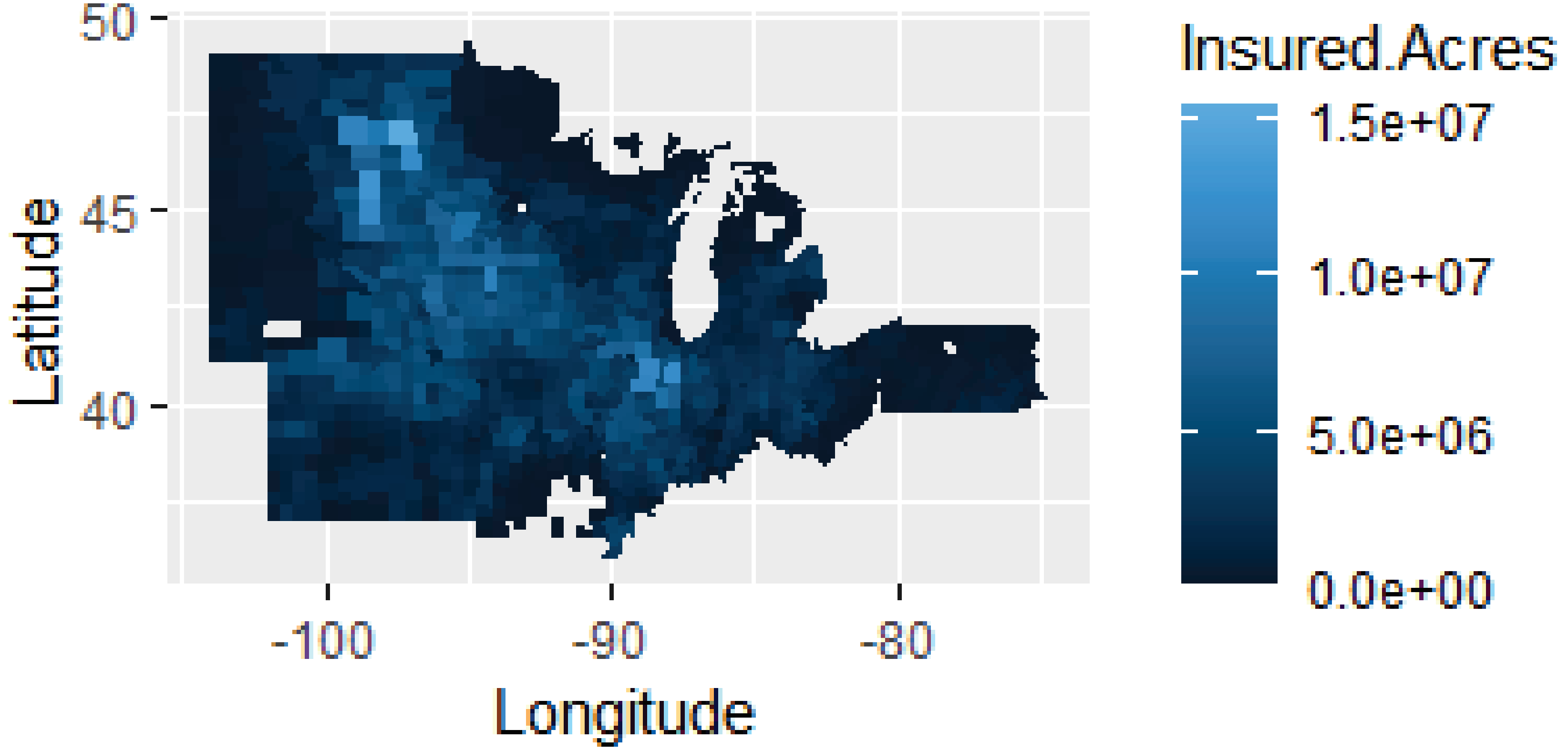


Table 14: Crop Indemnification Beta Regressions, County-Years

Variable	Corn Total		Corn Drought		Soy Total		Soy Drought	
	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio
% Insurance Coverage	1.266***	3.55	-0.039***	0.96	1.237***	3.45	-0.532***	0.59
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated	0.161***	1.17	0.400***	1.49	-0.942***	0.39	-1.081***	0.34
	(0.001)		(0.001)		(0.001)		(0.002)	
Drought	0.411***	1.51	0.700***	2.01	0.295***	1.34	0.510***	1.67
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated * Drought	-2.008***	0.13	-2.466***	0.08	-1.012***	0.36	-1.524***	0.22
	(0.000)		(0.000)		(0.000)		(0.000)	
Successive Drought	-0.056***	0.95	0.062***	1.06	0.019***	1.02	0.032***	1.03
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated * Successive Drought	0.972***	2.64	1.700***	5.47	0.349***	1.42	0.706***	2.03
	(0.000)		(0.000)		(0.001)		(0.001)	
Extreme Drought	0.253***	1.29	0.222***	1.25	0.041***	1.04	-0.009***	0.99
	(0.000)		(0.000)		(0.000)		(0.000)	
County FE	Y		Y		Y		Y	
Year FE	Y		Y		Y		Y	
Weighted By Insured Acres	Y		Y		Y		Y	
N	14929		12258		14117		11716	
Median Deviance Residual	-66.3		-137.7		-58.7		-123.2	
Pseudo R^2	0.40		0.57		0.50		0.56	

Table 15: Crop Indemnification Beta Regressions, Yield Catastrophic Coverage Only

Variable	Corn Total		Corn Drought		Soy Total		Soy Drought	
	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio
Proportion Irrigated	-1.241***	0.29	-0.756***	0.47	-0.638***	0.53	0.200***	1.22
	(0.004)		(0.006)		(0.006)		(0.008)	
Drought	0.371***	1.45	0.567***	1.76	0.278***	1.32	0.521***	1.68
	(0.001)		(0.001)		(0.001)		(0.001)	
Proportion Irrigated * Drought	-0.959***	0.38	-1.263***	0.28	-1.240***	0.29	-1.620***	0.20
	(0.002)		(0.003)		(0.004)		(0.006)	
Successive Drought	0.072***	1.07	0.349***	1.42	0.164***	1.18	0.484***	1.62
	(0.001)		(0.001)		(0.001)		(0.002)	
Proportion Irrigated * Successive Drought	0.468***	1.60	0.254***	1.29	0.279***	0.28	-0.226***	0.80
	(0.003)		(0.003)		(0.005)		(0.006)	
Extreme Drought	0.417***	1.52	0.577***	1.78	0.200***	1.22	0.249***	1.28
	(0.001)		(0.001)		(0.001)		(0.001)	
County FE	Y		Y		Y		Y	
Year FE	Y		Y		Y		Y	
Weighted By Insured Acres	Y		Y		Y		Y	
N	4321		2247		3272		1558	
Median Deviance Residual	-19.4		-19.9		-21.1		-20.8	
Pseudo R^2	0.46		0.62		0.42		0.57	

Table 16: Crop Indemnification Beta Regressions, Yield Buy-Up Coverage Only

Variable	Corn Total		Corn Drought		Soy Total		Soy Drought	
	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio
Proportion Irrigated	-0.483***	0.62	-0.668***	0.51	-0.314***	0.73	-0.643***	0.53
	(0.002)		(0.003)		(0.002)		(0.003)	
Drought	0.329***	1.39	0.524***	1.69	0.462***	1.59	0.627***	1.87
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated * Drought	-1.504***	0.22	-1.799***	0.17	-1.079***	0.34	-1.360***	0.26
	(0.001)		(0.001)		(0.001)		(0.001)	
Successive Drought	0.092***	1.10	0.349***	1.42	-0.120***	0.89	0.098***	1.10
	(0.000)		(0.001)		(0.000)		(0.000)	
Proportion Irrigated * Successive Drought	0.820***	2.27	0.846***	2.33	0.075***	1.08	-0.251***	0.78
	(0.001)		(0.001)		(0.001)		(0.001)	
Extreme Drought	0.306***	1.36	0.284***	1.33	0.157***	1.17	0.084***	1.09
	(0.000)		(0.000)		(0.000)		(0.000)	
County FE	Y		Y		Y		Y	
Year FE	Y		Y		Y		Y	
Weighted By Insured Acres	Y		Y		Y		Y	
N	13162		7867		12261		7398	
Median Deviance Residual	-18.17		-27.0		-16.1		-24.8	
Pseudo R^2	0.41		0.53		0.38		0.51	

Table 17: Crop Indemnification Beta Regressions, Revenue Insurance Pre-2011

Variable	Corn Total		Corn Drought		Soy Total		Soy Drought	
	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio
Proportion Irrigated	-0.764***	0.47	-1.372***	0.25	-0.557***	0.57	-0.752***	0.47
	(0.000)		(0.000)		(0.000)		(0.001)	
Drought	0.459***	1.58	0.595***	1.81	0.312***	1.37	0.441***	1.55
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated * Drought	-0.946***	0.39	-0.883***	0.41	-0.547***	0.58	-0.409***	0.66
	(0.001)		(0.001)		(0.001)		(0.001)	
Successive Drought	-0.042***	0.96	0.028***	1.03	-0.091***	0.91	-0.278***	0.76
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated * Successive Drought	0.715***	2.04	0.590***	1.80	0.009***	1.01	0.303***	1.35
	(0.001)		(0.001)		(0.001)		(0.001)	
Extreme Drought	0.313***	1.37	0.408***	1.50	0.263***	1.30	0.284***	1.33
	(0.000)		(0.000)		(0.000)		(0.000)	
County FE	Y		Y		Y		Y	
Year FE	Y		Y		Y		Y	
Weighted By Insured Acres	Y		Y		Y		Y	
N	8305		6815		7965		6663	
Median Deviance Residual	-33.8		-76.6		-31.9		-70.2	
Pseudo R^2	0.32		0.46		0.46		0.46	

Table 18: Crop Indemnification Beta Regressions, Revenue Insurance 2011 and Onwards

Variable	Corn Total		Corn Drought		Soy Total		Soy Drought	
	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio	Beta	Odds Ratio
Proportion Irrigated	0.075***	1.08	-0.392***	0.68	-0.283***	0.75	-0.695***	0.50
	(0.000)		(0.000)		(0.000)		(0.001)	
Drought	0.526***	1.69	0.858***	2.36	0.240***	1.27	0.510***	1.67
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated * Drought	-1.824***	0.16	-2.180***	0.11	-0.882***	0.41	-1.545***	0.21
	(0.001)		(0.001)		(0.001)		(0.001)	
Successive Drought	-0.147***	0.86	-0.164***	0.85	0.053***	1.05	0.026***	1.03
	(0.000)		(0.000)		(0.000)		(0.000)	
Proportion Irrigated * Successive Drought	-0.031***	0.97	0.510***	1.67	0.265***	1.30	0.702***	2.02
	(0.001)		(0.001)		(0.001)		(0.001)	
Extreme Drought	0.368***	1.44	0.449***	1.57	-0.024***	0.98	0.041***	1.04
	(0.000)		(0.000)		(0.000)		(0.000)	
County FE	Y		Y		Y		Y	
Year FE	Y		Y		Y		Y	
Weighted By Insured Acres	Y		Y		Y		Y	
N	7681		5927		7442		5715	
Median Deviance Residual	-64.3		-182.7		-47.8		-154.2	
Pseudo R^2	0.34		0.48		0.33		0.45	

Table 19: Loss Ratio Regression Results, Corn

Variable	Corn				
	County	Yield Catastrophic	Yield Buy Up	Revenue <=2010	Revenue >=2011
% Insured	1.389***				
	(0.072)				
Proportion Irrigated	0.954***	-0.706***	-0.551***	-0.553***	-0.037
	(0.283)	(0.105)	(0.060)	(0.054)	(0.105)
Drought	0.305***	0.244***	0.383***	0.318***	0.541***
	(0.031)	(0.053)	(0.031)	(0.026)	(0.056)
Proportion Irrigated * Drought	-2.100***	-0.797***	-1.240***	-0.584***	-2.29***
	(0.102)	(0.153)	(0.081)	(0.086)	(0.174)
Successive Drought	-0.099**	-0.172**	-0.010	-0.046	-0.297***
	(0.041)	(0.084)	(0.044)	(0.036)	(0.068)
Proportion Irrigated * Successive Drought	0.912***	0.398**	0.363***	0.232**	0.210
	(0.114)	(0.188)	(0.095)	(0.096)	(0.193)
Extreme Drought	0.574***	0.418***	0.487***	0.295***	0.931***
	(0.037)	(0.066)	(0.037)	(0.032)	(0.063)
Spatial FE	County	State	State	State	State
Year FE	Y	Y	Y	Y	Y
Weighted By Insured Acres	Y	Y	Y	Y	Y
N	15405	14786	16793	8870	8039
Median Deviance Residual	-13.7	-4.9	-8.3	-8.8	-23.8
Adjusted R^2	0.43	0.08	0.30	0.37	0.44

Table 20: Loss Ratio Regression Results, Soybeans

Variable	Soybean				
	County	Yield Catastrophic	Yield Buy Up	Revenue Early	Revenue Late
% Insured	1.138***				
	(0.073)				
Proportion Irrigated	0.16	-0.438***	-0.150**	-0.238***	-0.150**
	(0.029)	(0.093)	(0.069)	(0.063)	(0.069)
Drought	0.232***	0.094***	0.074**	0.284***	0.074**
	(0.021)	(0.034)	(0.030)	(0.024)	(0.030)
Proportion Irrigated * Drought	-0.773***	-0.310**	-0.392***	-0.347***	-0.392***
	(0.094)	(0.152)	(0.130)	(0.115)	(0.130)
Successive Drought	-0.135***	-0.092*	-0.015	-0.137***	-0.015
	(0.028)	(0.053)	(0.037)	(0.033)	(0.037)
Proportion Irrigated * Successive Drought	0.170	0.256	0.033	-0.068	0.033
	(0.111)	(0.198)	(0.153)	(0.130)	(0.153)
Extreme Drought	0.155***	0.103**	0.193***	0.146***	0.193***
	(0.027)	(0.049)	(0.036)	(0.032)	(0.036)
Spatial FE	County	State	State	State	State
Year FE	Y	Y	Y	Y	Y
Weighted By Insured Acres	Y	Y	Y	Y	Y
N	14355	13101	16793	8385	7697
Median Deviance Residual	-14.4	-4.8	-13.2	-7.6	-13.2
Adjusted R^2	0.37	0.09	0.27	0.42	0.27

Table 21: Loss Ratio Regressions, USDCI vs. Disaggregated

Variable	Loss Ratio ~ BX			
	Corn		Soybeans	
D0	-3.05***		-2.30***	
	(0.66)		(0.49)	
D1	5.70***		10.14***	
	(0.84)		(0.64)	
D2	2.70**		5.03***	
	(1.16)		(0.92)	
D3	17.60***		15.94***	
	(1.79)		(1.52)	
D4	7.63***		46.87***	
	(2.62)		(2.89)	
USDCI		1.73***		2.74***
		(0.29)		(0.22)
Omitted Power of 10	*10^-3			
Median Weighted Deviance Residual	-14.9	-14.5	14.4	15.6
Adj R^2	0.43	0.43	0.39	0.37

Chapter 3. Pricing Specialty Crop Insurance with Machine Learning

1. Introduction

The Federal Crop Insurance Program (FCIP) is a large and growing program, paying out over $142 billion in indemnities between 2000 and 2021. The FCIP aims to replace ad-hoc disaster assistance, particularly for widely planted commodities such as Corn or Wheat (Glauber 2004). FCIP premiums are heavily subsidized to achieve the market saturation necessary to replace ad-hoc programs, with 'catastrophic' coverage available for only an administrative fee. In recent decades, the FCIP has greatly expanded, but not all crops have available insurance products and not all insurable crops are insurable everywhere. These holes in insurance coverage risk creating market distortions beyond the usual concerns related to subsidies or insurance; offering crop insurance has both an option value and a subsidy value for takers, theoretically incentivizing farmers away from area-crop combinations without offered FCIP insurance. For these reasons, the USDA Risk Management Agency (RMA) has been instructed to expand the FCIP through various farm bills. While coverage for commodities is not total, the main holes in offered coverage have been and continue to be for 'specialty' crops such as fruit and vegetables.

The FCIP has made considerable progress expanding insurance coverage for specialty crops in recent decades; Figure 6 shows insured specialty crop acreage rising by three-fold from 1989 to the present day. However, this expansion has not been without missteps. Pilot programs for crops such as raspberries and winter squash were terminated during their pilots due to a lack of grower interest, unacceptably high loss ratios, or both (USDA, Risk Management Agency). Unacceptably high loss ratios are a direct consequence of underpricing insurance contracts, while a lack of grower uptake could be due to over-pricing these pilot offerings.

Part of the problem is the inherent difficulty of pricing new insurance products. For pre-existing policies, the RMA adjusts premiums by observing experienced loss ratios. The RMA's own business files are not the only source of data that could guide pricing decisions, but they are the most relevant, granular, and frequent. Market statistics are too aggregated to model crop risk's spatial heterogeneity. The USDA's Census and Surveys of Agriculture respectively run only once in every 5 years or only capture a small subsample. The USDA's premium setting mechanisms do not explicitly estimate demand for crop insurance nor construct a partial equilibrium model. Without these structural formulations, adverse selection can only be accounted for post-hoc. These complications are particularly dire for brand new products, across crops or types of insurance, but still exist when expanding existing products to new areas. FCIP insurance products are not generally available everywhere, and specialty crops have especially pock-marked geographic availability: see Figure 7 for insured specialty crop acreage by county, where dark grey signifies the lack of any specialty crop insurance recorded between 2000 and 2021.

In this essay, I utilize Machine Learning methods to model expected indemnification for Potato crop insurance using basic climate variables and the historical performances of commodities such as Grain Corn or Barley. Potatoes represent a specialty crop with an adequate dataset of insured records within a single crop species so are optimal for a testing demonstration. This essay demonstrates how correlations between crop yields can be exploited to approximate aggregate risk when data is scarce and research budgets prohibit structural agronomic modelling. This algorithm can be used to assist USDA RMA pricing decisions for new insurance products. It could also be used to model the risk of catastrophic loss where information is generally scarce, such as in developing countries. The success of this algorithm contributes a key addition to future models of food system resiliency.

Through careful cross-validation and variable selection the algorithm takes on the position of the RMA expanding an insurance product into new counties. To preview the results, this algorithm estimates indemnification for Potatoes with comparable total accuracy to the USDA's historical performance. This algorithm manages to do so without inter-temporal correcting mechanisms nor any geographic weighting, meaning that it does not rely on training observations near to the target area nor any prior experience with that insurance product in the new counties. This gives the algorithm operational flexibility when producing first-estimates for crop risk in entirely new areas.

2. Review of the Machine Learning Process

Machine Learning is a process of constructing a statistical model to solve some problem. The components of the process include the dataset, the learning algorithm that transforms the dataset and the model, which is the final product that can be used by decision makers.

For example, Ordinary Least Squares linear regression qualifies as a simple Machine Learning process. Here, the researcher decides ex-ante that their desired finished model should be a linear combination of covariates **X** to predict some numerical dependent variable y:

1) $\mathrm{E}(\mathrm{y}|\mathrm{X}) = \hat{\mathrm{y}} = f_{w,b} = \mathbf{b} + \mathbf{w} \cdot \mathbf{X}$

These researchers do not know which covariates **X** should be weighed by what coefficient values **w**, so they apply the linear regression learning algorithm to estimate parameters **w** and **b**. Here, the average square of the 'errors', the differences between y_i and $\mathbf{b} + \mathbf{w} \cdot \mathbf{X}_\mathbf{i}$, serves as the loss function that trains the model. The resulting model is the function $f_{w,b}$with estimated parameters **w** and **b**.

Such a project would encounter many of the same concerns that bedevil far more complicated Machine Learning processes. For example, if characterizing E(y|X) as a linear function is inappropriate for y's data generating process, then the model produced by this regression will not likely be a useful product. Alternatively, researchers may worry about producing a model that seems to achieve acceptable loss function values but fails when exposed to new data. This external validity could fail for multiple reasons, but issues such as 'overfitting' are especially common. The use-case of the model is typically for data that doesn't yet exist, so its information cannot help train the model; researchers attempting to achieve low values on their loss functions by increasing the number of covariates or the complexity of their functional forms may create spurious patterns out of statistical noise. The details of a Machine Learning process, including cross-validating model forms or referencing pre-existing literature, exist to prevent training spurious models.

This project's direct use case is to estimate indemnification for a crop insurance product in a new area. In such a new area, there would be no local historical information on insuring that specialty crop to base premiums on. Instead, the algorithms train the model based on a dataset of historical climate and commodity performance variables that would be frequently available in both the new area and alongside pre-existing performance data for said insurance product from other areas. This demonstration ultimately produces a model estimating expected indemnities from crop insurance for Potatoes. However, the product of this essay is its methodology more so than the parameter estimates produced along the way. The machine learning process in this essay can be reapplied by decision makers to other specialty crops or other areas.

That process begins with constructing a dataset of relevant variables to estimate Potato indemnification risk, including local climates and historical yields for commodity crops. A cross-validation procedure is then used to decide on final model form while avoiding overfitting. Because crop insurance indemnification takes the form of a point density at 0 and a spectrum of observed losses, the model includes two separate regressions, one each to estimate chance of loss and the severity of losses. Finally, the model is trained and tested against observations left out of the training process, imitating the use case.

3. Methodology: Model Forms

A natural description of crop risk is the probability density function of the proportion of acres indemnified; this base description of risk can be used to determine the 'Base Premium Rates' for a policy, after which futures prices, selected coverage levels, and standard subsidy rates would determine the true premium paid by farmers. This definition of crop risk can be naturally modelled by the flexible inflated beta distribution: the beta distribution allows for modelling a variable bounded between (0,1), while the 'inflation' means using jointly estimated probabilities of exactly equaling 0 or 1 to extend the distribution to [0,1].

Beta regressions to estimate these distributions as functions of covariates have been reviewed in the previous essay as well as in the literature (Ospina and Ferrari 2012). Due to the computational burden and interpretation complexity of calculating joint inflated beta Bayesian regressions, I separate those analyses into two separate regressions; I conduct logistic regressions to estimate the percent chance of some loss, and then a non-inflated beta regression estimating the mean and spread of actual losses. Losses equal to 100% of insured acres are rare and I discard them for this analysis.

Alternatively, the severity of Potato losses may be described in terms relative to the USDA's expectations from its pre-existing methodologies. In this case, we use the loss ratios experienced for Potato insurance products as the key dependent variable. To investigate this modelling choice, I use linear regression for non-zero losses, and logistic regression to estimate the chance of loss[13].

[13] Unlike the beta regression, the linear regression I use here could theoretically be used to estimate the entirety of the data, including the point density at 0 losses. However, early exploratory analysis confirmed that there is too great of a point density at 0 losses to rely on linear regression to accomplish both tasks.

The rest of the algorithm training methodology description is split into 3 subsections, each detailing a specific type of regression and how I use them in the cross-validation process and final test.

Logistic Regression

Logistic regression is a classification learning algorithm that estimates by maximum likelihood the influence of covariates on the probability of a 'yes' result from a binary dependent variable. The model fits a conditional sigmoid function:

2) $f_{a,b}(\boldsymbol{X}) = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-(\boldsymbol{a}+\boldsymbol{b}*\boldsymbol{X})}}$

where $\boldsymbol{a}$ + $\boldsymbol{b}$ ***X** is the traditional linear regression. Parameters **a** and **b** are fit to maximize the likelihood, given actual binary data y_i. Equivalently we can maximize the log likelihood:

3) $Log\left(L_{a,b}(\boldsymbol{X})\right) = \sum_{i=1}^{N}[y_i * \ln\left(f_{a,b}(\boldsymbol{X})\right) + (1 - y_i) * \ln\left(1 - f_{a,b}(\boldsymbol{X})\right)]$

I conduct maximum-a-posteriori logistic regressions to further regularize these regressions. This means that instead of taking the **a** and **b** that maximize the likelihood function, I take the coefficients that maximize the posterior function; the posterior function multiplies the likelihood by a prior, which I take as a normal distribution N(0,2).[14] Because this prior is centered on zero, it has the effect of reducing the magnitude of any coefficients estimated, metaphorically similar to ridge regression in the linear context.

[14] Logistic regressions in this essay were estimates using the rstanarm R package by Goodrich et al. (2022). ROC curves were calculated using the pROC package by Robin et al. (2021). Maximum-a-posteriori estimates were calculated using the bayestestR package by Makowski et al. (2019)

For a review of analyzing the performance and validity of logistic regressions, see (Arboretti Giancristofaro and Salmaso 2007). Classification algorithms such as logistic regression may error by creating false negatives (1-'Specificity') or false positives (1 – 'Sensitivity'). For any given fitted logistic model, the selection of the cut-off point of $f_{a,b}(\boldsymbol{X})$ for classification decisions includes a tradeoff between these problems. A Receiver Operating Characteristic (ROC) curve displays the trade off the model encounters: the area under the ROC curve, equivalent to the "C-statistic", marks how well the logistic regression serves as a classifier: the range of values for the C-statistic is [0,1], with 0.5 being equivalent to a random classifier and 0.7 being the minimum accepted standard in the literature for a classifier model. In this way, the area under the ROC curve serves as a metric of logistic model fit even if a decision rule is not literally being instituted. The C-statistics/ areas under the ROC curve for the tested algorithm formulations are shown in Table 24.

Beta Regression

Ospina and Ferrari (2012) demonstrate a general model for beta regression models, including zero inflation. A beta distribution takes parameters μ (mean) and φ (precision) and can be parameterized as:

$$4)\ f(y;\mu,\varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma\left((1-\mu)\varphi\right)} y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1}, y \epsilon\ (0,1)$$

where Γ represents the gamma function. Beta regression parameterizes the above based on maximum likelihood, with μ and φ modeled as conditional on a linear function of covariates. A linear function would allow for invalid values of μ and φ, so link functions h() are used to expand the parameters to the real number line:

5) $h_1(\mu) = \boldsymbol{b_1} \cdot \boldsymbol{X_1}$

6) $h_2(\varphi) = \boldsymbol{b_2} \cdot \boldsymbol{X_2}$

The sets of covariates used for estimating the mean and precision parameters may be the same or distinct: in older literature, such as Ferrari and Cribari-Neto (2004), readers may see φ assumed to be constant. For these analyses[15], $\boldsymbol{X_1}$ includes the full set of covariates, as described in section 5. $\boldsymbol{X_2}$ includes only two variables that clearly influence the 'spread' of plausible losses relative to a degree of risk: the number of policies insured, and the average coverage level of the Potato insurance product bought.

Linear Regression

Linear regression parametrizes coefficients **β** on a direct linear sum of covariates **X** to minimize the sum of squared errors y_i - **β**·**X**. While linear regression is traditionally taught in the frequentist context, its estimation has an immediate bayesian interpretation. For regressions measuring loss experience using observed loss ratios, I conduct maximum-a-posteriori estimation, assuming a prior on each coefficient of N(0,2).

4. Methodology: Cross Validation Procedure

The recommended procedure to avoid overfitting, used across Machine Learning projects (Chen 2021; Burkov 2019; Aboretti Giancristofaro and Salmaso 2007) , is to divide the data into training and testing sub-samples, and rely on cross-validation to make 'researcher decisions' such as hyper-parameter values or which covariates to include. First, I allocate ~10% of the counties featuring observed Potato insurance products to the

[15] The beta regressions used in this essay were estimated using the betareg R package by Cribari-Neto and Zerileis (2010).

testing subsample. I leave records from these counties out of the process until the final testing procedure. Within the remaining ~90%, I conduct ‘leave-one-out’ repeated data-splitting. In each round, eight ninths of the training counties train a model. The model then projects estimates onto the left out ninth, which is called the ‘validation’ sub-sample. For each regression variation that enters the cross-validation process, the produced model fit diagnostics are averaged across 9 runs of this process, each with a distinct and mutually exclusive validation sub-sample.

These results are analyzed alongside research objectives to decide on the final model. As is standard in Machine Learning processes, I look for a combination of good fit onto the validation sub-sample and a minimal loss of fit between the training and validation sub-samples. Minimizing the amount of contemporaneous information necessary to achieve model fit is a key metric of this project’s contribution. While it is possible and useful to describe specialty crop loss in terms of its correlation with other crop losses within-year, it requires less processing and fewer assumptions on the part of the researcher if the process can fit specialty crop loss using only historical information. Therefore, with comparable and acceptable results across model formulations, I select the model with the least contemporaneous information and less information from other USDA beliefs and operations.

I then test that model by re-training it on the totality of the training set, and cross-validating it against the test set. That model’s performance on this test is the final diagnostic for this project. By leaving out the test set until this point, I imitate the position of the USDA expanding the product to the test set for the first time.

The training, validation and testing subsamples should be subdivided to model the use case of the algorithm, so that its validation procedure best models the external validity needed by the use case. The target use case of this algorithm is to expand existing insurance products to new counties. Therefore, I subdivided the data by county: I randomly assign each county a number from (0,100), and I designate those within (91,100) to compose the test set while the rest compose the training set. All yearly observations then follow their counties to their assigned groups.

There is a theoretically infinite quantity of models to search through during the cross-validation phase. I conduct cross-validation across three key dimensions: how many commodities should be included as potential signals, whether indemnification severity should be measured by the percent of acres indemnified or by the recorded loss ratio, and whether commodity loss experience should be measured by contemporaneous loss ratios or historical yield statistics. Adding data from less commonly available commodities and/or balancing loss experience against pre-existing USDA expectations adds potential information to the model at the expense of additional noise and potential for overfitting.

5. Data

From the USDA's Summary of Business public data files, I obtained data on insured acreage, pre-subsidy premiums, and total indemnities by crop, county and year. From the USDA's Cause of Loss data files, I obtained data on acres loss by crop, county, year and the cause of that loss. From the USDA's National Agricultural Statistics Service's Quick Stats query tool, I take survey data on yields for commodity crops. For climate variables, I use data and code from Schlenker and Roberts (2009). This code calculates precipitation and degree days for each county, extrapolated from nearby PRISM data and weighted over known agriculture areas. These variables are calculated over the standard growing season from March through November. For heat variables, I consider degree days over 10°C to represent heat and degree days over 30°C to represent extreme heat; these values are taken using advice from the Commercial Potato Production in North America (2010) handbook, with extreme heat moderated to better represent variation in the data. I also include total precipitation in mm, and the square of that precipitation. Wherever weather variables are included, they are represented not by the direct observation of that county-year, but by historical averages and variances. For each of the above variables for each county-year, I take the mean and variance of that statistic in that county over the previous 20 years.

The selection of Potatoes as the example specialty crop and which associated commodity crops to use as signals are not trivial design choices. The most important criteria for the specialty crop were 1) a large sample size of county-year observations, from itself and with commodities, and 2) evidence of shared loss drivers with commodities. Following criteria (1), candidate specialty crop groups included Berries, Root Bulbs, Legumes, and Tree Fruits including Citrus. Table 22 shows the percentage of indemnities owed to each damage cause group for each crop type. Tree Fruits was thus disqualified due to differences in causes of loss. From the remaining candidates, Potatoes stand out with a large and widely spread quantity of observations within a single crop; Figure 8 shows the counties with recorded insured Potato acres by 2021.

I ordered commodity crop candidates by their quantity of shared county-year tuples with insured Potato acres to select those commodities with the best data overlap. After filtering out dissimilar-in-kind insurance products (e.g. livestock insurance, insured trees), I selected the following eight commodities: Wheat, Corn, Barley, Oats, Soybeans, Sunflowers, Canola and Grain Sorghum To validate the number of commodities to include, I conduct tests while decreasing the number of included commodities, leaving out the commodities with the least overlap in the observed data.

Which commodities are used in each regression depends on whether that regression uses historical yield statistics or contemporaneous loss ratios. Using contemporaneous loss ratios from purchased insurance contracts, the commodities in order of quantity of observed records in the same county-years as insured Potatoes are Wheat, Corn, Barley, Oats, Soybeans, Sunflowers, Canola and Grain Sorghum. However, county-level yield from NASS data is more limited for some commodities. County-year yield tuples only exist up to 2008 for Sunflowers, and they only exist for Canola back to 1999. The available commodities in order of observed historical yield records alongside insured Potatoes are Corn, Wheat, Soybeans, Oats, Barley, then Grain Sorghum. Due to the public insurance data coding all Wheat varieties together, loss ratios on Wheat refer to the total loss ratio for all Wheat products, while Wheat in the yield data refers to the more common winter Wheat.

There are two immediate variables measuring loss experience for any insured crop: % Acres Lost and Loss Ratio. % Acres Lost measures the percent of insured acreage that paid out indemnities due to low yield, while the Loss Ratio is the quantity of total indemnities paid out divided by the calculated pre-subsidy premiums: the USDA metric for a well-priced FCIP is a long-run averages loss ratio of approximately 1. Summary loss statistics for Potatoes and the eight chosen commodities are listed in Table 23. Additionally, crop harvests may be described by their Yields instead of referencing their performance in the FCIP. From the NASS data, I consider the average yield and the variance in yields for a given crop within that county over the past 20 crop-years

All of these metrics are constrained to non-negative values, and insurance loss statistics typically have a right tailed skew. As is standard practice for Machine Learning techniques, I standardize the crop performance and weather control variables (Burkov 2019 ; Chen 2021). For loss ratio variables, I take the natural logarithm of the base data prior to standardization. This makes these variables closer to a 'bell-curve' distribution but would transform 0s into undefined values; these values are set to 0 post-standardization and are marked with a binary variable representing zero indemnification. For logged loss ratios as well as the other variables, I difference observed values from their observed sample means and divide by their sample standard deviations:

1) $X^*_{i,k} = (X_{i,k}-\mu_k) / \sigma_k$

Standardizing these variables aids the learning algorithms to converge on an estimate. It also aids interpretation, as the resulting coefficients compare a standard deviation change in a covariate to a standard deviation change in the dependent variable.

At least some of the eight commodities listed in Table 23 lack insured acreage and/or surveyed yields in most observations alongside insured Potato acreage. There are several methods for dealing with missing data, summarized by Donders et al. (2006) and

Emmanuel et al. (2021). Due to the proliferation of missing values for some commodities, I reject restricting analyses to complete cases or imputing by regression; instead, post-standardization missing values are replaced with 0s. These values thus have no interaction with coefficients. Because these variables are standardized, this effectively imputes missing values by sample means.

Except where noted otherwise, the set of covariates used for each regression in this essay includes degree days for 10°C and 30°C, as well as precipitation (mm) and its square. These variables take the form of the mean and variance of these statistics in that county over the previous 20 years. These also include the number of Potato policies sold and the weighted average coverage level of sold Potato policies for that crop-year in that county. Each regression also includes a set of variables describing commodity crop performance. These include either the commodity loss ratios and a binary 1(Zero Indemnification) or the historical means and variances of the commodities' yields. Except for the binary variable, all above covariates are standardized, and missing values are imputed as described above.

Finally, I include two variables describing the policies bought to make up these observations of Potato indemnification: the number of policies bought and the weighted average coverage level of Potato policies. Theoretically, these variables would not be known to the USDA in the initial price-setting phase; however, the former should be relatively easy to estimate based on known local plantings, and the latter is self-controlling by price differences between coverage levels being pre-set by standing policy. Moreover, these are critical variables describing the data generating process. For example, for any given level of risk to each policy, the chance that at least one policy earned indemnities increases with the number of policies that were sold.

6. Cross Validation Results

Table 24 displays the results for the repeated data-splitting for all three regression types, across variations. These variations include the choice of commodity variable set and the use of historical yields or contemporaneous loss ratio. The reader should be aware that while the fit of both the beta and linear regressions are judged by Mean Squared Error, their dependent variables are on different ranges: the former within (0,1), the latter is standardized. Furthermore, for the beta and linear regressions the MSE should be minimized, while for the logistic regressions the C-statistic should be maximized.

The most immediate result from these validation runs is that the commodities with the fewest shared data tuples with Potatoes are not substantially contributing to the fit of these regressions. In some cases, this seed resulted in the more parsimonious models creating a *better* model fit than the full set of commodity variables. As such, the clear result is to conduct the final model using only the commodities with the greatest overlap in insurance availability with Potatoes.

Which metric of commodity performances better signaled the performance of Potato insurance depends on the type of regression. For the logistic regression, the historical yields slightly overperformed the modern loss ratios, while for the linear regressions the modern loss ratios resulted in slightly smaller MSE than the historical yields. However, these differences are small. Therefore, I choose the historical yields data, as it eases the model's use and widens the process's possible use cases.

Finally, I must choose between the linear and beta regressions to predict the severity of Potato losses. To judge between the linear and beta regressions, I transform the average MSE's so the two regressions' metrics are on equivalent scales. The linear system measures error against a standardized variable where an absolute difference of 1 equals the standard deviation of the logged loss ratio ~ 1.34: ~0.9 of these standard deviations implies a difference between logs of actual and fitted of approximately 1.2, or a MSE in magnitude of 120%! The conditional average proportion of acres lost for Potatoes was ~22 percentage points, so an absolute MSE of 4 percentage points at the mean is a MSE in magnitude of 18-19 percentage points. Clearly, the beta regression system results in smaller errors.

Therefore, the results of these cross-validations imply that the final model format shall be a zero-inflated beta regression system. I include historical yields information for winter Wheat, Grain Corn, Oats, and Soybeans, the commodities with the most overlap between their observed yields data and insured Potato acres.

7. Testing the Final Model

The above process selected a model form that achieved external validity across nine runs, within each the training observations had to train a model for a unique validation set. From this, I conclude that a model using historical commodity yields, beta regression, and the four commodities with the greatest data overlap will achieve reasonable fit without succumbing to 'overfitting'. In this section, that model form is tested against data that was reserved until now. I use this testing as a proxy for how this process will perform on future data; I do this by calculating the ‘loss ratio’ generated by the model over the test set. Specifically, I let the model estimate expected indemnification for the test set; assuming that the test set policies were priced based on these expected indemnities, I take the sum of those expected indemnities as the Total Premiums assigned. The actual observed losses are taken as is, and the ratio between the actual losses and the sum of expected losses becomes the Loss Ratio for the model and for this process that generated it. Like the USDA, I aim for a Loss Ratio of 1, which would signify an actuarially fair program. The metric for this process’s credibility is the distance between the model’s tested Loss Ratio and that target of 1.

I parametrize the models based on the totality of the training set, and project their results onto the testing set. The coefficients from the totality of the training set represent the final parameterization of the model and will be discussed in the proceeding section. The external validity of the model is judged from how well the projection fits the testing set. Finally, I conduct a back-of-the-envelope estimate of how well this model predicts indemnities, and therefore how promising it is as a tool to set premium rates.

Projected probabilities and mean severity of losses are assigned to observations in the test set, and expected indemnities are calculated as so:

7) E(Indemnities) = E(Probability of Loss) * E(Severity of Loss) *

Weighted Average Coverage Level * Liabilities

By summing these expected indemnities over the test set and dividing by the actual indemnities, I derive the proportional error in premium setting for this model.

Previewing those results, I find a magnitude of error similar to historical USDA performance for Potatoes. However, I achieved this baseline completely out of geographic sample, without reliance on updating cycles nor within-state weights like the standard USDA cycle, making this model less resource intensive to train than the USDA's current methods.

8. Final Model Results

Table 25 and 26 present the final models' coefficients for commodity yields and climate information respectively.

Figure 9 displays the ROC curves from both the training set and the test set for the final logistic regression. This figure shows that within-training set, the model classifies risky observations quite well, achieving a C-statistic of over 0.8. The fit onto the test set is poorer, with the C-statistic falling to just below 0.7, the traditional bound for acceptable discrimination (Hosmer and Lemeshow 2000; Arboretti Giancristofaro and Salmaso 2007). However, the logistic regression performs far better than a random classifier.

Figure 10 shows the density of predicted loss severities from the beta regression plotted against the true observed values of % acres indemnified. The black diagonal line represents y=x, and is the target for the model. We can see that there are no significant densities far from this line, which repeats the good MSE statistics from the cross-validation tests. The model tends to slightly overestimate losses when actual losses are small and underestimates the severity of large losses. In other words, the true data has a more positive skew than the beta regression's best fit

The coefficients shown in Tables 25 and 26 show high statistical significance across explanatory variables for both the logistic and beta regressions. Average precipitation emerges as a key climate variable with surprising signs on itself and its square. Historically dry and historically wet areas both generate low estimated losses and decreased chance of loss, while the counties within the middle of the sample had the worst loss experience. This could be due to more extreme climates resulting in Potato plantings only alongside crop practices prepared for that extreme, or similar forces influencing selection into the sample.

Among the selected commodities, Winter Wheat Yield shows the lowest associated changes in Potato crop risk, possibly due to the lack of signaling power of non-growing season months on Potato losses. Larger signaling power is found in average Oats Yield, for which a standard deviation increases signals an 89% greater increased chance of loss and a 25% increase in loss severities. Meanwhile, larger Corn Yield signals a decreased chance of any Potato loss, while larger Soybean Yield signals lower severities of Potato losses. Historical yield variances for commodities do not generally give signals of the same magnitude, but historically high variance in Soybean yields signals a lowered chance of Potato losses and a lower severity of average losses of nearly 25%.

Assuming that premiums are set to equal predicted losses, the model's predicted losses in the test set yielded an effective loss-ratio of 0.58. The model greatly overestimates losses in its current form. However, this issue is also apparent in the USDA's historical performance with Potatoes, which was 0.69 within the test sample and 0.64 for the entire sample.

9. Conclusion

The goal of this essay is to assist in pricing specialty crop insurance in new areas, where neither within-area nor region weights are available for that specialty crop. The Machine Learning process I developed uses historical commodity yields and climate data to generate baseline expectations for indemnification for that specialty crop.

The results show a capacity to fit risk profiles that is comparable to the USDA's historical performance. However, this model has two primary advantages against the USDA's record. First, it is decidedly low cost to initiate, with little ad-hoc decision making. Second, the model is operating under stricter conditions than the USDA benchmark. The USDA's pricing process over the test set utilized State weights and an updating cycle. The former requires within-state observations of the same crop to generate State risk statistics, while this model does not. The latter gave the USDA pricing regime additional chances to update county risk profiles over time using loss experience, while my model only allows for marginal shifts in predicted losses as historical yield/climate information updates.

Some of the usual challenges when judging crop risk remain present in this model’s performance. For example, the right-tail of actual losses that bedevils fair pricing is under-modeled by this algorithm. However, this procedure demonstrates an efficient and objective method to generate baseline risk profiles in scenarios where no local data is available for the crop in question.

Figure 6: Specialty Crop Acreage Insured under the FCIP

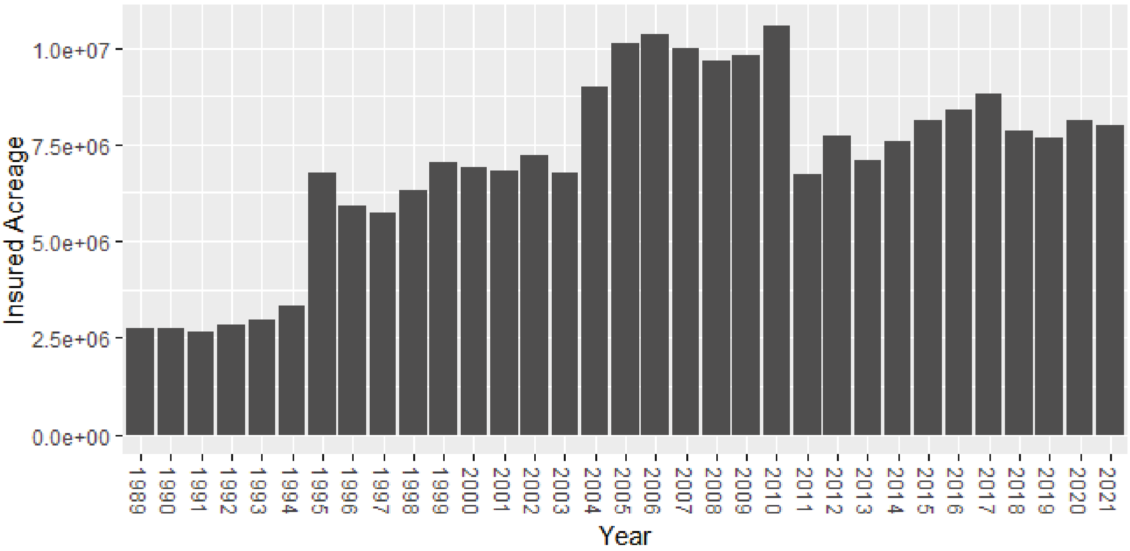


Figure 7: % of FCIP Liabilities to Specialty Crops by County, 2000-2021

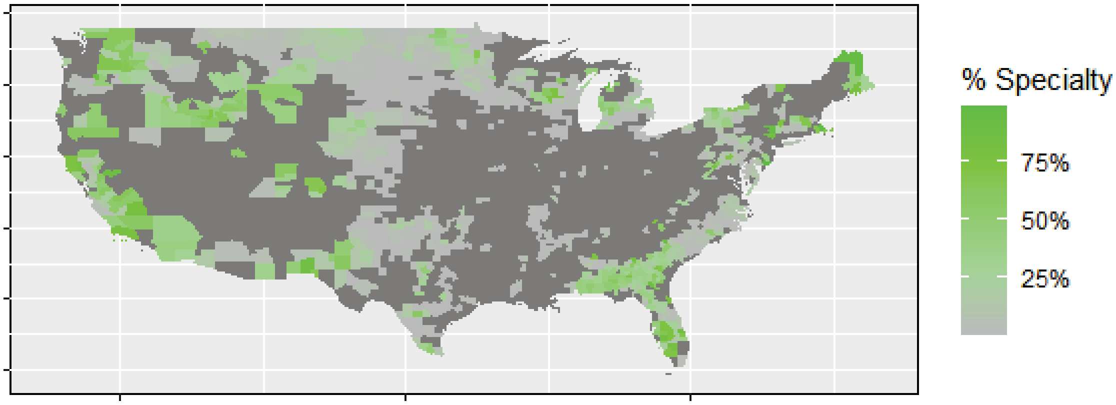


Table 22: Causes of Loss Among Commodities and Specialty Crop Groups

% of Indemnities to...	Tree Fruit	Potatoes/Root Bulbs	Berries	Legumes	Commodities
Drought	0.17	0.20	0.07	0.43	0.43
Flood	0.00	0.01	0.00	0.00	0.01
Hurricane	0.09	0.00	0.00	0.02	0.01
Fire	0.00	0.00	0.00	0.00	0.00
Wind	0.03	0.02	0.03	0.01	0.02
Rain/Humidity	0.14	0.47	0.14	0.32	0.26
Tornadoes	0.00	0.00	0.00	0.00	0.00
Winter	0.41	0.16	0.50	0.06	0.05
Disease/Pests	0.01	0.09	0.03	0.02	0.01
Hail	0.12	0.04	0.09	0.10	0.06
Other	0.03	0.02	0.12	0.03	0.14

Figure 8: Continental U.S. Counties with Insured Potato Acreage

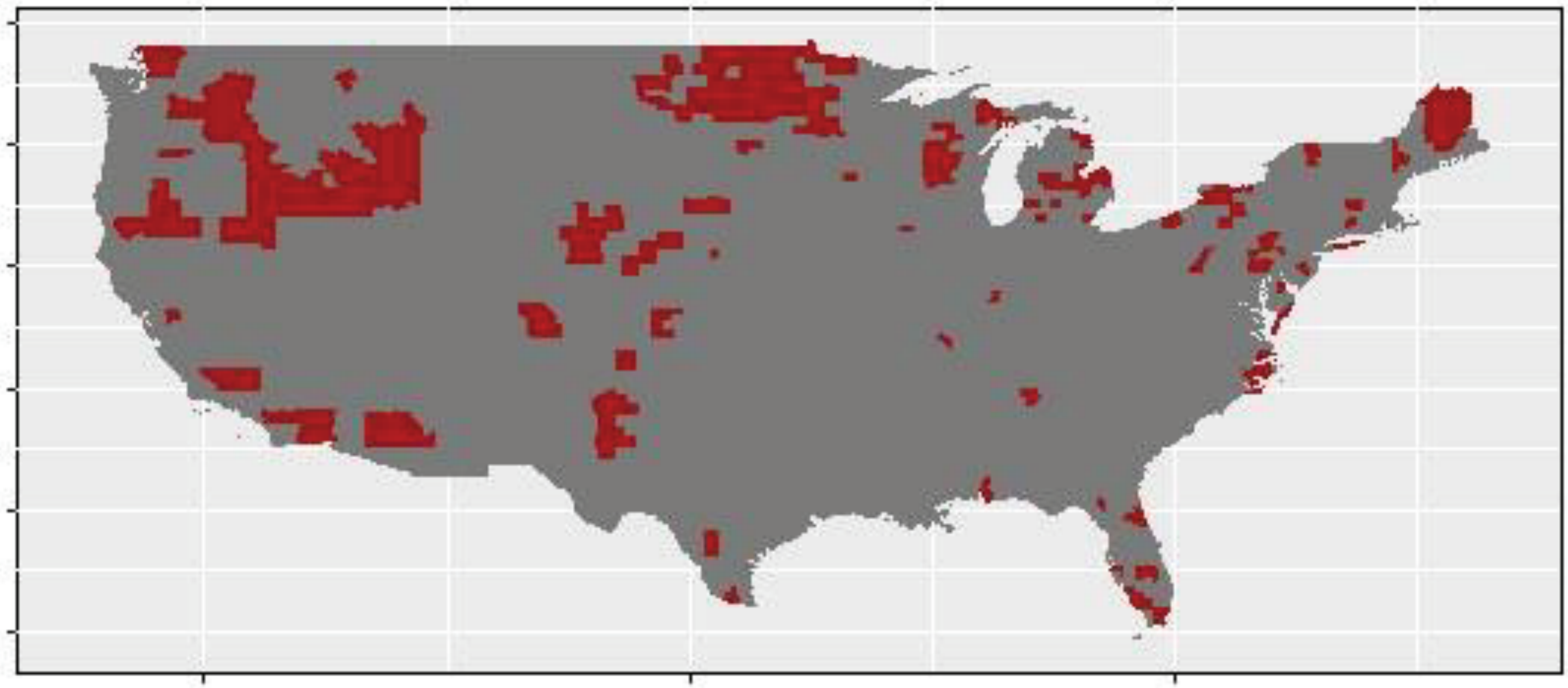


Table 23: Average Loss Experience Summary Statistics, Crops Used in this Essay

Variable	Potatoes	Wheat	Corn	Barley	Oats	Soybeans	Sunflowers	Canola	Grain Sorghum
Total Loss Ratio	0.641	0.843	0.771	0.877	0.93	0.653	1.177	0.869	0.9
% Acres Loss to Drought	0.025	0.116	0.058	0.052	0.131	0.052	0.067	0.04	0.31
% Acres Lost to Hail	0.004	0.022	0.02	0.021	0.017	0.019	0.027	0.03	0.016
% Acres Lost to Floods	0	0	0.001	0	0	0.001	0	0	0
% Acres Lost to Hurricanes	0	0	0.001	0	0	0.001	0	0	0.001
% Acres Lost to Fire	0	0	0	0	0	0	0.004	0	0
% Acres Lost to Wind	0.001	0.008	0.003	0.003	0.004	0.001	0.011	0.017	0.012
% Acres Lost to Excessive Precipitation/Humidity	0.041	0.062	0.067	0.065	0.081	0.08	0.209	0.157	0.024
% Acres Lost to Tornadoes	0	0	0	0	0	0	0	0	0
% Acres Lost to other Winter Weather	0.01	0.026	0.017	0.012	0.012	0.03	0.024	0.029	0.016
% Acres Lost to Disease and Pests	0.007	0.004	0.002	0.003	0.002	0.001	0.023	0.008	0.001
% Acres Lost to Other Causes	0.001	0.017	0.03	0.004	0.001	0.026	0.022	0.006	0.033

Table 24: Cross-Validation Results

Round	Full	Less Sorghum	& Less Canola	& Less Sunflowers
1	0.860	0.860	0.860	0.860
2	0.860	0.860	0.860	0.860
3	0.800	0.800	0.800	0.800
4	0.840	0.840	0.840	0.850
5	0.850	0.850	0.860	0.860
6	0.850	0.850	0.850	0.850
7	0.880	0.880	0.880	0.880
8	0.860	0.860	0.860	0.860
9	0.850	0.850	0.850	0.850
Average	0.850	0.850	0.851	0.852

Linear, Modern Loss Ratios, Mean Squared Error, Within Training Set

Round	Full	Less Sorghum	& Less Canola	& Less Sunflowers
1	0.830	0.840	0.840	0.840
2	0.850	0.840	0.850	0.850
3	1.270	1.270	1.290	1.270
4	0.970	0.980	0.960	0.970
5	0.890	0.900	0.890	0.880
6	0.900	0.910	0.910	0.890
7	0.590	0.590	0.600	0.590
8	0.830	0.820	0.810	0.820
9	0.960	0.960	0.980	0.970
Average	0.899	0.901	0.903	0.898

Linear, Modern Loss Ratios, Mean Squared Error, Projected onto Validation Set

Continued

Table 24: Continued

Round	Full	Less Sorghum	& Less Barley
1	0.860	0.860	0.860
2	0.860	0.870	0.870
3	0.800	0.800	0.800
4	0.840	0.840	0.840
5	0.850	0.850	0.860
6	0.850	0.860	0.860
7	0.880	0.880	0.880
8	0.870	0.870	0.870
9	0.870	0.870	0.870
Average	0.853	0.856	0.857

Linear, Historical Yields, Mean Squared Error, Within Training Set

Round	Full	Less Sorghum	& Less Barley
1	0.880	0.870	0.880
2	0.830	0.810	0.810
3	1.340	1.350	1.390
4	1.050	1.060	0.990
5	0.950	0.950	0.950
6	0.920	0.930	0.920
7	0.630	0.610	0.610
8	0.780	0.770	0.770
9	0.840	0.870	0.860
Average	0.913	0.913	0.909

Linear, Historical Yields, Mean Squared Error, Projected onto Validation Set

Continued

Table 24: Continued

Round	Full	Less Sorghum	& Less Canola	& Less Sunflowers
1	0.038	0.038	0.038	0.039
2	0.039	0.039	0.039	0.040
3	0.035	0.035	0.035	0.036
4	0.037	0.037	0.037	0.037
5	0.036	0.037	0.036	0.037
6	0.039	0.039	0.039	0.039
7	0.037	0.037	0.037	0.037
8	0.037	0.037	0.037	0.038
9	0.038	0.038	0.037	0.038
Average	0.037	0.037	0.037	0.038

Beta, Modern Loss Ratios, Mean Squared Error, Within Training Set

Round	Full	Less Sorghum	& Less Canola	& Less Sunflowers
1	0.021	0.021	0.021	0.021
2	0.024	0.024	0.024	0.025
3	0.061	0.061	0.060	0.059
4	0.043	0.043	0.043	0.044
5	0.052	0.052	0.051	0.052
6	0.032	0.032	0.032	0.033
7	0.039	0.039	0.038	0.040
8	0.043	0.044	0.043	0.042
9	0.034	0.034	0.035	0.033
Average	0.039	0.039	0.039	0.039

Beta, Modern Loss Ratios, Mean Squared Error, Projected onto Validation Set

Continued

Table 24: Continued

Round	Full	Less Sorghum	& Less Barley
1	0.038	0.038	0.038
2	0.037	0.037	0.037
3	0.035	0.036	0.036
4	0.036	0.037	0.037
5	0.036	0.036	0.036
6	0.038	0.038	0.039
7	0.036	0.037	0.036
8	0.037	0.037	0.037
9	0.037	0.037	0.037
Average	0.037	0.037	0.037

Beta, Historical Yields, Mean Squared Error, Within Training Set

Round	Full	Less Sorghum	& Less Barley
1	0.026	0.025	0.026
2	0.029	0.029	0.029
3	0.051	0.052	0.051
4	0.041	0.041	0.041
5	0.048	0.048	0.048
6	0.030	0.030	0.030
7	0.041	0.040	0.041
8	0.047	0.049	0.048
9	0.035	0.037	0.036
Average	0.039	0.039	0.039

Beta, Historical Yields, Mean Squared Error, Projected onto Validation Set

Continued

Table 24: Continued

Round	Full	Less Sorghum	& Less Canola	& Less Sunflowers
1	81.8	81.7	81.6	81.7
2	81.4	81.3	81.3	81.5
3	82.0	82.0	82.0	82.0
4	81.6	81.5	81.6	81.7
5	81.8	81.6	81.7	81.8
6	82.0	81.9	81.9	82.0
7	82.9	82.7	82.7	82.8
8	82.4	82.3	82.3	82.4
9	81.6	81.5	81.5	81.6
Average	81.9	81.8	81.8	81.9

Logistic, Modern Loss Ratios, C Statistics, Within Training Set

Round	Full	Less Sorghum	& Less Canola	& Less Sunflowers
1	83.7	84.1	84.0	84.4
2	86.6	86.6	86.6	86.4
3	80.2	79.9	79.9	80.3
4	81.9	81.8	81.7	82.0
5	82.4	82.4	82.9	83.1
6	77.0	76.8	76.8	75.8
7	64.8	64.8	64.9	64.7
8	78.0	78.1	78.2	78.3
9	76.8	76.7	76.6	76.1
Average	79.0	79.0	79.1	79.0

Logistic, Modern Loss Ratios, C Statistics, Projected onto Validation Set

Continued

Table 24: Continued

Round	Full	Less Sorghum	& Less Barley
1	82.3	82.2	82.2
2	81.8	81.7	81.7
3	82.4	82.3	82.2
4	81.9	81.8	81.8
5	81.9	81.7	81.7
6	82.2	82.2	82.1
7	83.3	83.2	83.1
8	82.6	82.5	82.5
9	81.8	81.6	81.6
Average	82.2	82.1	82.1

Logistic, Historical Yields, C Statistics, Within Training Set

Round	Full	Less Sorghum	& Less Barley
1	83.2	83.0	83.2
2	86.3	86.3	86.2
3	80.2	80.7	81.0
4	81.3	81.1	81.1
5	83.7	83.7	83.8
6	75.0	74.3	74.7
7	67.5	66.6	66.7
8	77.6	77.7	77.6
9	78.9	79.0	79.2
Average	79.3	79.2	79.3

Logistic, Historical Yields, C Statistics, Projected onto Validation Set

Figure 9: ROC Curves, Final Training Set (Blue) and Test Set (Red)

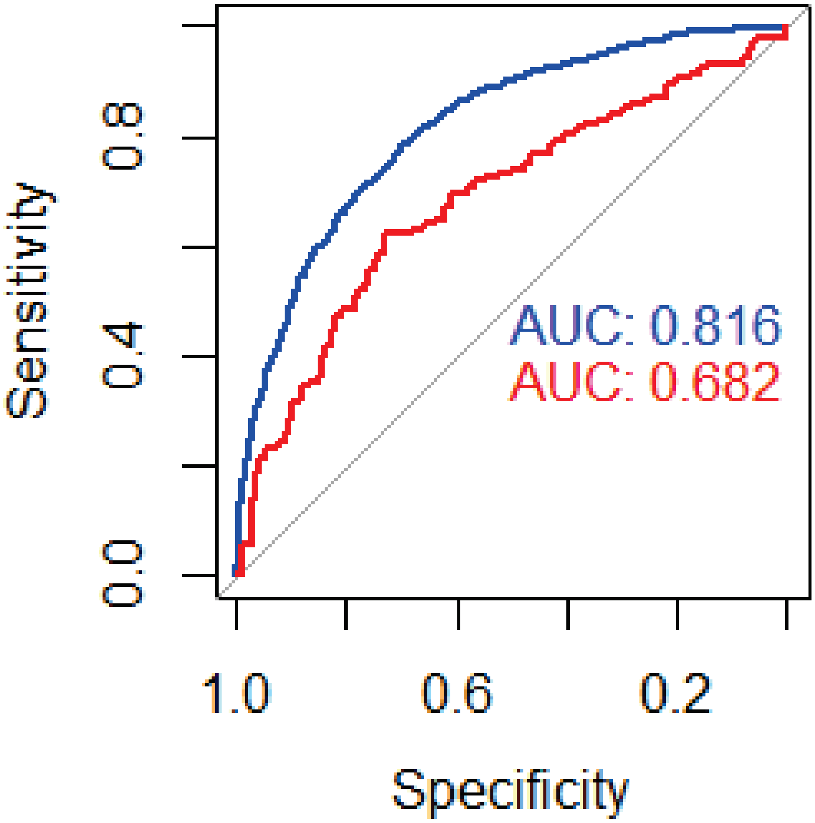


Figure 10: Residuals against Observed Values, Final Beta Regression

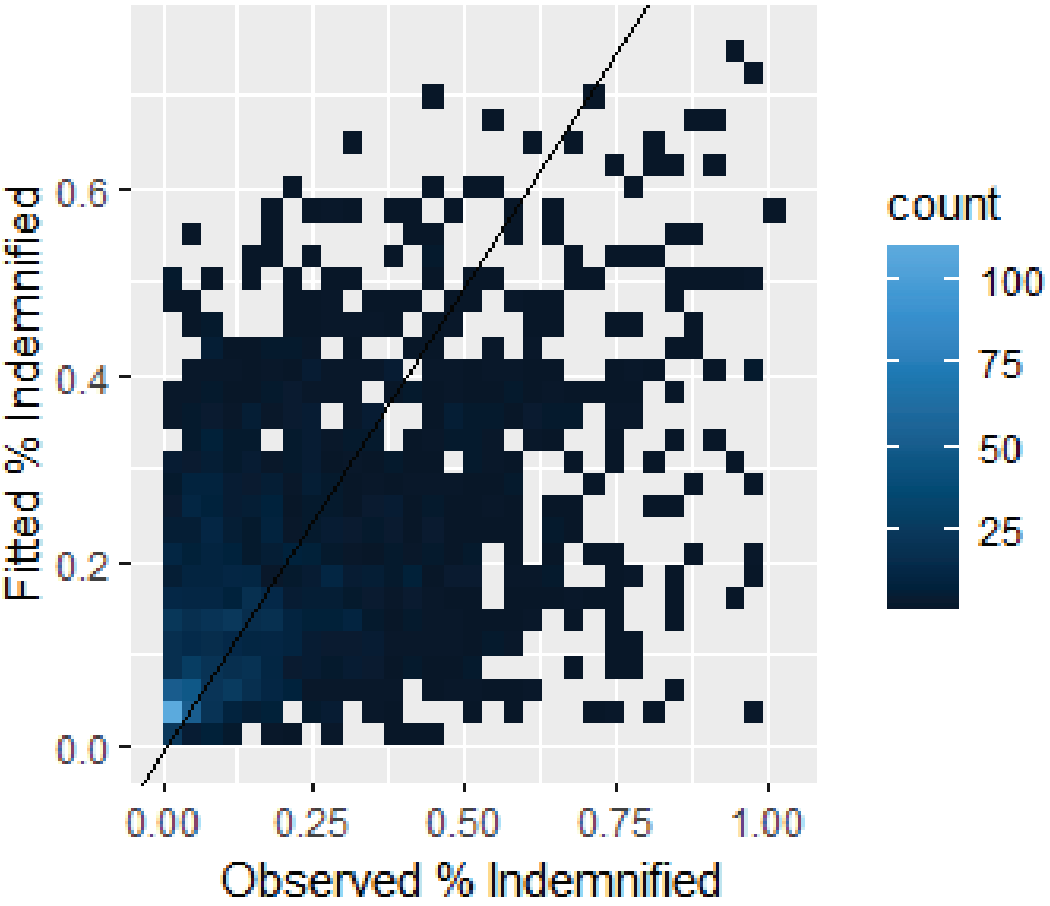


Table 25: Final Model Commodity Coefficients

Variable	Logistic Coefficients	Odds Ratio	Beta Coefficients	Odds Ratio
Wheat, Yield				
Mean	-0.012***	0.988	-0.105***	0.900
	(0.001)		(0.000)	
Variance	-0.111***	0.895	-0.105***	0.900
	(0.001)		(0.000)	
Oats, Yield				
Mean	0.636***	1.889	0.219***	1.245
	(0.002)		(0.000)	
Variance	-0.161***	0.851	0.011***	1.011
	(0.001)		(0.000)	
Corn Yield				
Mean	-0.471***	0.624	0.022***	1.022
	(0.001)		(0.000)	
Variance	0.194***	1.214	-0.009***	0.991
	(0.001)		(0.000)	
Soybeans Yield				
Mean	0.133***	1.142	-0.476***	0.621
	(0.002)		(0.001)	
Variance	-0.263***	0.769	-0.240***	0.787
	(0.002)		(0.001)	
Model Fit	Mean PPD		Pseudo R Squared	
	0.6		0.26	

Table 26: Final Model Weather Coefficients

Variable	Logistic Coefficients	Odds Ratio	Beta Coefficients	Odds Ratio
Precipitation				
Mean	2.429***	11.348	0.845***	2.328
	(0.006)		(0.002)	
Variance	-0.503***	0.605	-0.368***	0.692
	(0.006)		(0.002)	
Precipitation Squared				
Mean	-2.155***	0.116	-0.596***	0.551
	(0.009)		(0.003)	
Variance	1.034***	2.812	0.432***	1.540
	(0.008)		(0.002)	
10°C Degree Days				
Mean	0.246***	1.279	-0.147***	0.863
	(0.003)		(0.001)	
Variance	-0.694***	0.500	-0.148***	0.862
	(0.003)		(0.001)	
30°C Degree Days				
Mean	-0.115***	0.891	0.172***	1.188
	(0.002)		(0.001)	
Variance	0.064***	1.066	0.144***	1.155
	(0.003)		(0.001)	
Model Fit	Mean PPD		Pseudo R Squared	
	0.6		0.26	

Conclusion

In the first essay of this dissertation, I show that designation as a Scenic River is attractive to potential buyers. In the second and third essays, I tackle market distortions in the Federal Crop Insurance Program: first by identifying successive droughts as an opportunity for adverse selection, and second by demonstrating an algorithm to expedite the pricing process to offer all crops equal access to subsidized insurance.

The OSRP and the FCIP are both public policies attempting to fulfill societal needs where private markets fail to exist. It is not reasonable to expect atomized individuals to manage and protect riparian corridors, even though demand for that preserved green space is substantial. Similarly, as Duncan and Myers (1997) explain, correlated losses prohibit private insurance solutions for most crop risks.

However, these programs encounter their own challenges and ‘market failures’. Local communities voting to designate a Scenic River are betting that the regulations on public construction will not deprive them of critical future infrastructure. How well the FCIP reaches its stated goal of actuarial fairness is only the simplest metric for how such a program can distort United States agriculture. A poorly designed public crop insurance program may still offer farms financial security, but risks sacrificing food system resiliency to do so. Even after the mission statements are written and the tradeoffs summarized, the details of these programs face decision makers with an array of uncertainties and complications.

These essays contribute answers to some of those complications. By estimating the capitalization effect of designation on nearby homes, my first essay both contributes to the field’s understanding of the market value of regulatory protection and shows to local governments that designating a local scenic river is valuable to potential neighbors. My second essay identifies successive droughts as a source of FCIP inefficiencies as well as showing the importance of considering the phenomenon when analyzing crop risk across time series. Finally, my third essay offers a low-cost method to deal with missing data, as is common for specialty crops and in developing economies; using other crops as risk signals and careful cross-validation the model can estimate crop risk for a crop in an entirely new area, where measurements are not or cannot be taken.

References

Abadie, Alberto, and Guido W. Imbens. 2006. “Large Sample Properties of Matching Estimators for Average Treatment Effects.” *Econometrica* 74 (1): 235–67. https://doi.org/10.1111/j.1468-0262.2006.00655.x.

———. 2011. “Bias-Corrected Matching Estimators for Average Treatment Effects.” *Journal of Business & Economic Statistics* 29 (1): 1–11. https://doi.org/10.1198/jbes.2009.07333.

Annan, Francis, and Wolfram Schlenker. 2015. “Federal Crop Insurance and the Disincentive to Adapt to Extreme Heat.” *American Economic Review* 105 (5): 262–66. https://doi.org/10.1257/aer.p20151031.

Bernhardt, Emily S., and Margaret A. Palmer. 2007. “Restoring Streams in an Urbanizing World.” *Freshwater Biology* 52 (4): 738–51. https://doi.org/10.1111/j.1365-2427.2006.01718.x.

Bowker, JM, and JoHN C Bergstrom. 2017. “Wild and Scenic Rivers: An Economic Perspective.” *International Journal of Wilderness* 23 (2): 22–33.

Bundy, Logan R., Vittorio A. Gensini, and Mark S. Russo. 2022. “Insured Corn Losses in the U.S. from Weather and Climate Perils.” *Journal of Applied*

Meteorology and Climatology, May. https://doi.org/10.1175/JAMC-D-21-0245.1.

Burkov, Andriy. 2019. *The Hundred-Page Machine Learning Book*. Polen: Andriy Burkov.

Çakir, Recep. 2004. “Effect of Water Stress at Different Development Stages on Vegetative and Reproductive Growth of Corn.” *Field Crops Research* 89 (1): 1–16. https://doi.org/10.1016/j.fcr.2004.01.005.

Caparas, Monica, Zachary Zobel, Andrea D A Castanho, and Christopher R Schwalm. 2021. “Increasing Risks of Crop Failure and Water Scarcity in Global Breadbaskets by 2030.” *Environmental Research Letters* 16 (10): 104013. https://doi.org/10.1088/1748-9326/ac22c1.

Chen, James Ming. 2021. “An Introduction to Machine Learning for Panel Data.” *International Advances in Economic Research* 27 (1): 1–16. https://doi.org/10.1007/s11294-021-09815-6.

Cohen, Itay, Sara I. Zandalinas, Clayton Huck, Felix B. Fritschi, and Ron Mittler. 2021. “Meta-analysis of Drought and Heat Stress Combination Impact on Crop Yield and Yield Components.” *Physiologia Plantarum* 171 (1): 66–76. https://doi.org/10.1111/ppl.13203.

Cribari-Neto F, Zeileis A (2010). “Beta Regression in R.” *Journal of Statistical Software*, **34**(2), 1–24. doi:10.18637/jss.v034.i02.

Douma, Jacob C., and James T. Weedon. 2019. “Analysing Continuous Proportions in Ecology and Evolution: A Practical Introduction to Beta and Dirichlet Regression." Edited by David Warton. *Methods in Ecology and Evolution* 10 (9): 1412–30. https://doi.org/10.1111/2041-210X.13234.

Duncan, John, and Robert J. Myers. 2000. "Crop Insurance under Catastrophic Risk." *American Journal of Agricultural Economics* 82 (4): 842–55. https://doi.org/10.1111/0002-9092.00085.

Ferrari, Silvia, and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7): 799–815. https://doi.org/10.1080/0266476042000214501.

Ganguli, Poulomi, and Auroop R. Ganguly. 2016. "Space-Time Trends in U.S. Meteorological Droughts." *Journal of Hydrology: Regional Studies* 8 (December): 235–59. https://doi.org/10.1016/j.ejrh.2016.09.004.

Glauber, Joseph W. 2004. "Crop Insurance Reconsidered." *American Journal of Agricultural Economics* 86 (5): 1179–95. https://doi.org/10.1111/j.0002-9092.2004.00663.x.

Goodrich B, Gabry J, Ali I, Brilleman S (2022). "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.21.3, https://mc-stan.org/rstanarm/.

Henckel, P A. 1964. "Physiology of Plants Under Drought." *Annual Review of Plant Physiology* 15 (1): 363–86. https://doi.org/10.1146/annurev.pp.15.060164.002051.

Hupp, Cliff R, Gregory B Noe, Edward R Schenk, and Adam J Benthem. 2013. "Recent and Historic Sediment Dynamics along Difficult Run, a Suburban Virginia Piedmont Stream." *Geomorphology* 180: 156–69.

Keith, John, Paul Jakus, and Jacoba Larsen. 2008. "Impacts of Wild and Scenic River Designation." Utah State University.

Ker, Alan P., and McGowan, Pat. 2000. "Weather-Based Adverse Selection and the U.S. Crop Insurance Program: The Private Insurance Company Perspective," 26.

Kieschnick, Robert, and B D McCullough. 2003. "Regression Analysis of Variates Observed on (0, 1): Percentages, Proportions and Fractions." *Statistical Modelling* 3 (3): 193–213. https://doi.org/10.1191/1471082X03st053oa.

Kuminoff, Nicolai V., Christopher F. Parmeter, and Jaren C. Pope. 2010. "Which Hedonic Models Can We Trust to Recover the Marginal Willingness to Pay for Environmental Amenities?" *Journal of Environmental Economics and Management* 60 (3): 145–60. https://doi.org/10.1016/j.jeem.2010.06.001.

Kuminoff, Nicolai V., and Jaren C. Pope. 2014. "DO 'CAPITALIZATION EFFECTS' FOR PUBLIC GOODS REVEAL THE PUBLIC'S WILLINGNESS TO PAY?: DO CAPITALIZATION EFFECTS REVEAL WTP?" *International Economic Review* 55 (4): 1227–50. https://doi.org/10.1111/iere.12088.

Kuwayama, Yusuke, Alexandra Thompson, Richard Bernknopf, Benjamin Zaitchik, and Peter Vail. 2019. “Estimating the Impact of Drought on Agriculture Using the U.S. Drought Monitor.” *American Journal of Agricultural Economics* 101 (1): 193–210. https://doi.org/10.1093/ajae/aay037.

Lobell, David B., Graeme L. Hammer, Greg McLean, Carlos Messina, Michael J. Roberts, and Wolfram Schlenker. 2013. “The Critical Role of Extreme Heat for Maize Production in the United States.” *Nature Climate Change* 3 (5): 497–501. https://doi.org/10.1038/nclimate1832.

Mafoua, Edouard, Calum G. Turvey, Edouard Mafoua, and Calum G. Turvey. 2004. “EFFECTS OF WEATHER EVENTS ON LOSS RATIOS FOR CROP INSURANCE PRODUCTS: A COUNTY-LEVEL PANEL DATA ANALYSIS.” https://doi.org/10.22004/AG.ECON.20113.

Makki, Shiva S., and Agapi Somwaru. 2001. “Evidence of Adverse Selection in Crop Insurance Markets.” *The Journal of Risk and Insurance* 68 (4): 685. https://doi.org/10.2307/2691544.

Makowski D, Ben-Shachar M, Lüdecke D (2019). “bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework.” *Journal of Open Source Software*, **4**(40), 1541. doi:10.21105/joss.01541, https://joss.theoj.org/papers/10.21105/joss.01541.

Moore, Roger L, and Christos Siderelis. 2002. “USE AND ECONOMIC IMPORTANCE OF THE WEST BRANCH OF THE FARMINGTON RIVER.”

———. 2003. “Use and Economic Importance of the Wild and Scenic Chattooga River.” *Washington, DC: American Rivers, Inc. and National Park Service Park Planning and Special Studies and Rivers, Trails and Conservation Assistance Programs*.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M (2011). “pROC: an open-source package for R and S+ to analyze and compare ROC curves.” *BMC Bioinformatics*, **12**, 77.

Rosen, Sherwin. 1974. “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.” *Journal of Political Economy* 82 (1): 34–55. https://doi.org/10.1086/260169.

Scanlon, Bridget R., Claudia C. Faunt, Laurent Longuevergne, Robert C. Reedy, William M. Alley, Virginia L. McGuire, and Peter B. McMahon. 2012. “Groundwater Depletion and Sustainability of Irrigation in the US High Plains and Central Valley.” *Proceedings of the National Academy of Sciences* 109 (24): 9320–25. https://doi.org/10.1073/pnas.1200311109.

Schlenker, Wolfram, and Michael J. Roberts. 2009. “Nonlinear Temperature Effects Indicate Severe Damages to U.S. Crop Yields under Climate Change.” *Proceedings of the National Academy of Sciences* 106 (37): 15594–98. https://doi.org/10.1073/pnas.0906865106.

Smithson, Michael, and Jay Verkuilen. 2006. “A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables.” *Psychological Methods* 11 (1): 54–71. https://doi.org/10.1037/1082-989X.11.1.54.

Smucker, Nathan J., and Naomi E. Detenbeck. 2014. “Meta-Analysis of Lost Ecosystem Attributes in Urban Streams and the Effectiveness of Out-of-Channel Management Practices: Out-of-Channel Restoration of Urban Streams.” *Restoration Ecology* 22 (6): 741–48. https://doi.org/10.1111/rec.12134.

Towe, Charles, H. Allen Klaiber, Joe Maher, and Will Georgic. 2021. “A Valuation of Restored Streams Using Repeat Sales and Instrumental Variables.” *Environmental and Resource Economics* 80 (2): 199–219. https://doi.org/10.1007/s10640-021-00575-9.

Walsh, Christopher J, Allison H Roy, Jack W Feminella, Peter D Cottingham, Peter M Groffman, and Raymond P Morgan. 2005. “The Urban Stream Syndrome: Current Knowledge and the Search for a Cure.” *Journal of the North American Benthological Society* 24 (3): 706–23.

Wang, Enli, and Chris J. Smith. 2004. “Modelling the Growth and Water Uptake Function of Plant Root Systems: A Review.” *Australian Journal of Agricultural Research* 55 (5): 501. https://doi.org/10.1071/AR03201.

White, Eric M., and Larry A. Leefers. 2007. “Influence of Natural Amenities on Residential Property Values in a Rural Setting.” *Society & Natural Resources* 20 (7): 659–67. https://doi.org/10.1080/08941920601171998.

Zhang, Wendong, and David M Wishart. 2019. “A Report on Results of Surveys to Measure the Demand for Designating the Mad River in Clark County Ohio a Recreational River.”