Local Crop Diversity Slows Pest Diffusion: Evidence from the Boll Weevil*

Tristan Emma Du Puy, Marguerite Obolensky†

December 2025

Abstract: Modern agricultural landscapes are becoming increasingly homogeneous, driven by the rise of monoculture and the pursuit of economies of scale through larger, more contiguous tracts of land. The ecological consequences of this landscape simplification remain poorly quantified. We use the historical diffusion of the boll weevil across the U.S. Cotton Belt to causally link crop diversity to pest resilience. By developing a harmonized dataset that links historical agricultural censuses over time and leveraging terrain ruggedness as an instrument for local crop diversity, we show that more diverse croplands significantly slow pest spread. A one-standard-deviation increase in crop diversity reduces annual contamination risk by 12 percentage points, or 36 percent relative to the mean annual risk. This study demonstrates that enhancing crop diversity can serve as an important component of sustainable pest management.

Classification: Social Sciences, Economic Sciences
Keywords: Crop diversity, Landscape Complexity, Landscape Simplification, Pest Control

*This paper benefited from feedback at various stages from Pierre Bodéré, Michael Crossley, Eyal Frank, Joséphine Gantois, Nicolas Longuet-Marx, Anouch Missirian, Suresh Naidu, Wolfram Schlenker, and Paul Rhode. We are grateful for comments from numerous participants in seminars and conferences at Columbia University, the LSE Shifting Landscape Conference, and AERE OSWEET. We are particularly indebted to Christopher Muller for sharing data and to Frederik Noack for his generous feedback at a critical point in our project. All remaining errors are our own.

†Du Puy: Center for Political Economy, Columbia University (td2631@columbia.edu); Obolensky: Kellogg School of Management, Northwestern University (marguerite.obolensky@kellogg.northwestern.edu). Both authors contributed equally to this work. Both designed the research, performed the research, contributed analytic tools, analyzed the data, and wrote the paper. The authors declare no competing interests.

1 Introduction

Modern transformations in agricultural markets have reshaped rural landscapes. Agricultural trade liberalization, mechanization, irrigation, and chemical inputs have increased productivity but have also encouraged monoculture and landscape simplification—larger contiguous fields, fewer crop types, and lower edge-to-area ratios (Tilman, Fargione, et al., 2001; Tilman, Cassman, et al., 2002; Crossley et al., 2020). While the ecological consequences of these changes for pollination and biodiversity are well-documented (Schroeder et al., 2024; Hemberger et al., 2021; Palmu et al., 2014), their causal effect on agricultural vulnerability to pests remains largely untested.

In this paper, we show that local crop diversity slows pest spread. To do so, we exploit one of the most damaging pest shocks to U.S. agriculture: the diffusion of the boll weevil across the Cotton Belt between 1890 and 1930 (Hunter, 1923; Lange et al., 2009; Bloome et al., 2017). This historical setting allows us to estimate the unmitigated effect of crop diversity on pest diffusion as the boll weevil shock predates the introduction of commercial pesticides (Hunter, 1923).

We estimate that a one-standard-deviation increase in the county-level Gini-Simpson index of crop diversity reduces the annual hazard of first contamination by about 12 percentage points, or 36 percent relative to the mean annual risk. Moving a county from the 25th to the 75th percentile of crop evenness reduces the annual hazard by roughly 16 percentage points and, under a simple constant-hazard approximation, delays median contamination by two years.

Our empirical strategy combines a new long-run measure of crop diversity with quasi-experimental variation in terrain ruggedness. First, we construct a county-level panel of planted acres for major crops from 1850 to 2007 by developing a machine-learning algorithm that predicts historical crop distributions to harmonize the U.S. Census of Agriculture across shifting county boundaries. Second, we use terrain ruggedness as an instrument for crop diversity. Rugged landscapes affect drainage, erosion risk, soil depth, and the ease of plowing and irrigation, and in our data they are strongly associated with a less even mix of crops: a one–standard-deviation increase in ruggedness lowers the Gini–Simpson index of crop diversity by about 0.22 standard deviations.

To address the concern that ruggedness might also influence boll weevil diffusion through other channels—such as cotton suitability, local climate, or access to transportation networks—we condition on historical cotton intensity, total agricultural area, long-run temperature and rainfall, and state-by-decade fixed effects. We show that our results are robust to adding controls for distance to rail and market access, and to restricting attention to neighboring counties along the moving frontier of diffusion. Section 3 develops this

IV strategy and its identifying assumptions in detail, and reports additional robustness checks which demonstrate that results are not sensitive to alternative choices of diversity index.

A key feature of our setting is that the boll weevil only feeds on cotton. This characteristic narrows the range of potential channels which can link its diffusion to landscape structure. The first of these is the resource concentration hypothesis: more diverse agricultural landscapes tend to be more fragmented, and host plants—here cotton—tend to be less contiguous, slowing pest reproduction and migration (Tilman, Cassman, et al., 2002; Larsen and Noack, 2017). By controlling for initial cotton shares, we isolate how the spatial distribution of cotton and non-cotton land affects pest spread, independent of host plant availability. The second is the natural enemies hypothesis: more diverse landscapes also support non-crop habitat for predators such as birds, bats, and beneficial insects that suppress pest populations (Hunter, 1923; Gurr et al., 2003; Lin, 2011; McDaniel et al., 2014; Landis et al., 2008; Sirami et al., 2019). A third, complementary mechanism is that crop diversity often correlates with grazing practices that reduce overwintering sites for pests (Hunter, 1923).

This paper makes three main contributions. First, we develop a new historical dataset that traces county-level crop diversity in the United States from 1850 to 2007. To overcome shifting county boundaries and the heterogeneous spatial distribution of cropland within counties, we construct a machine-learning linking algorithm that predicts the fine-scale location of cultivated land using topography and climate. This approach substantially improves on the area-weighted crosswalks commonly used in economic and ecological work (e.g., Hornbeck and Naidu, 2014; Crossley et al., 2020; Eckert et al., 2020), and provides a more accurate, spatially consistent series of planted acres for the full U.S. agricultural census. Using this harmonization, we document a 48 percent decline in county-level crop diversity since the mid-twentieth century, a number in line with prior accounts of long-run landscape homogenization (Aguilar et al., 2015; Hijmans et al., 2004; Crossley et al., 2020).

Second, we provide the first causal evidence that higher crop diversity slows the spatial diffusion of an agricultural pest, complementing ecological research showing that biodiversity can buffer disease emergence (Weitzman, 2000) and transmission (Hartfield et al., 2011; Civitello et al., 2015; Rohr et al., 2020; Manning and Ando, 2022). Our empirical setting allows us to quantify this relationship at scale and complements theoretical, experimental and field-based ecological studies (Berry et al., 2019; Gurr et al., 2003; Lin, 2011; McDaniel et al., 2014; Landis et al., 2008) by showing that these mechanisms aggregate to measurable reductions in pest spread across counties.

Third, we contribute to the economics of sustainable pest management practices (Eiswerth

et al., 2018; Liebhold and Griffin, 2016; Epanchin-Niell et al., 2021) and the broader literature on spatial externalities in agriculture. This literature is developed against a backdrop of increasingly well-documented health costs of pesticide exposure (Frank, 2024; Dias et al., 2023; Taylor, 2022; Calzada et al., 2023). Because farmers currently address pest pressure primarily through pesticide inputs, the spatial diffusion of pests generates both economic losses and downstream health externalities. Our results show that individual farmers' land-use decisions generate sizable externalities by altering the likelihood and speed with which a pest reaches neighboring fields. Yet existing agricultural policy instruments fail to internalize these landscape-level interactions. More efficient policy design could promote diversified planting or preserve non-crop habitat. A simple back-of-the-envelope calculation suggests that moving a county from the 25th to the 75th percentile of crop evenness would have increased expected annual cotton revenues by roughly $2,300 (in 1890 dollars, corresponding to 5 percent of annual county-level cotton revenues). Beyond these direct gains, greater diversification could also lower reliance on chemical pesticides. While our analysis focuses on quantifying the pest-diffusion channel, these complementary health and ecological gains further strengthen the case for incorporating landscape externalities into agricultural policy.

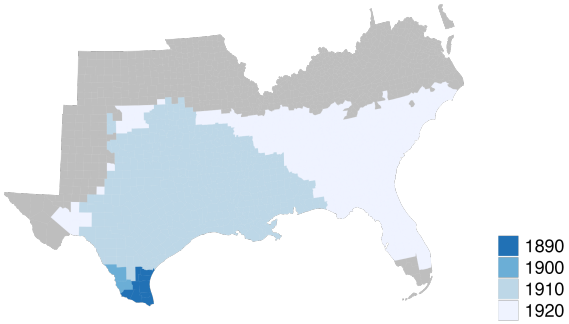
2 Context and Data

The boll weevil is an insect native to Mexico and Central America. It was first observed in the U.S. in 1892 in Brownsville, Texas, and later spread across the entire Cotton Belt, where it caused substantial agricultural damage (Lange et al., 2009).[1] We track its county-level advance using digitized historical maps, which reveal a concentric diffusion pattern consistent with its primarily wind-driven spread (Figure 1; Bloome et al., 2017; Lange et al., 2009). Critically, this pest shock predates the widespread use of effective pesticides—calcium arsenate was not introduced until the late 1920s (Hunter, 1923). As a result, the pest's spatial dynamics during this period reflect relatively unmediated interactions between its behavior and landscape features, making it a uniquely clean setting to study the ecological consequences of crop diversity.

Measuring crop diversity. We measure crop diversity using the Gini-Simpson index, defined as the probability that two randomly selected hectares in a county are planted with different crops. The index increases with both the number of crops grown and the evenness

[1] It is estimated that after five years, cotton production in contaminated counties declined by 40%. Impacted counties experienced significant out-migration, and by transforming the prevalence of tenant farming, the boll weevil even reduced marriage rates among young African Americans in the South (Bloome et al., 2017).

Figure 1: Boll Weevil Contamination over Time



Notes: The figure shows the decade each county was first contaminated by the boll weevil, based on digitized historical maps (Bloome et al., 2017). The sample includes all counties where cotton was grown at the start of the twentieth century.

of land allocation (Jost, 2006).[2][3] While many dimensions of biodiversity may influence the spread of pests, we chose county-level crop diversity as our primary metric for three key reasons. First, crop diversity shapes the abundance and spatial distribution of host plants, which in turn affects pest survival and transmission (Strobl, 2021; Tilman, Cassman, et al., 2002). Second, it correlates with broader landscape characteristics relevant to pest ecology: more diverse croplands tend to be more fragmented and provide greater non-crop habitat, which can support natural enemies of agricultural pests (Gurr et al., 2003; Lin, 2011; McDaniel et al., 2014; Landis et al., 2008; Palmu et al., 2014). As such, crop diversity serves as a useful proxy for key landscape configurations. Finally, crop diversity is one of the few landscape structure-related variables that can be consistently measured across the entire twentieth century in the U.S., thanks to the systematic coverage of the U.S. Census of Agriculture.

County boundaries have shifted over time, complicating longitudinal comparisons within the U.S. Census of Agriculture. To make historical data comparable across years, researchers typically reassign historical data to a fixed map of counties. However, standard crosswalk methods (e.g., Eckert et al., 2020) assume cropland is evenly distributed within counties—an unrealistic assumption that can bias temporal comparisons.[4] We address this limitation by developing a novel linking algorithm. First, we train a machine learning model on modern fine-grained cropland data, using topography and climate as predictors.[5] Second, we use this model to predict where agricultural activity was most likely

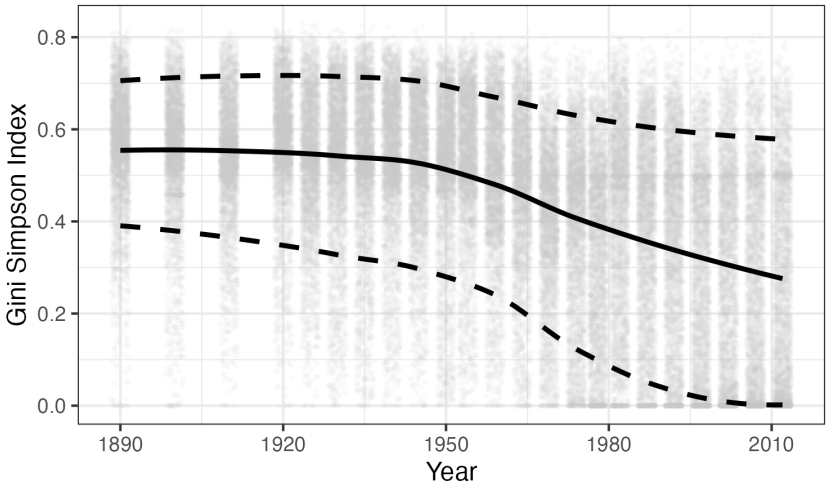
[2] In a two-crop landscape, the index can never rise above 0.5. In a four-crop landscape, the maximum increases to 0.75. Holding the number of crops constant, the index increases with the evenness of land allocation.

[3] In robustness exercises, we use the Pielou index, which captures evenness of crop abundance within a county.

[4] Figure A.1 shows sub-county cultivated areas in Kansas, illustrating that this homogeneity assumption fails for most counties.

[5] We use the Cropland DataLayer that has a 30-meter resolution to recover current cropland distribution (United States Department of Agriculture and National Agricultural Statistics Service, 2023). We obtain data on historical and contemporary weather from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) group (PRISM Climate Group, Oregon State University, 2025).

Figure 2: Crop Diversity over Time



Notes: The figure shows the evolution of crop diversity in the U.S. over time. Each dot represents a U.S. county, the solid line is the national average, and the dotted lines are the 10th and 90th percentiles. We construct the crop diversity index from the U.S. Census of Agriculture. See Appendix A.1 for construction details.

concentrated within each historical county between 1890 and 1930, the period of expansion of the boll weevil. Finally, we use the predicted distributions to construct reweighting schemes that refine traditional area-based methods. Appendix A.1 provides full details.

Using the harmonized county-level dataset, we construct a balanced panel of planted acres for 12 major crops.[6] Consistent with prior work (Crossley et al., 2020), we find a sharp decline in crop diversity across the U.S. beginning over the second half of the twentieth century (Figure 2). Between 1950 and 2010, crop diversity fell by 48 percent from 0.56 to 0.29. This sustained loss of crop diversity reinforces the importance of understanding its ecological consequences, particularly its potential role in shaping the spatial dynamics of pest diffusion.

3 Empirical Strategy

A key empirical challenge is that many long-run determinants of cropping patterns—such as the suitability of land for cotton, climate conditions, or access to markets and transportation—can influence both crop diversity and the diffusion of the boll weevil. Simple correlations between diversity and contamination are therefore unlikely to be causal.

We address this concern using an instrumental-variables (IV) approach in a linear probability model.[7] Specifically, we estimate a two-stage least squares (2SLS) model that

[6] Rice, cane sugar, wheat, corn, cotton, potato, sweet potato, barley, rye, tobacco, buckwheat, and oats are included. Notice the exclusion of soybeans, which is only accounted for in the U.S. Census of Agriculture from 1920 onwards. Results are robust to using a larger set of crops in an unbalanced panel.

[7] In robustness checks, we verify that results are largely unchanged when using a logit model with a control function.

leverages county-level terrain ruggedness as a source of quasi-exogenous variation in the Gini–Simpson index of crop diversity:[8]

$$\text{Gini-Simpson}_{ct} = \gamma_1 \text{Ruggedness}_{ct} + \mathbf{X}_{ct}\gamma_2 + \eta_{s(c)t} + \varepsilon_{ct}. \tag{1}$$

We define the terrain ruggedness index, Ruggedness_{ct}, as the mean of the absolute differences in elevation between a focal grid cell and its eight neighbors (Wilson et al., 2007). Elevation data come from the Landfire 2020 raster provided by the USGS (U.S. Department of the Interior, U.S. Geological Survey and U.S. Department of Agriculture, 2013). The fitted values from this regression are used in the second stage to estimate the effect of crop diversity on pest diffusion:

$$\text{BollWeevil}_{ct} = \beta_1 \widehat{\text{Gini-Simpson}}_{ct} + \mathbf{X}_{ct}\beta_2 + \zeta_{s(c)t} + \nu_{ct}. \tag{2}$$

The outcome variable is a binary indicator equal to 1 if county c is contaminated by the boll weevil in decade t. Because the boll weevil feeds exclusively on cotton, we restrict our analysis to counties that cultivated cotton at least once between 1890 and 1930. Once contaminated, counties remain so in future years. $\eta_{s(c)t}$ and $\zeta_{s(c)t}$ are state-by-decade fixed effects and X_{ct} corresponds to a vector of controls. Standard errors are clustered at the county level to account for autocorrelation.[9] Summary statistics for the main variables used in the analysis are shown in Table B.3.

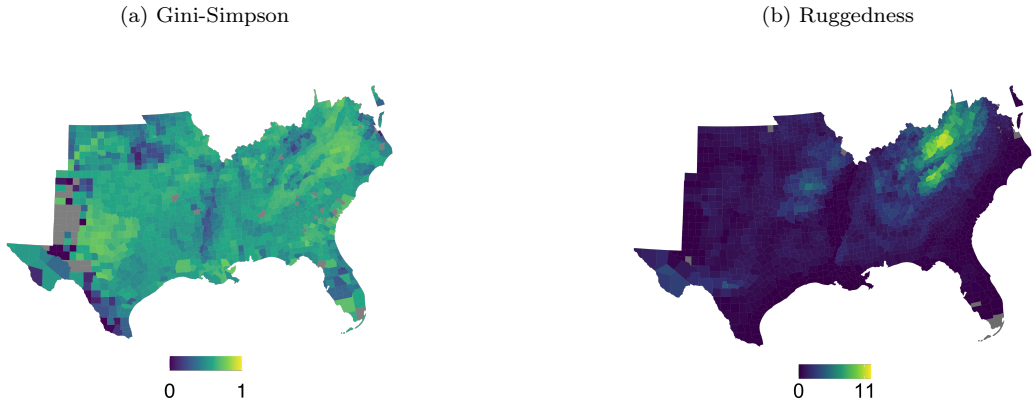
Relevance. Terrain ruggedness imposes physical constraints on agriculture by affecting drainage, irrigation feasibility, soil depth, soil organic content, and erosion risk, all of which influence the ease and profitability of plowing and cultivation (Patton et al., 2019; Vollering et al., 2025; Khawlie et al., 2002). As a result, rugged landscapes tend to have less flexible cropping possibilities and more limited scope for diversified rotations. In our data, more rugged counties have systematically lower crop diversity. A one–standard-deviation increase in terrain ruggedness reduces the Gini–Simpson index of crop diversity by about 0.01 (0.22 standard deviations; Table 1, column (2)). The first-stage F-statistic of 37.4 exceeds the conventional threshold for instrument strength. Figure 3 maps the spatial distribution of both variables and illustrates this negative association.

Exclusion. A valid instrument must not affect boll weevil diffusion through any channel other than crop diversity. In our context, the main concern is that ruggedness could be correlated with other determinants of pest spread—such as the density and spatial

[8] We use the Gini–Simpson index as our main measure of crop diversity and show in Table C.4 that results are robust when using the Pielou index.

[9] We test in the appendix the robustness of our results to allowing for spatial correlation via Conley standard errors with different distance cut-offs.

Figure 3: Crop Diversity and Terrain Ruggedness



Notes: Panel (a) shows the Gini-Simpson index across cotton-growing counties in 1890. See Appendix A.1 for details on the construction of this crop diversity index. Panel (b) shows the spatial distribution of the ruggedness index, defined as the terrain ruggedness index, or the mean of the absolute differences in elevation between a focal grid cell and its eight neighbors (Wilson et al., 2007). A higher index value corresponds to a more uneven terrain. Elevation data come from the Landfire 2020 raster provided by the USGS (U.S. Department of the Interior, U.S. Geological Survey and U.S. Department of Agriculture, 2013).

configuration of cotton fields, local climate, or access to transportation networks—even after conditioning on crop diversity.

We address these threats in several ways. First, we explicitly control for agronomic and climatic determinants of boll weevil reproduction that might co-vary with ruggedness. The boll weevil depends exclusively on cotton to reproduce, thrives under warm and wet conditions, and suffers from cold winters (Hunter, 1923; Lange et al., 2009). To account for these mechanisms, we compare counties with a similar baseline share of cropland devoted to cotton and total agricultural acreage in 1880 (the decade of first boll weevil entry into the United States), as well as long-run averages of annual rainfall and minimum and maximum daily temperatures.[10] Conditioning on both cotton's share of cropland and overall agricultural area helps absorb much of the variation in host-crop availability and crop density that could otherwise provide a separate pathway from ruggedness to pest diffusion.

Second, we include state-by-decade fixed effects, which flexibly control for unobserved policy shocks, pest-management practices, and broader regional climate patterns that evolve over time. This structure absorbs, for example, state-level responses to contamination, extensions services, or large-scale climate anomalies that might be correlated with ruggedness at broad spatial scales. While we do not directly observe historical wind speeds or directions, these fixed effects and weather controls substantially limit the scope for terrain-induced differences in climate to bias our estimates outside of their influence on cropping patterns.

[10] Baseline share of cotton acreage is computed in 1880, at the time of first entry of the boll weevil in the U.S. We compute weather averages between 1895–1997 using data from PRISM (PRISM Climate Group, Oregon State University, 2025). Results are robust to using alternative windows, such as 1895–1930.

Third, we consider additional channels explicitly in the results section. In particular, ruggedness could also be correlated with access to trade and transportation networks, allowing boll weevils to travel along rail or water routes. To address this, we show that our IV estimates are robust to including controls for distance to rail and Donaldson and Hornbeck (2016)-style market access and to restricting attention to counties on the moving frontier of infestation, where exposure risk is similar across counties. These exercises suggest that the variation in boll weevil diffusion used for identification is not primarily driven by differences in market access or location relative to the initial entry point.

Taken together, the host specialization of the boll weevil, the controls for baseline cotton intensity and agricultural area, the long-run climate variables, and the rich set of fixed effects greatly reduce the range of alternative channels through which terrain ruggedness could affect pest diffusion. Our key identifying assumption is that, once these factors and crop diversity are accounted for, any remaining influence of ruggedness on boll weevil spread operates only through the cropping patterns it induces. In the next section, we document that the estimated effect of diversity is stable across a wide set of specifications and samples, which is consistent with this assumption.

4 Results

Table 1 reports estimates of the effect of crop diversity on pest diffusion for the full sample. Column (1) presents OLS estimates, which are small and statistically insignificant, while column (2) reports the baseline 2SLS specification. The IV estimates indicate that a one-standard-deviation increase in the Gini-Simpson index of crop diversity reduces the annual probability of boll weevil contamination by roughly 12 percentage points—about 36 percent relative to the mean annual risk.[11] The difference between the OLS and IV estimates suggests that the OLS coefficients are attenuated, reflecting downward bias arising from the endogeneity of crop diversity.

To interpret magnitude, consider a shift from the 25th to the 75th percentile of predicted evenness in the first stage (0.49 to 0.57). This interquartile-range increase reduces the annual hazard of contamination by approximately 16 percentage points. Under a constant-hazard approximation it delays the median time to pest arrival by roughly 2 years and increases the expected time to arrival by about 2.9 years.[12]

[11] Following standard IV practice, all summary statistics used to interpret IV coefficients refer to the predicted endogenous variable from the first stage, since this reflects the variation used for identification.

[12] Under a constant annual hazard h, the probability that a county remains uncontaminated after t years is $(1-h)^t$. The median arrival time is therefore $T_{0.5} = \ln(0.5)/\ln(1-h)$, and the expected arrival time is $1/h$. Applying these expressions to the baseline hazard ($h_0 = 0.33$) and to the hazard implied by the 25th–75th percentile increase in diversity ($h_1 = 0.17$) yields a median delay of about 2 years and an increase in expected arrival time of roughly 2.9 years.

Table 1: Diversity and Boll Weevil Diffusion

Dependent Variable:	Boll Weevil Presence			
Model:	(1) OLS	(2) IV	(3) OLS	(4) IV
Crop Diversity (Gini-Simpson)	-0.033	-1.975***	-0.027	-1.715***
	(0.047)	(0.735)	(0.047)	(0.563)
State × Decade FE	Yes	Yes	Yes	Yes
Observations	3,895	3,895	3,895	3,895
Number of Clusters	1,080	1,080	1,080	1,080
R^2	0.86	0.72	0.86	0.76
Mean Boll Weevil	0.33	0.33	0.33	0.33
SD Boll Weevil	0.47	0.47	0.47	0.47
Mean Gini-Simpson	0.54	0.53	0.54	0.53
SD Gini-Simpson	0.10	0.06	0.10	0.06
First Stage		-0.011 (0.003)		-0.011 (0.003)
Mean Ruggedness		1.34		1.34
SD Ruggedness		1.27		1.27
F-statistic		37.4		54.1

Notes: This table reports results from OLS and IV estimation of the effect of crop diversity on boll weevil contamination, measured by the Gini-Simpson index. The sample contains observations of counties in the cotton belt between 1890 and 1930. The dependent variable is equal to 1 if a county is affected by the boll weevil in decade t, and 0 otherwise. In all specifications, we control for state by decade fixed effects. In columns (1) and (2), we control for the share of cotton in agricultural acreage in 1880, agricultural acreage in 1880, average rainfall, average daily minimum, and maximum temperature. In columns 3 and 4, we additionally control for the richness index measured in 1880. Standard errors are clustered at the county level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Using NASS's estimate that total U.S. cotton revenues in 1890 were approximately $43 million (about $40,000 per cotton-growing county) (USDA-NASS, 2008), and historical evidence that boll weevil infestation reduced cotton output by roughly 40–50 percent (Lange et al., 2009), this reduction in annual hazard corresponds to an expected annual revenue gain of about $2,300 per county. Interpreted through the timing channel, the two-year delay in median arrival implies avoided cotton losses on the order of $36,000 per county (1890 dollars), given typical annual losses of roughly 45 percent of revenues once the pest becomes established.

Mechanisms Several mechanisms may explain the negative relationship between crop diversity and pest spread. Increased crop diversity may reduce pest spread by fragmenting the spatial continuity of host crops—an effect consistent with the resource dilution hypothesis (Larsen and Noack, 2017; Clough et al., 2020). The boll weevil depends exclusively on cotton and remains largely sedentary outside of its brief August migration period, during which it spreads out to other locations (Hunter, 1923). In more diverse agricultural landscapes, cotton fields tend to be more spatially fragmented, reducing the

size and contiguity of suitable habitat and making it harder for migrating weevils to locate food. Historical USDA reports from the 1920s recognized this principle, proposing the use of non-cotton "barrier crops" to slow the advance of the pest. This mechanism is supported by the results in columns (3) and (4) of Table 1, where we include baseline crop richness as a control. The estimate remains large and statistically significant: a one-standard-deviation increase in evenness reduces the annual probability of contamination by 10 percentage points (31 percent relative to the mean annual risk). This suggests that the reduction in pest spread is primarily driven by evenness in land allocation—i.e., lower crop homogeneity—rather than by the number of crops grown.[13] Importantly, we control for the share of cotton in 1880 in all specifications, ensuring that these results are not driven by variation in cotton availability across counties.

Second, diverse cropping systems may suppress pest populations by supporting a greater abundance of natural predators—a mechanism known as the natural enemy hypothesis (Larsen and Noack, 2017; Fahrig et al., 2011). A more heterogeneous cropland provides varied habitat and foraging resources for beneficial insects, birds, and bats that prey on agricultural pests (Hiller et al., 2025). Historical USDA reports identified dozens of natural insect enemies of the boll weevil and noted that birds were frequent predators (Hunter, 1923). These predator groups are known to decline in more homogeneous agricultural landscapes (Hemberger et al., 2021; Hiller et al., 2025).

Third, crop diversity may correlate with more varied land management practices, including fall grazing, which inhibits boll weevil hibernation by removing plant residues (Hunter, 1923). Historical studies document that agricultural specialization at the county level often involved a shift toward either cropland or livestock production (Winsberg, 1982; Gregson, 1996; MacDonald and Newton, 2018), potentially reducing the presence of grazing animals in crop-dominant areas.

These last two mechanisms likely complement the role of host plants homogeneity in explaining our results. However, the lack of data on historical predator abundance, management intensity, or grazing practices prevents us from directly quantifying their role in our context.

Accounting for transportation networks A key identifying assumption of our strategy is that terrain ruggedness affects pest diffusion only through its influence on crop diversity. One potential threat to this assumption is that rugged terrain may also be correlated with access to trade networks. While previous research has shown that the boll weevil spread primarily via wind rather than trade routes (Lange et al., 2009), we test the robustness of our estimates by controlling for proxies of historical market access.

[13] Due to limited spatial variation in crop count during our study period, we do not use richness as an alternative crop diversity index, but include it as a control to help isolate the specific contribution of agricultural homogeneity.

Table 2: Diversity and Boll Weevil Diffusion—Transportation Networks

Dependent Variable:	Boll Weevil Presence		
Model:	(1)	(2)	(3)
	IV	IV	IV
Crop Diversity (Gini-Simpson)	-1.975***	-2.153**	-2.391**
	(0.735)	(0.879)	(1.105)
State × Decade FE	Yes	Yes	Yes
Additional Controls		Distance to rail	Market Access
Observations	3,895	3,895	2,949
Number of Clusters	1,080	1,080	1,055
R^2	0.72	0.70	0.67
Mean Boll Weevil	0.33	0.33	0.43
SD Boll Weevil	0.47	0.47	0.50
Mean Gini-Simpson	0.53	0.54	0.52
SD Gini-Simpson	0.06	0.07	0.08
First Stage	-0.011 (0.003)	-0.015 (0.0023)	-0.018 (0.005)
Mean Ruggedness	1.34	1.34	1.33
SD Ruggedness	1.27	1.27	1.25
F-statistic	37.4	27.7	20.7

Notes: This table reports IV estimates of the effect of crop diversity, measured by the Gini-Simpson index, on boll weevil contamination. Column (1) reproduces the baseline IV specification. Column (2) adds a control for the distance from the county centroid to the nearest rail segment. Column (3) includes a control for county-level market access (Donaldson and Hornbeck, 2016). All specifications control for state-by-decade fixed effects, as well as for the 1880 share of cotton in agricultural acreage, total cultivated area in 1880, average rainfall, and average daily minimum and maximum temperature. Standard errors are clustered at the county level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

In Table 2, we report IV estimates controlling for two transportation-related variables. In column (2), we control for the evolving distance between each county's centroid and the nearest rail segment. In column (3), we instead use the market access measure developed by Donaldson and Hornbeck (2016) (Donaldson and Hornbeck, 2016), which accounts for a county's connectivity to national transport networks via rail, canals, and natural waterways. Across both specifications, the estimated effect of crop diversity remains negative and statistically significant. Effects increase slightly in magnitude to 13-14 percent for the one-standard-deviation increase in diversity observed in our main sample, compared to 12 percent in the baseline.

Alternative samples To further support the validity of our identification strategy, we estimate the model on increasingly restrictive samples that reduce the risk of reverse causality and strengthen identification. One concern is that farmers may have responded to boll weevil contamination by diversifying their crops. If this response is correlated with terrain ruggedness, including post-contamination years could increase the magnitude of our estimates. Column (1) in Table 3 reproduces our baseline IV estimates. In column

Table 3: Diversity and Boll Weevil Diffusion—Alternative Samples

Dependent Variable:	Boll Weevil Presence		
Model:	(1)	(2)	(3)
	IV	IV	IV
Crop Diversity (Gini-Simpson)	-1.975***	-1.551***	-1.728**
	(0.735)	(0.491)	(0.729)
State × Decade FE	Yes	Yes	Yes
Sample	Full	Drop After Contamination	Drop Neighboring Contaminated
Observations	3,895	3,440	804
Number of Clusters	1,080	1,080	685
R^2	0.72	0.74	0.61
Mean Boll Weevil	0.33	0.25	0.61
SD Boll Weevil	0.47	0.43	0.49
Mean Gini-Simpson	0.53	0.54	0.51
SD Gini-Simpson	0.06	0.07	0.09
First Stage	-0.011 (0.0023)	-0.016 (0.003)	-0.020 (0.005)
Mean Ruggedness	1.34	1.40	1.67
SD Ruggedness	1.27	1.31	1.51
F-statistic	37.4	52.7	24.0

Notes: This table reports IV estimates of the effect of crop diversity on boll weevil contamination—measured by the Gini-Simpson index—using progressively more restrictive samples. Column (1) reproduces the baseline IV specification (Equation 2. Column (2) drops counties in all decades after their first contamination. Column (3) further restricts the sample to counties with a neighboring county contaminated in the previous period. All specifications control for state-by-decade fixed effects, the 1880 share of cotton in agricultural acreage, total cultivated area, average rainfall, and daily minimum and maximum temperature. Standard errors are clustered at the county level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

(2), we restrict the sample to counties that are either not yet contaminated or in their first decade of contamination, when such adaptive responses were rarely observed (Hunter, 1923). The estimated effect remains negative and statistically significant: a one-standard-deviation increase in crop diversity reduces the annual likelihood of contamination by 11 percentage point, or 33 percent of the average contamination probability. The slightly smaller magnitude relative to the baseline estimate is consistent with an adaptation scenario.

In column (3), we further restrict the sample to counties on the boundary of the contaminated zone, defined as those with at least one neighboring county contaminated in the previous decade. This allows for comparisons among counties with similar exposure risk. This more restrictive sample helps address the possibility that more diverse counties were systematically located farther from the initial zone of boll weevil contamination. Even in this most conservative sample, the effect of crop diversity remains large and statistically significant: a one-standard-deviation increase in crop diversity, using the baseline standard deviation value for comparison, reduces the likelihood of contamination by 16 percentage points.

Table 4: Diversity and Boll Weevil Diffusion - Nonlinear Models

Dependent Variable:	Boll Weevil Presence		
Model:	(1)	(2)	(3)
	IV	Logit + CF	Probit + CF
Crop Diversity (Gini-Simpson)	-1.975***	-78.45***	-38.10***
	(0.734)	(19.02)	(10.17)
State × Decade FE	Yes	Yes	Yes
Observations	3,895	1,364	1,364
Number of Clusters	1,080	691	691
Marginal effect (1 s.d.)	-0.124	-0.226	-0.206
Mean Boll Weevil	0.33	0.33	0.33
SD Boll Weevil	0.47	0.47	0.47
Mean Gini-Simpson	0.53	0.53	0.53
SD Gini-Simpson	0.06	0.06	0.06
First Stage	-0.011 (0.003)	-0.011 (0.003)	-0.011 (0.003)
Mean Ruggedness	1.32	1.32	1.32
SD Ruggedness	1.22	1.22	1.22

Notes: This table compares estimates of the effect of crop diversity on the probability of boll weevil contamination across three models. Column (1) presents the baseline linear IV specification estimated by 2SLS, where crop diversity (Gini–Simpson index) is instrumented with terrain ruggedness. Columns (2) and (3) report logit and probit models estimated using a control-function approach: the first-stage residual from regressing crop diversity on ruggedness and controls (with state-by-decade fixed effects) is included as an additional regressor to correct for endogeneity. All models include state-by-decade fixed effects and the same set of covariates as in the baseline specification. Standard errors are clustered at the county level. The logit and probit marginal effects (about 20–22 pp for a one-standard-deviation increase in predicted diversity) closely match the linear IV results, indicating robustness to functional form. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Robustness The main IV estimates are robust to allowing for nonlinear link functions. Table 4 reports logit and probit specifications estimated using a control-function approach, which corrects for the endogeneity of crop diversity by including the first-stage residual from the ruggedness regression. The coefficients in the nonlinear models are substantially larger in magnitude, as expected given the steepness of the logit and probit response functions in the range of baseline contamination risks observed in our sample. However, the implied effects on the annual probability of contamination remain similar to our baseline linear IV estimates: a one-standard-deviation increase in predicted crop diversity reduces the annual hazard of contamination by roughly 20–22 percentage points (compared to 12 percentage points in the linear model).

In Appendix C, we present three additional robustness exercises. First, we re-estimate our main specification using the Pielou index as an alternative measure of crop diversity, which isolates the evenness of crop distribution, in contrast to the Gini-Simpson index, which reflects both richness and evenness (Pielou, 1966). The results—a 11–14 percentage point reduction in contamination risk—are statistically significant and closely aligned with our baseline estimates. Second, we test the sensitivity of our results to alternative fixed effect specifications. The estimated effect of crop diversity remains stable whether

we control only for state fixed effects or only for decade fixed effects, rather than their interaction. Finally, we evaluate the robustness of our results to spatial correlation using Conley standard errors. The coefficients remain statistically significant when we consider a threshold of spatial correlation below 50 km. To put this threshold in perspective, counties in our sample have an average diameter of 41 km.

5 Conclusion

We provide causal evidence that local crop diversity dampens the spatial spread of a specialist pest. In our preferred IV specification, a one-standard-deviation increase in the Gini-Simpson index reduces the annual hazard of first boll weevil contamination by about 12 percentage points, or roughly 36 percent relative to the mean annual risk. Under a constant-hazard approximation, this corresponds to delaying the median time to contamination by two years and increasing the expected time to arrival by about 2.8 years.

The magnitude of the effect is consistent with two well-established ecological mechanisms. More even cropland allocations reduce the contiguity of host plants, limiting the ability of a specialist herbivore to reproduce and spread, and more heterogeneous landscapes tend to support higher abundance of natural enemies. Although our county-decade data do not allow us to disentangle these channels, the consistency of the estimates across specifications suggests that landscape configuration plays a central role in shaping pest diffusion.

Our study focuses on a primarily wind-borne, specialist herbivore in a pre-pesticide era. Many modern pests are mobile along trade networks, and pesticides alter population dynamics. Even so, where pest movement is local (e.g., many beetles, moths, and fungal vectors) and where enemies benefit from field margins and mixed land use, we expect qualitatively similar effects, though magnitudes may vary with patchiness, crop phenology, and resistance traits.

As agricultural landscapes continue to simplify, our evidence highlights a cost that mitigates the benefits of scale economies in production. The historical boll weevil diffusion shows that modest increases in crop evenness materially slow diffusion. Overall, the results suggest that landscape diversification can be a valuable complement to chemical control.

References

Aguilar, Jonathan et al. (2015). "Crop Species Diversity Changes in the United States: 1978–2012". In: *PlosONE*.

Berry, Kevin, Eli P. Fenichel, and Brian E. Robinson (2019). "The Ecological Insurance Trap". In: *Journal of Environmental Economics and Management*.

Bloome, Deirdre, James Feigenbaum, and Christopher Muller (2017). "Tenancy, marriage, and the boll weevil infestation, 1892–1930". In: *Demography* 54.3, pp. 1029–1049.

Calzada, Joan, Meritxell Gisbert, and Bernard Moscoso (2023). "The hidden cost of bananas: the effects of pesticides on newborns' health". In: *Journal of the Association of Environmental and Resource Economists* 10.6, pp. 1623–1663.

Civitello, David J. et al. (2015). "Biodiversity inhibits parasites: Broad evidence for the dilution effect". In: *Proceedings of the National Academy of Sciences* 112.28.

Clough, Yann, Stefan Kirchweger, and Jochen Kantelhardt (2020). "Field sizes and the future of farmland biodiversity in European landscapes". In: *Conservation Letters* 13.6.

Crossley, Michael S. et al. (2020). "Recent Collapse of Crop Belts and Declining Diversity of US Agriculture since 1840". In: *Global Change Biology* 27.1, pp. 151–164.

Dias, Mateus, Rudi Rocha, and Rodrigo R. Soares (2023). "Down the River: Glyphosate Use in Agriculture and Birth Outcomes of Surrounding Populations". In: *The Review of Economic Studies* 90.6, pp. 2943–2981.

Donaldson, Dave and Richard Hornbeck (2016). "Railroads and American Economic Growth: A "Market Access" Approach". In: *Quarterly Journal of Economics* 131.2, pp. 799–858.

Eckert, Fabian et al. (2020). "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790". In: *NBER Working Paper*.

Eiswerth, Mark, Chad Lawley, and Michael H Taylor (2018). "Economics of invasive species". In: *Oxford Research Encyclopedia of Environmental Science*.

Epanchin-Niell, Rebecca et al. (2021). "Biological invasions and international trade: Managing a moving target". In: *Review of Environmental Economics and Policy* 15.1, pp. 180–190.

Fahrig, Lenore et al. (2011). "Functional landscape heterogeneity and animal biodiversity in agricultural landscapes". In: *Ecology Letters* 14.2, pp. 101–112.

Frank, Eyal (2024). "The Economic Impacts of Ecosystem Disruptions: Private and Social Costs From Substituting Biological Pest Control". In: *Science*.

Gregson, Mary Eschelbach (1996). "Long-Term Trends in Agricultural Specialization in the United States: Some Preliminary Results". In: *Agricultural History* 70, pp. 90–101.

Gurr, Geoff M., Stephen D. Wratten, and John Michael Luna (2003). "Multi-function agricultural biodiversity: pest management and other benefits". In: *Basic and Applied Ecology* 4, pp. 107–116.

Hartfield, Matthew, K. A. Jane White, and Klaus Kurtenbach (2011). "The role of deer in facilitating the spatial spread of the pathogen Borrelia burgdorferi". In: *Theoretical Ecology* 4.1, pp. 27–36.

Hemberger, Jeremy, Michael S. Crossley, and Claudio Gratton (2021). "Historical decrease in agricultural landscape diversity is associated with shifts in bumble bee species occurrence". In: *Ecology Letters* 24.9, pp. 1800–1813.

Hijmans, Robert J., Hyeyeong Choe, and Joshua Perlman (2004). "Spatiotemporal Patterns of Field Crop Diversity in the United States, 1870-2012". In: *Agricultural and Environmental Letters*.

Hiller, Thomas, Friederike Gall, and Ingo Grass (2025). "Enhanced crop diversity but not smaller field size benefit bats in agricultural landscapes". In: *Landscape Ecology* 40.1, pp. 1–12.

Hornbeck, Richard and Suresh Naidu (2014). "When the Levee Breaks: Black Migration and Economic Development in the American South". In: *American Economic Review* 104.3, pp. 963–90.

Hunter, Walter David (1923). "The Boll-Weevil Problem". In: *U.S. Department of Agriculture, Farmer's Bulletin No1329*.

Jost, Lou (2006). "Entropy and Diversity". In: *Oikos* 113, pp. 363–375.

Khawlie, M. et al. (2002). "Remote Sensing for Environmental Protection of the Eastern Mediterranean Rugged Mountainous Areas, Lebanon". In: *Photogrammetry and Remote Sensing* 57, pp. 13–23.

Landis, Douglas A. et al. (2008). "Increasing corn for biofuel production reduces biocontrol services in agricultural landscapes". In: *PNAS* 105, pp. 20552–20557.

Lange, Fabian, Alan L. Olmstead, and Paul W. Rhode (2009). "The Impact of the Boll Weevil, 1892-1932". In: *The Journal of Economic History*.

Larsen, Ashley E. and Frederick Noack (2017). "Identifying the Landscape Drivers of Agricultural Insecticide Use Leveraging Evidence from 100,000 Fields". In: *Proceedings of the National Academy of Sciences* 114, pp. 5473–5478.

Liebhold, Andrew M and Robert L Griffin (2016). "The legacy of Charles Marlatt and efforts to limit plant pest invasions". In: *Bulletin of the Entomological Society of America* 62.4, pp. 218–227.

Lin, Brenda (2011). "Resilience in Agriculture through Crop Diversification". In: *BioScience* 61, pp. 183–193.

MacDonald J.M., R.A. Hoppe and D. Newton (2018). "Three Decades of Consolidation in U.S. Agriculture". In: *U.S. Department of Agriculture, Economic Research Service*.

Manning, Dale T and Amy Ando (2022). "Ecosystem services and land rental markets: Producer costs of bat population crashes". In: *Journal of the Association of Environmental and Resource Economists* 9.6, pp. 1235–1277.

McDaniel, M D, L K Tiemann, and A S Grandy (2014). "Does agricultural crop diversity enhance soil microbial biomass and organic matter dynamics? A meta-analysis". In: *Ecological Applications* 24, pp. 560–70.

Palmu, Erkki et al. (2014). "Landscape-scale crop diversity interacts with local management to determine ground beetle diversity". In: *Basic and Applied Ecology* 15.

Patton, Nicholas R. et al. (2019). "Topographic Controls of Soil Organic Carbon on Soil-Mantled Landscapes". In: *Scientific Reports* 9.1.

Pielou, Evelyn (1966). "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13.

PRISM Climate Group, Oregon State University (2025). *PRISM Climate Data*. URL: https://prism.oregonstate.edu/.

Rohr, Jason R. et al. (2020). "Towards common ground in the biodiversity–disease debate". In: *Nature Ecology and Evolution* 4, pp. 24–33.

Schroeder, Hayley et al. (2024). "Agricultural landscape simplification affects wild plant reproduction indirectly through herbivore-mediated changes in floral display". In: *Scientific Reports* 14.

Sirami, Clélia et al. (2019). "Increasing crop heterogeneity enhances multitrophic diversity across agricultural regions". In: *Proceedings of the National Academy of Sciences* 116.33, pp. 16442–16447.

Strobl, Eric (2021). "Preserving Local Biodiversity through Crop Diversification". In: *American Journal of Agricultural Economics* 104, pp. 1140–1174.

Taylor, Charles (2022). "Cicadian Rythm: Insecticides, Infant Health, and Long-Term Outcomes". In: *CEEP Working Paper*.

Tilman, David, Kenneth G. Cassman, et al. (2002). "Agricultural sustainability and intensive production practices". In: *Nature* 418, pp. 671–677.

Tilman, David, Joseph Fargione, et al. (2001). "Forecasting Agriculturally Driven Global Environmental Change". In: *Science* 292.5155, pp. 281–284.

U.S. Department of the Interior, U.S. Geological Survey and U.S. Department of Agriculture (2013). *LANDFIRE Existing Vegetation Type Layer*. URL: https://landfire.gov/vegetation/evt.

United States Department of Agriculture and National Agricultural Statistics Service (2023). *Cropland Data Layer, 2008–2023*. USDA NASS, Washington, D.C. URL: https://nassgeodata.gmu.edu/CropScape/.

USDA-NASS (2008). *Crop production historical track records; April 2019*.

Vollering, Julien et al. (2025). "Terrain is a Stronger Predictor of Peat Depth than Airborne Radiometrics in Norwegian Landscapes". In: *EGUSphere Working Paper*.

Weitzman, Martin L. (2000). "Economic Profitability versus Ecological Entropy". In: *Quarterly Journal of Economics* 115.1.

Wilson, Margaret F.J. et al. (2007). "Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope". In: *Marine Geodesy* 30, pp. 3–35.

Winsberg, Morton D. (1982). "Agricultural Specialization in the United States since World War II". In: *Agricultural History*.

A Data and Methods

A.1 Linking the U.S. Census of Agriculture over time

The boundaries of U.S. counties have undergone many changes over time, making the linking of U.S. Census of Agriculture data challenging. In order to account for these changes, one needs to use a stable map of U.S. counties and reallocate the data—as collected in each U.S. Census of Agriculture wave—to that map.

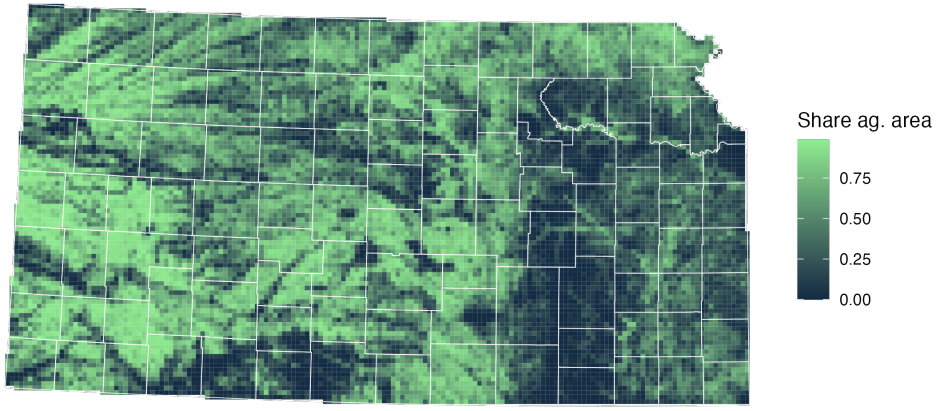
Two traditional methods exist to perform this reallocation exercise. The first method entails identifying the smallest time-invariant geographic units—which could be much larger than a county. This method is assumption-free, and therefore quite useful, but leads to significant aggregation if one wishes to look at U.S Census of Agriculture data over more than a few decades.[14]

The second method distributes the data to a selected target map assuming homogeneous distribution of the measured variables over the old geographic units. In what follows, we call this method the *area-based method*. Historically, this algorithm is preferred for ecological and economic work based on the U.S. Census or the U.S. Census of Agriculture (Hornbeck and Naidu, 2014; Crossley et al., 2020). Recently, Eckert et al. (2020) published readily available datasets of aggregation weights for applying this method to all U.S. counties since 1790. While easy to implement, this method relies on a strong spatial homogeneity assumption, unlikely to hold in the agricultural context. Figure A.1 shows the 2021 distribution of agricultural coverage in Kansas. White outlines correspond to the boundaries of the 1997 U.S. counties. Agricultural coverage is far from being homogeneous within each shape. This makes the *area-based linking method* assumption implausible, leading to mismeasurement bias when applying this method. In a regression context, this mismeasurement is likely to cause an omitted variable bias rather than a classical measurement error, and also create heteroskedasticity in the errors.

We propose a *machine-learning-based method* to relax the homogeneity assumption of the *area-based linking method* and reduce the bias in agricultural census linking exercises. We start by training a cropland coverage prediction model using fine-grid land use from the Cropland Data Layer (United States Department of Agriculture and National Agricultural Statistics Service, 2023), weather data from PRISM (PRISM Climate Group, Oregon State University, 2025) and topographic data from the Landfire dataset (U.S. Department of the Interior, U.S. Geological Survey and U.S. Department of Agriculture, 2013). This model can then be applied to any historical weather dataset to predict where crops were grown. The spatial resolution of the weather data, however, needs

[14] Horan and Hargis developed such a county crosswalk for 1840–1990, which is freely available on the ICPSR website.

Figure A.1: Spatial Distribution of Agricultural Land in Kansas—2021



Notes: Lighter shades of green indicate a higher agricultural share within the 4km x 4km pixel. White shapes correspond to the boundaries of the U.S. counties in 2021. Agricultural land is not homogeneously distributed within counties.

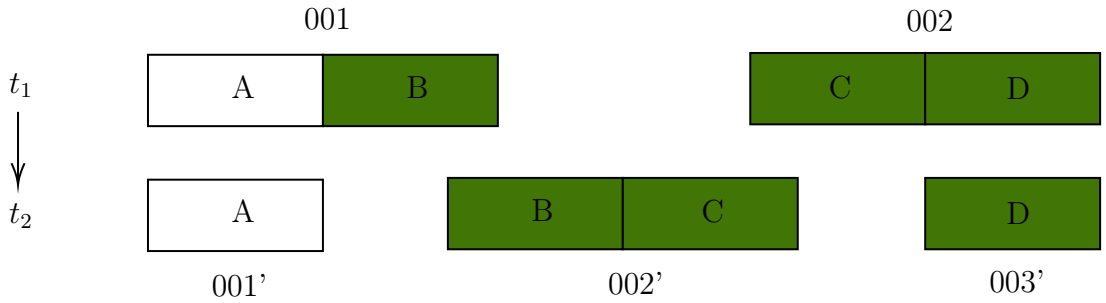
to be smaller than a county, which is the case for the historical PRISM series that we use. With land use predictions at hand, we refine the *area-based* aggregation weights, only taking into account the predicted agricultural areas to compute reallocation weights. Figure A.2 provides a stylized example comparing the outcomes of the *area-based* and the *machine-learning-based* linking methods.

A.1.1 Machine-learning-based agricultural census linking method

The *machine-learning-based* agricultural census linking method relies on the observation of the spatial distribution of agricultural area within counties. Unfortunately, data on sub-county cropland fractions do not exist for the period of interest (1880–2012). These cropland fractions can, however, be predicted at a fine scale using historical weather data, as well as topographic data. In what follows, we describe the steps of the *machine-learning-based* linking algorithm.

First, we build a model of agricultural area spatial distribution using fine-scale land use data from the Cropland Data Layer (CDL) for the period 2008–2020. Predictive features are PRISM weather data available at a 4 km x 4 km grid for the contiguous U.S., and elevation and slope data from the Landfire dataset (U.S. Department of the Interior, U.S. Geological Survey and U.S. Department of Agriculture, 2013). An XGBoost model is used to predict the presence of agriculture over each pixel. Out-of-sample, we reach an accuracy of 85.5 percent for 2021. In Table A.1, we provide a discretized confusion matrix to test the accuracy of our model. We predict the percentage of agricultural coverage in each PRISM cell for 2021, using 2021 PRISM weather data, and then compare it with

Figure A.2: Stylized example—Area-based versus Machine-learning-based methods to linking agricultural Censuses



Assume that two counties 001 and 002 at time t_1 are split in half by time t_2 and rearranged to create counties 001', 002', and 003'. Second, assume that the agricultural area is not homogeneously distributed over county 001; green shapes denote agricultural areas. We are interested in creating a time series of an agricultural stock variable X, e.g. corn acreage, and we need to reallocate measures of X at t_1 to counties observed at t_2. The *area-based linking method* is blind to the distribution of agricultural area and we get: $X_{001'} = 0.5X_{001}$, $X_{002'} = 0.5X_{001} + 0.5X_{002}$ and $X_{003'} = 0.5X_{002}$. The *machine-learning-based linking method* aims at predicting the agricultural coverage at a sub-county level and we get: $X_{001'} = 0$, $X_{002'} = X_{001} + 0.5X_{002}$ and $X_{003'} = 0.5X_{002}$. Note that while improving on the first method, the second method is not hypothesis-free: we still need to assume that within areas B, C, and D, which are the agricultural areas in the picture above, the variable X is homogeneously distributed.

Table A.1: Confusion Matrix for 2021 CDL

	0 - 25%	25% - 50%	50% - 75%	75% - 100%
0 - 25%	275511	23495	5666	238
25% - 50%	21491	15523	5902	447
50% - 75%	7349	15671	11232	1415
75% - 100%	1957	10428	19309	7975

Notes: Number of PRISM cells per bin of agricultural coverage. Rows indicate the true categorization, and columns are the predicted category. We also obtain an R^2 of 0.63 and a RMSE of 0.17. Note that assuming a homogeneous complete agricultural coverage of the U.S. would lead to an RMSE of 0.87.

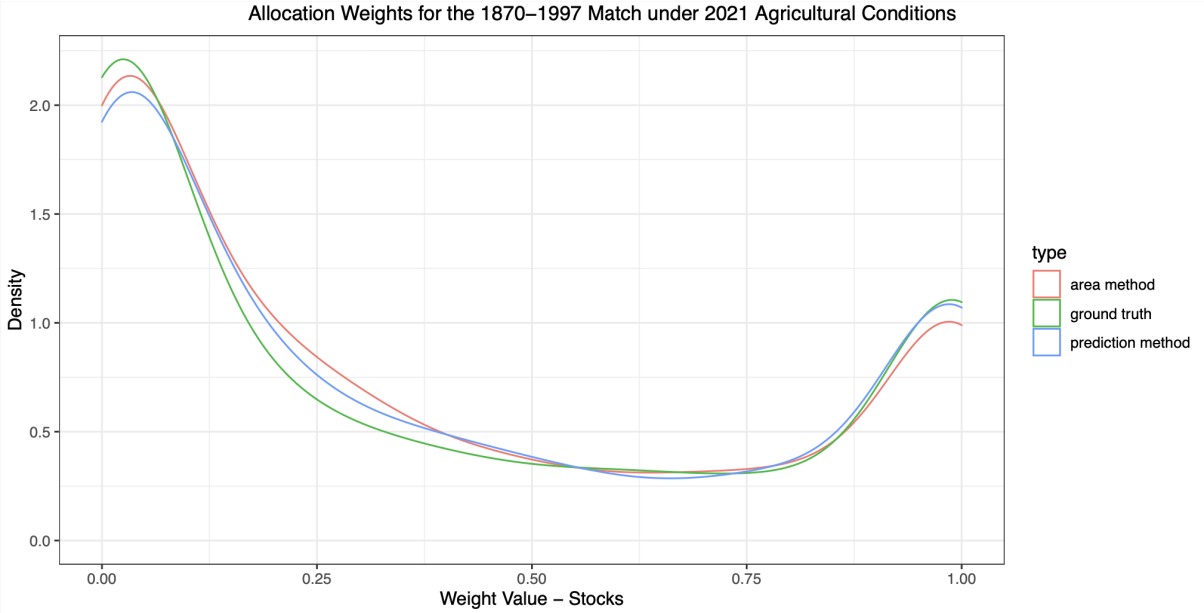
the 2021 cropland data layer. Predicted and target values are continuous. We thus bin the values to produce the confusion matrix. We see that the diagonal contains the largest number of cells and that numbers decrease as we move further off the diagonal—a sign of the model's accuracy. However, the model is not so accurate as to precisely predict the acreage in medium-range cells, with coverage ranging between 25 percent and 75%.

Second, we use this model to predict cropland presence for each U.S. census of Agriculture between 1880 and 1997. Because PRISM only goes back to 1895, we use the 1895–1924 average weather for censuses prior to 1900.

Finally, we choose the 1997 U.S. county map as our target map and intersect it with every historical county map. For stock variables[15], we use the agricultural area predictions and compute the fraction of agricultural area for all intersected areas. Reallocation weights are given by $\frac{\widehat{Ag.Area}_j}{\widehat{Ag.Area}_{\mathcal{J}:j\in\mathcal{J}}}$ where $\mathcal{J}$ is the set of counties in 1997.

[15]We note that mean variables would use the origin county total agricultural area, rather than the destination one, as a normalization. We provide both sets of weights in our dataset.

Figure A.3: Comparing the Distribution of Agricultural Fraction across Linking Methods



Notes: We compute true 1870/2021 county-level agricultural fractions. The predictions based on 1870 counties use either the area-based method or our machine-learning-based one. We note that our distribution is almost always closer to the true distribution.

A.1.2 Comparing area-based versus machine-learning-based linking methods

The CDL and other fine-grain cropland coverage data sources only exist for relatively recent years, whereas the county-level map of the U.S. mostly fluctuates in the early years of the U.S. Census of Agriculture (end of the nineteenth and beginning of the twentieth century). As such, the biases introduced by the intersected area method will be higher for U.S. Census of Agriculture data gathered in the earlier years.

To test our algorithm against the *area-based linking method*, we choose to build an example using the 1870 map of U.S. counties. We do not observe sub-county agricultural data for 1870, and thus use the fine-resolution 2021 cropland data (CDL) to construct a "counterfactual ground truth" for this 1870 map, i.e., we attribute 2021 land use patterns to 1870 counties and use this fake map as our ground truth. Our goal is then to convert this 1870/2021 map into a map of the same land cover with 1997 county boundaries.

First, we compare the accuracy performance of the two linking methods when reallocating the total cropland area. Figure A.3 plots the distribution of reallocation weights obtained using the two algorithms as well as the true weights (one for stock weights, and one for averages). The prediction method outperforms the traditional one at almost every point of the distribution. The reason for this improvement is that while most of the cells probably have positive agricultural area, most are also only partially covered by agriculture, thus making the homogeneous land cover assumption too coarse.

Second, we check the performance of the *machine-learning-based* algorithm in correctly

Table A.2: Crop-Specific Matching Improvement

	Area-based	ML-based	RMSE gain
Corn	0.09	0.01	**88%**
Wheat	0.10	0.04	**60%**
Soy	0.03	0.04	-33%

Notes: The table presents the sum of squared distances between true county crop shares, and the ones obtained using either the crosswalk assuming homogeneous distribution, or our crosswalk based on agricultural areas

reallocating crop-specific areas. We compute the mean squared errors of the two methods: first, when aggregating the data using the traditional area-based crosswalk (all area), and second when using the agricultural area-based crosswalk. Results are presented in Table A.2. Our method outperforms the area-based one for corn and wheat, and slightly under-performs for soy, where error margins are smaller. The variation in performance across crops is due to the relative spatial homogeneity of counties where these crops are grown. It appears that soy is grown more homogeneously in counties that cultivate soy, relative to wheat and corn.

B Descriptives

Table B.3: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Gini-Simpson	6,315	0.52	0.12	0.00	0.82
Ruggedness	6,315	1.32	1.22	0.06	10.99
Boll Weevil Presence	4,043	0.33	0.47	0	1
Distance to Rail	6,315	9	14	0	139
Minimum Temperature	6,315	10.75	2.37	−0.23	17.83
Maximum Temperature	6,315	23.73	2.19	15.35	31.53
Rainfall	6,315	1,104	335	77	2,199
County Agricultural Area	6,315	268,517	210,515	645	3,580,524

Notes: The tables show summary statistics for the main variables used in the analysis in the paper. Distance to rail is expressed in kilometers. Rainfall is expressed in mm, temperatures in °C. Agricultural area is expressed in acres. See Section 2 for details on data sources.

C Robustness

Table C.4: Diversity and Boll Weevil Diffusion – Alternative Crop Diversity Index

Dependent Variable:	Boll Weevil Presence			
Model:	OLS	IV	OLS	IV
Crop Diversity (Pielou Index)	-0.0159	-2.695***	-0.0061	-2.275***
	(0.0576)	(1.0430)	(0.0574)	(0.7592)
State × Decade FE	Yes	Yes	Yes	Yes
Observations	3,895	3,895	3,895	3,895
Number of Clusters	1,080	1,080	1,080	1,080
R^2	0.8608	0.6973	0.8608	0.7479
Mean Boll Weevil	0.33	0.33	0.33	0.33
SD Boll Weevil	0.47	0.47	0.47	0.47
Mean Pielou Index	0.40	0.39	0.40	0.39
SD Pielou Index	0.09	0.05	0.09	0.05
First Stage (Ruggedness)		-0.008 (0.002)		-0.008 (0.002)
Mean Ruggedness		1.34		1.34
SD Ruggedness		1.27		1.27
F-statistic		31.8		49.0

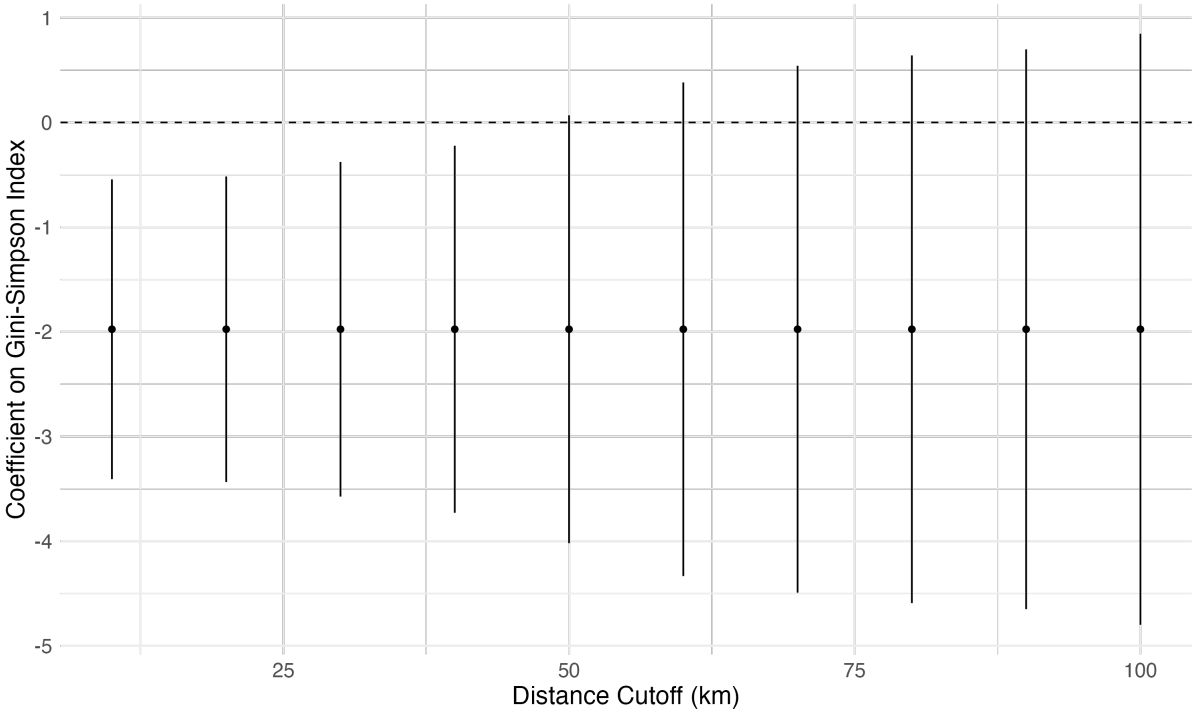
Notes: This table reports results from OLS and IV estimation of the effect of crop diversity on boll weevil contamination, measured by the Pielou index. The sample contains observations of counties in the cotton belt between 1890 and 1930. The dependent variable is equal to 1 if a county is affected by the boll weevil in decade t, and 0 otherwise. In all specifications, we control for state by decade fixed effects. In columns (1) and (2), we control for the share of cotton in agricultural acreage in 1880, agricultural acreage in 1880, average rainfall, average daily minimum and maximum temperature. In columns (3) and (4), we additionally control for the richness index measured in 1880. Standard errors are clustered at the county level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table C.5: Diversity and Boll Weevil Diffusion—Alternative Fixed Effects

Dependent Variable:	Boll Weevil Presence		
Model:	(1)	(2)	(3)
Crop Diversity (Gini-Simpson)	-1.975***	-1.874***	-1.773***
	(0.7345)	(0.6064)	(0.6065)
State × Year FE	Yes		
State FE		Yes	Yes
Year FE		Yes	Yes
State × Year Time Trend			Yes
Observations	3,895	3,895	3,895
Number of Clusters	1,080	1,080	1,080
R^2	0.724	0.573	0.651
Mean Boll Weevil	0.33	0.33	0.33
SD Boll Weevil	0.47	0.47	0.47
Mean Gini-Simpson	0.53	0.53	0.53
SD Gini-Simpson	0.06	0.05	0.06
First Stage	-0.011 (0.003)	-0.010 (0.003)	-0.011 (0.003)
Mean Ruggedness	1.32	1.32	1.32
SD Ruggedness	1.22	1.22	1.22
F-statistic	37.4	44.3	45.7

Notes: This table reports results from the IV estimation of the effect of crop diversity on boll weevil contamination, measured by the Gini-Simpson index. The sample contains observations of counties in the cotton belt between 1890 and 1930. The dependent variable is equal to 1 if a county is affected by the boll weevil in decade t, and 0 otherwise. In all specifications, we control for state by decade fixed effects. We control for the share of cotton in agricultural acreage in 1880, agricultural acreage in 1880, average rainfall, average daily minimum and maximum temperature. We vary the set of fixed effects included in the three specifications, the one shown in column (1) being our baseline specification. Standard errors are clustered at the county level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Figure C.4: Crop Diversity and Pest Diffusion: Alternative Standard Errors



Notes: This figure presents the robustness of our results to Conley standard errors accounting for spatial auto-correlation, varying the distance threshold for the standard errors.