Local Crop Species Diversity and Pest Diffusion: Evidence from the US Census of Agriculture*

Tristan Du Puy, Marguerite Obolensky†

January 2024

Abstract

We provide evidence of the role of local agricultural crop diversity, measured by the number of crops grown locally and the homogeneity of their land allocation, in hampering the diffusion of pests. First, we build consistent county-level diversity measures using a new machine-learning-based method to link the US Census of Agriculture between 1880-2007. Second, we show large declines in local crop diversity over the second half of the 20th century, consistent with previous findings. Finally, we examine the impact of crop diversity on the spread of two significant pest outbreaks in US agricultural systems—the boll weevil (1890-1930) and the imported fire ant (1940-1997). We address reverse causality concerns by instrumenting local crop diversity with the pre-planting expected standard deviation in crop revenues. To the best of our knowledge, this constitutes the first causal inference study of the role of crop diversity on pest diffusion. We find that lower local crop diversity favored the diffusion of these two pests.

*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 of our project. All remaining errors are ours.

†Du Puy: School of International and Public Affairs, Columbia University (td2631@columbia.edu); School of International and Public Affairs, Columbia University (m.obolensky@columbia.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 interest.

1 Introduction

Transformations in agricultural production processes have been a major driver of land use changes. Increased reliance on irrigation, pesticides, and fertilizers impact local climate [1], the sustainability of water resources [2], and lead to significant air, soil, and water pollution [3] [4]. These shifts in land management have also affected the structure and layout of rural landscapes.

We study the consequences of the homogenization of cropland on the spatial diffusion of agricultural pests in the US over the twentieth century. Pest shocks are extremely harmful to the stability of ecosystems and can lead to significant agricultural losses [5] [6] [7]. Pesticides are efficient at curbing pest-related risks, but pests develop product-specific resistance over time [8], and exposure to pesticides causes significant harm to human and non-human health [9] [10] [11] [12] [13] [14]. Understanding the determinants of pest diffusion is therefore important to finding alternatives to pesticide use and improving the resilience of agricultural systems.

Our analysis spans two major pest shocks of the twentieth century in the U.S. The U.S. provides an ideal setting for studying the role that crop diversity plays in pest diffusion for several reasons. First, the United States government has consistently recorded highly detailed agricultural data starting in the late nineteenth century. We construct two measures of crop diversity at the county level over time: the number of crops grown locally (crop richness) and the homogeneity of land allocation across the crops grown (Gini-Simpson index or crop evenness).[1] Second, the land management practices of American farmers have transformed during the twentieth century. As a result, the average farm size increased from 150 acres in 1920 to 446 acres in 2022 and croplands have become more spatially homogeneous [18] [19] [20]. Third, meticulous records of pest diffusion were created throughout the twentieth century. The diffusion of the boll weevil between 1890 and 1930 was monitored annually [21]. A few years later, in 1940, the imported fire ant started proliferating in the Southeast, and the county-level quarantine records allow us to track its diffusion [22].

We show that higher crop diversity hampers the local diffusion of the boll weevil and the imported fire ants using an instrumental variable approach. To the best of our knowledge, this constitutes the first causal inference study for this mechanism. This instrumental variable strategy could be used to study further consequences of local cropland homogeneization. Results from our preferred specification indicate that increasing the crop richness (resp. evenness) of a county such that it moves from the 25^{th} to the 75^{th} percentile of the crop richness distribution would reduce the yearly likelihood of contam-

[1] Using multiple diversity indices is in line with the recommendation of the literature on the local biodiversity [15] [16] [17].

ination by the imported fire ant by 64% (resp. 22%)[2]. In the case of the boll weevil, a county moving from the 25th to the 75th percentile of the evenness distribution would see its likelihood of contamination in a given year decreased by 71%. We do not find any effect of crop richness on the boll weevil diffusion, in part due to a lack of identifying power. The mechanisms through which crop diversity acts on pest diffusion most likely differ between the two considered pests, explaining in part the difference in effect size between pest events.[3] The boll weevil feeds mainly on cotton and is thus negatively impacted by the fragmentation of cotton fields [23]. By contrast, the omnivorous imported fire ant is less likely to depend on the spatial contiguity of specific crops. Factors hampering its progress relate to competition for niches in a given ecosystem [24]. County-level crop richness and evenness positively correlate with both land fragmentation and the competitiveness of ecological niches, explaining why we find they slow down the diffusion of pests locally in most of our specifications.[4]

There are many threats to identification in this context, which may persist even after accounting for systematic differences between counties and years. For example, areas of intense agriculture are both more likely to have homogeneous cropland and rely heavily on pesticides. This may downward bias the correlation between crop diversity and pest invasion. We address these concerns by instrumenting local crop diversity with the standard deviation in pre-planting expected agricultural revenues across crops. Profit-maximizing farmers optimize their crop mix and, everything-else-equal, grow fewer crops when the distribution of expected revenues across crops grows wider. The spread of expected revenues is therefore correlated with local crop diversity, making the instrument valid. Additionally, after controlling for county and decade fixed effects, variation in the instrument is due to national shocks in crop demand and is exogenous from local drivers of pest diffusion.

When examining the relationship between crop diversity and pest diffusion, a key concern is that pests may travel with commodity trade. The boll weevil appeared in the southeast of the United States at a time when the rail was being developed. However, previous research has established that the diffusion of the boll weevil was not significantly impacted by local trade networks, but rather by wind patterns [5]. On the contrary, the diffusion of the imported fire ant is known to have been facilitated by commodity trade, especially to California [6]. Our results are robust to dropping the state from the sample as well as controlling for the evolution of the United States interstate highway network over our period of analysis.

[2]The spread in richness over counties is larger – 4 crops difference between the 25th and 75th centiles, than for the Gini-Simpon index, with resp. 0.31 and .58.

[3]A formal comparison of the effect size across pest events is difficult due to differences in sample composition.

[4]We explain the lack of responsiveness of the boll weevil to the richness index by the lack of variation in crop mix in the early twentieth century United States South – and thus the lack of statistical power.

Biodiversity estimates are heavily influenced by sampling variation. This implies that cross-space and time comparisons require consistency in sampling methods, ideally at a local scale. In the absence of fine-grained data on landscape complexity and biodiversity over our period of study, the Census of Agriculture is a remarkable source of county-level data given its completeness and relative homogeneity over time. To further improve the comparison across census waves, we develop a new method to link the different census waves over time. For each county shape for which census data was ever collected, we predict the within-county location of cropland areas at the data collection moment. We do so using flexible functions of historical climate and topography. We use this spatial distribution of agricultural activity to construct aggregation weights and reallocate the data collected in various waves to a stable map of US counties. This allows for the comparison of county-level observables across time. Our method relaxes the usual assumption that agricultural areas are homogeneously distributed within a county. We show evidence that our algorithm outperforms traditional area-based linking algorithms [25].

This paper contributes to a growing literature documenting the consequences of biodiversity loss. The decline in species richness in managed agricultural environments has been shown to allow fungal pathogens to occupy additional ecological niches and reduce dilution effects [26], as well as to increase the use of pesticides [27]. In line with our results, recent research shows that larger agricultural areas and larger agricultural fields are associated with more intense pesticide use [28] [29]. Noncrop habitat has however been shown to have no systematic relation with pest suppression [30]. Our paper also relates to the epidemiology literature that has established that more diverse ecological communities are more immune to diseases [31][32][33]. To the best of our knowledge, this paper is the first to use causal inference methods to show how high crop diversity can slow down the spread of pests. Second, we contribute to the literature analyzing the consequences of economic activity on biodiversity. Recent examples include the study of the drivers of the collapse of vulture populations in India [34], and the bison population in the Great Plains of the United States [35]. Liang et al [36] also show how local economic shocks in production negatively impact species abundance, diversity, and stability by increasing air pollution. We build on these findings to construct an instrumental variable for local crop diversity. Finally, we make a methodological contribution by developing a new data-driven approach to link the 1840-2007 United States Census of Agriculture over time, accounting for the changing county boundaries for which the Census data was collected. Our new dataset replicates the declining trends previously established by the literature [20].

2 Methodology

2.1 Measuring Pests Diffusion

Boll Weevil The boll weevil is an insect native to Mexico and Central America. It entered the United States in 1892 through Brownsville, Texas, and later expanded to the entire cotton belt [5]. The lack of natural enemies and effective pesticides against the boll weevil in the United States until 1918 favored its expansion in cotton fields where it caused significant damages.[5]

We get boll weevil contamination data from historical maps [21]. Panel (a) of Figure 1 displays the year of arrival of the boll weevil in each county of the United States South. The diffusion follows a concentric pattern, consistent with the evidence of the spread of the boll weevil following dominant winds [5].

Imported Fire Ants The two species of imported fire ants we study arrived in the United States around the late 1930s. They have since spread through the United States South, and more recently to some areas in California (Panel (b) Figure 1). The ants originate from South America and know few natural enemies in the United States [6]. The fire ants are omnivorous and can significantly affect local ecosystems by displacing other species of ants and invertebrates through competition for resources, as well as inflicting damage to reptiles, mammals, and ground-nesting birds.[6]

Starting in 1940, the United States Department of Agriculture (USDA) published county-level quarantine decisions to help prevent the spread of the imported fire ant [22]. Our data is such that we know the date of the introduction for each county-level quarantine. All the quarantines introduced are still active today. We assume that the date of introduction of the quarantine is a reasonable proxy for the arrival of the ants and that once contaminated the counties remain contaminated.

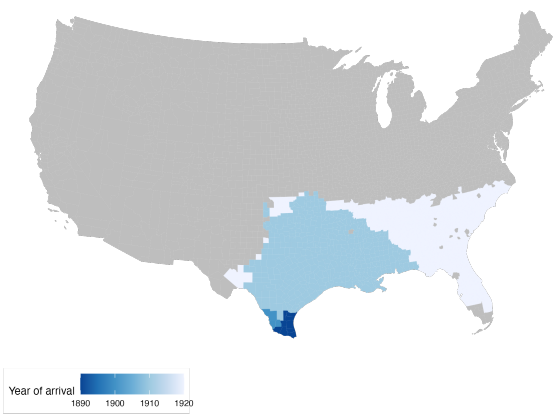
In the 2000's the USDA started a campaign of information among truck drivers to limit the accidental spread of the ants. Simultaneously, pesticides were developed to curb the invasion. Because these adaptive measures cannot be fully accounted for in our models, we focus on the pre-2000 period and show that results hold when accounting for changes in transportation networks.

[5] It is estimated that after five years of contact, a contaminated county's production would have declined by 50%. Impacted counties experienced significant out-migration, and by transforming the prevalence of tenant farming, the boll weevil even reduced the rates of marriage among young African Americans in the South [21].

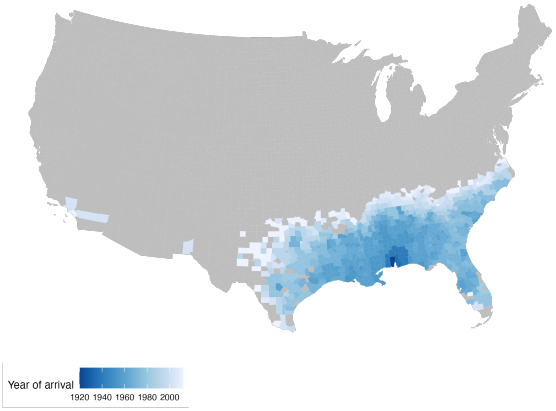
[6] The ants are also harmful to agricultural workers, working in crops that are manually harvested like vineyards, orchards, and vegetable field crops. Their poison causes a burning sensation, and in a few cases can cause life-threatening anaphylactic shocks. Additionally, the ants can cause extensive damage to irrigation lines, electrical equipment, and harvesting and mowing equipment by creating large mounded nests in the middle of fields. The state of California estimated that a state-wide spread of the ants could generate yearly damages ranging from $387 to $989 million.

Figure 1: Pests Contamination Over Time

(a) Boll Weevil



(b) Imported Fire Ant



Notes: Data comes from the digitized map of the spread of the boll weevil, from Bloom et al (2017) [21], and the USDA Aphis website [22] for the imported fire ant.

2.2 Measuring Crop Diversity

Biodiversity is a complex concept. It can be measured at various scales in a given ecosystem and can be defined in different ways.[7] In this paper, we focus on local crop diversity. While other variables related to biodiversity are important to explain the spread of pests, we focus on crop diversity at the county level for several reasons. First, crop diversity can be consistently measured throughout the twentieth century in the United States. Second, aggregated measures of crop diversity are correlated with more granular variables also related to biodiversity [38] [39]. Third, crop diversity is directly influenced by agricultural practices that respond to changes in the economic environment across space and over time [40] [41].[8] This measure allows us to test by proxy the links between landscape configurations related to human activity and pest diffusion.

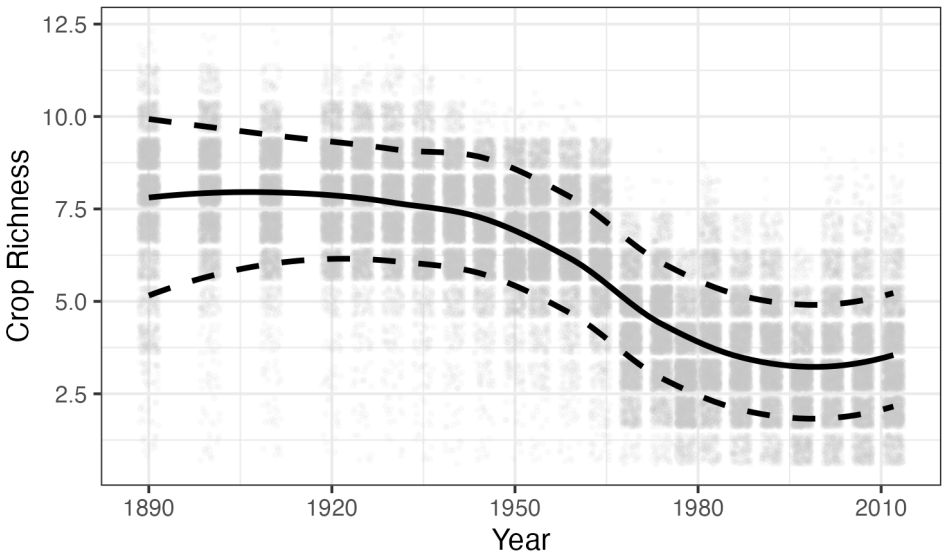
We construct two county-level measures of crop diversity between 1880 and 2012 using the Census of Agriculture [42]. The richness index counts the number of distinct crops grown in a county. The Gini-Simpson index or crop evenness is the probability that two acres sampled at random within a county are covered with different crops. This second index ranges from 0 (two random acres are systematically covered with the same crop) to 1 (all acres are planted with distinct crops).[9] Appendix A.1 describes the diversity indices in detail.

[7] Biodiversity is a concept that covers different realities, from phenotypical diversity, and genetic diversity, to species diversity. Diversity can also correspond to the diversity of a specific community or ecosystem, or measure the differences across them. This becomes an empirical issue given that these different dimensions often have a low statistical correlation [37].

[8] For example, crop diversity is correlated with decreased pesticide use [28] [29] and land fragmentation.

[9] These two diversity indices are amongst the most frequently used. Another popular index, the Pielou index, is highly correlated with the Gini-Simpson index (correlation of 0.97) and is omitted from the analysis.

Figure 2: County-Level Richness over Time



Notes: The solid line represents the average value in every wave of the Census. The dotted lines represent the 10th and 90th percentile of the county-level values for the same wave.

Constructing both the richness and evenness indices requires a precise time series of local cropland coverage. We develop a new method to link the different waves of the census together, to account for the changing county geographic boundaries. Our machine learning algorithm uses historical weather, as well as topography data to predict the location of agricultural areas within counties and over time, to reallocate all observed census data to a common map of United States counties. Appendix A.2 provides more detail on the methodology. We provide evidence in the appendix that our method outperforms traditional geographical census linking methods.

We obtain a balanced panel of planted acres for 12 crops, including the most common crops in the US: corn, soybean, wheat, and cotton.[10] The data are county-level observations every five to ten years, depending on the frequency of the Census waves. We can confirm previous findings of large-scale decreases in local crop diversity in the twentieth century [20]. In Appendix C, we provide more context for these decreasing trends. The increase in crop concentration is largely attributed to the reduction in the size of the crop mix, rather than the reallocation of land within a fixed set of crops. Concentration is jointly caused by the growing importance of corn and wheat in the United States agricultural system, and the decline of oats, rye, barley, and potatoes, generating the sharp decline in crop richness starting in the 1950s (Figure 2)[11].

[10] The complete list of crops included is the following: rice, cane sugar, wheat, corn, cotton, potato, sweet potato, barley, rye, tobacco, buckwheat, and oats. Results are robust to using a larger set of crops in an unbalanced panel.

[11] Our favored set of crops over which we measure diversity does not include soy, which is only accounted for in the Census from 1920 onwards. Hence, our measures are orthogonal to the evolution of soy acreage. We show in the annex that an expanded index also accounting for soy is very highly correlated with soy.

2.3 Identification

Using the data described in Section 2.1 and 2.2, we obtain a balanced panel of all contiguous United States counties, recording pest contamination and the diversity of cropland areas over time. One empirical challenge is that crop diversity is endogenous. Factors like soil quality, weather patterns, pest management practices, or transportation infrastructure are susceptible to influence both the pest diffusion and the crop mix of a given county.

We address these endogeneity concerns with an instrumental variable approach combined with a two-way fixed effects model. The inclusion of county and decade fixed effects in the regression removes the effects of confounding variables stable respectively across space and time. Some of the confounders listed above, however, are not absorbed by the fixed effects. The goal of the instrument is to shift county-year-level crop diversity exogenously to estimate the causal impact of crop diversity on pest contamination.

We use the variance of the distribution of crop revenues to instrument for crop diversity at time t. We expect that a profit-maximizing farmer – everything else equal – will cultivate fewer crops more intensively in counties and years characterized by an increase in revenue variance across crops. Figure A.5 shows that as the standard deviation of expected crop revenues increases, diversity decreases, confirming our hypothesis.

The distribution of expected county-year-level revenues is obtained by multiplying national prices with local potential yields for all crops. We obtain national-level crop prices by averaging state-level prices reported in the census for all twelve crops in our sample. Next, we compute crop-specific county-level theoretical output densities from the FAO GAEZ models [43]. We choose to use a modeled measure of potential yields rather than reported yields. The reason is that we want to measure expected revenues for all twelve crops in our sample, for all US counties – and not only for the sample of crops that were effectively planted in any given county. This ensures that the instrument captures all the information that may have incentivized farmers to diversify their plantations.

For the instrumental variable to be valid, it needs to satisfy the exclusion restriction, meaning that the impact of the instrument on pest diffusion must only be mediated by crop diversity. We argue that the standard deviation in expected agricultural revenues meets this criterion. The first component of the instrument, US crop prices, are determined by the global demand and supply of crops. After controlling for decade fixed effects, national prices are plausibly orthogonal to the local diffusion of pests. We pay particular attention to the impact of the boll weevil on cotton prices. Cotton prices at the national level remained globally stable between 1890 and 1915, the first twenty-five years of the boll weevil crisis. Prices only spiked in 1920, following what the literature

has considered as a US-wide commodity price crisis after the end of World War I [44].[12] We then consider that the use of US-level prices, rather than state or county level, is sufficient to recover price fluctuations that are exogenous to the pest shocks. Local expected yields constitute the second part of the instrument. Taken from the FAO GAEZ models, they are based on climate and topographic observations combined with agronomic models. These yields are precisely exogenous because they are not built using any relevant information that would be associated with the diffusion of our pests, and not captured by fixed effects. We compute expected output densities in low input and no irrigation scenarios for all crops and counties.

Despite the inclusion of fixed effects, and our reliance on a plausibly exogenous instrumental variable, transitory and local shocks may still bias the estimation of the effect of crop diversity on pest diffusion. To address such concerns, we also control for county-level total cropland area. This should account for the propensity of a county to be contaminated by a pest. Additionally, we test the robustness of our findings in an additional exercise: controlling for the expansion of respectively the railway and the highway networks (see 3).

3 Results

Our two-stage least-square approach amounts to estimating:

$$\text{Crop Diversity}_{ct} = \alpha + \gamma_1 Z_{ct} + X_{ct}\gamma_2 + \eta_c + \eta_{d(t)} + \varepsilon_{ct} \qquad \text{First Stage} \quad (1)$$

$$\text{Pest Diffusion}_{ct} = \alpha + \beta_1 \widehat{\text{Crop Diversity}}_{ct} + X_{ct}\beta_2 + \eta_c + \eta_{d(t)} + \nu_{ct} \qquad \text{Second Stage} \quad (2)$$

where Z_{ct} is the instrument, η_c are county fixed effects, $\eta_{d(t)}$ are decade fixed effects for Census year t and X_{ct} corresponds to a vector of controls. In our main specification, X_{ct} is the county-level total cropland area. The variable Crop diversity$_{ct}$ corresponds either to crop richness or to crop evenness. Pest Diffusion$_{ct}$ is a dummy equal to 1 if county c at period t is contaminated by the pest. Contamination is an absorbent state in both cases. Standard errors are heteroskedastic robust.

Columns 1 and 2 of Table 1 present the results of the first stage in the boll weevil sample. The instrument is significantly and negatively correlated with the Gini-Simpson index (column 2). This is consistent with the hypothesis that profit-maximizing farmers choose to grow a smaller set of crops when expected revenues across crops become more

[12] This spike in cotton prices has been specifically discussed as part of the general commodity crisis.

dispersed. An increase by one standard deviation in the spread of crop revenues at the county level will decrease crop evenness by 0.011, or 2% of its average value.[13] We cannot detect a statistically significant correlation between the spread of crop revenues and crop richness in a county (column 1), due to the relative homogeneity in crop mix across boll weevil-affected counties over 1890-1920. We will not interpret the boll weevil richness results because of this lack of power. In the case of the imported fire ant (columns 3 and 4), both indices of local crop diversity show a statistically significant negative relation with the cross-crop standard deviation in expected revenues. An increase by one standard deviation in the spread of revenue will decrease richness by .14, or 2.6% of its average value, and evenness by .02, or 5% of its average value.

Table 1: Diversity and Pest Diffusion: Revenues IVs (First Stage)

	Richness (1)	Gini-Simpson (2)	Richness (3)	Gini-Simpson (4)
Std. Revenue per Ha	0.0003	-0.0003**	-0.0002***	-3.21×10^{-5}***
	(0.0009)	(0.0001)	(2.54×10^{-5})	(3.84×10^{-6})
Observations	4,043	4,043	25,052	25,052
Counties	1142	1142	3018	3018
R^2	0.81153	0.60860	0.81460	0.57231
F-test (1st stage)	0.13432	7.8474	40.542	88.271
County fixed effects	✓	✓	✓	✓
Decade fixed effects	✓	✓	✓	✓
Sample	1895-1930	1895-1930	1940-1997	1940-1997

Notes: This table contains the first stage results from the instrumental regression shown in Table 2. Our instrumental variable corresponds to the county-level standard deviation in crop revenue. All regressions control for county-level total agricultural area, as well as county and decade fixed effects. The samples are limited to 1890-1930 for the boll weevil, and to 1940-1997 for the imported fire ant. Standard errors are heteroskedasticity robust. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

We report results in Table 2. In columns 1 and 3 of Panels A and B, we control for decade and county fixed effects, as well as total planted area. Columns 2 and 4 report our preferred specification with the inclusion of the instrumental variable. This specification absorbs any temporal and location-specific factors that might be correlated with crop diversity. Additionally, in panel A column 4 as well as panel B columns 2 and 4, the instrument effectively carves out some exogenous variation from the crop diversity indices (Table 1).

While the IV results – when significant – always show a negative relation between local crop diversity and pest diffusion, this is not the case for the OLS results. Specifically, for the imported fire ants, the OLS regressions show a positive statistically significant

[13] The standard deviation of the instrument over the period 1890-1920 is 55.

relation between diversity and spread. We interpret this as the presence of an omitted variable bias that distorts the sign of the results. For the more recent periods, where pesticide use is more frequent, it is more likely that low-diversity counties are also more pesticide-intensive and more protected from the spread of imported fire ants. This bias is less likely in the case of the boll weevil, as up to the 1920s, no pesticide was efficient at curbing its spread, and the take-up of pesticides in agriculture is largely a post-WW2 phenomenon.

Overall, crop diversity is significantly and negatively associated with the spread of both pests. Increasing the evenness of a county's crop allocation by 0.1 reduces the likelihood of contamination in a given year by the boll weevil by 61%. This large effect may be explained by the specialized nature of the pest and two competing mechanisms: 1) a less frequent encounter with cotton for any diffusion path through the landscape (more fragmented land, less spatial connection in the location of cotton) and 2) a more fragmented land which might leave fewer opportunity for the boll weevil to hibernate over the winter. We cannot differentiate between these two channels and our results should be interpreted as their joint impact.

The imported fire ant is slowed down by both crop richness and evenness: one additional crop type grown in a county reduces the likelihood of imported fire ant contamination by 16%, while a 0.1 increase in crop evenness decreases the likelihood by 8.3%. Both richness and evenness act as proxies for the fragmentation of landscapes, which comes with increased competition for ecological niches, and act as a barrier to the imported fire ant diffusion.

Robustness One may still be worried about the omission of a variable influencing both crop diversity and our outcome of interest, biasing our results. For example, the development of the interstate highway network between 1940 and the late 1970s [45] poses a threat to identification, as the imported fire ants are known to have traveled with agricultural commodities.[14] We expect better coverage by transportation networks to induce both a higher likelihood of pest contamination as well as a higher crop diversity – as farmers can take advantage of easier market access. This down-biases the estimates of the effect of crop diversity on pest diffusion. While transportation networks are less of a concern in the case of the boll weevil – as its spread mostly followed wind patterns [5] – the expansion of the rail network might still pose a threat. Table B.4 in the Appendix B presents results where we control for respectively the expansion of the rail and the interstate highway networks.

Specifically, in Panel A, we include the distance between the centroid of a county and the

[14]The rail network developed mostly before 1930, and market access via rail is absorbed by county fixed effects in the imported fire ants sample.

Table 2: Diversity and Pest Diffusion

Panel A	Boll weevil (0/1)			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Richness	-0.0423***	5.200		
	(0.0068)	(16.46)		
Gini-Simpson			-0.4499***	-6.137*
			(0.0704)	(3.134)
Observations	4,043	4,043	4,043	4,043
Counties	1,142	1,142	1,142	1,142
R^2	1.0000	0.99964	1.0000	0.99999
F-test (1st stage)		0.13432		7.8474
County fixed effects	✓	✓	✓	✓
Decade fixed effects	✓	✓	✓	✓
Panel B	Imported Fire Ants (0/1)			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Richness	0.0074***	-0.1595***		
	(0.0016)	(0.0430)		
Gini-Simpson			0.1649***	-0.8264***
			(0.0131)	(0.2252)
Observations	25,052	25,052	25,052	25,052
Counties	1,988	1,988	1,988	1,988
R^2	0.59843	0.40394	0.60129	0.48373
F-test (1st stage)		40.542		88.271
County fixed effects	✓	✓	✓	✓
Decade fixed effects	✓	✓	✓	✓

Notes: This table contains the second stage results from the instrumental regression. We instrument crop diversity with county-decade-level standard deviation in expected crop revenue. All specifications control for county-level total agricultural area, as well as county and decade fixed effects. In Panel A, the sample corresponds all counties growing cotton between 1890-1930 for the boll weevil. In Panel B, we include observations for all counties between 1940 and 1997. The standard errors shown are heteroskedasticity robust. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

nearest rail segment. In Panel B, we control for the distance of the county to the nearest interstate segment. Results are broadly unchanged. In the imported fire ants sample, instrumental variable estimates are larger (in absolute value) than in Table 2, suggesting that the highway network was biasing the effect towards zero as expected.

We replicate the analysis for the imported fire ant after removing the counties located in California and Arizona. Results are presented in Table B.5. These counties were likely contaminated through commodity trade, and as such the relation between their

agricultural diversity and likelihood of contamination is not informative of the channels we want to bring forward. The OLS coefficients are almost identical to our main results, the ones of the IV are slightly smaller in absolute value, moving from -.1595 to -.1459 for richness, and from -.8264 to -.7876 for the Gini-Simpson index, and still highly significant.

Finally, in Figure B.6, we test the robustness of the results after controlling for spatial correlation. Results generally stay significant for small Conley threshold values (30km). However, coefficients become noisy when considering higher spatial correlation thresholds.

4 Discussion

In this paper, we develop a novel method to link the United States Census of Agriculture over time. We use this data to corroborate recent results showing significant declines in local crop richness and evenness in the United States over the twentieth century. We then document the causal relationship between local crop diversity and pest diffusion for two of the most important United States agricultural pest shocks: the boll weevil (1890-1920), and the imported fire ant (1940-1997). A higher level of crop evenness slows down the progression of the boll weevil, a specialized pest with few natural predators. One potential mechanism behind this relationship is the role played by the fragmentation of agricultural landscapes, specifically the lesser spatial contiguity of cotton fields. Both crop richness and evenness hinder the diffusion of imported fire ants, an omnivorous pest. The diversity indices serve as effective proxies for the local saturation of ecological niches, explaining the challenging implantation of imported fire ants in this context.

To our knowledge, we are the first to estimate the effects of crop diversity on pest diffusion. We also advance and complement different strands of the literature. First, we contribute to the growing literature studying the consequences of the sharp decline in local crop diversity in the United States. Second, we bring a new perspective to the analysis of the development of intensive agriculture over the twentieth century, bringing evidence that it may have contributed to the increased vulnerability to pest contamination. Finally, we add to the literature linking biodiversity to the limited spread of diseases and pests.

Our findings open the door to several interesting questions. Our results hint at the important role of pest characteristics – such as their specialization, the spatial distribution of their enemies and hosts, and their favored modes of transport – on pests' diffusion patterns. More research is needed to understand the mechanistic interactions between crop diversity, biodiversity more generally, and pest contamination. This understanding is essential for enhancing cropland resilience to pest contamination.

References

[1] T. Braun and W. Schlenker, "Cooling externality of large-scale irrigation," *NBER Working Paper*, 2023.

[2] M. Hornbeck and P. Keskin, "The historically evolving impact of the ogallala aquifer: Agricultural adaptation to groundwater and drought," *American Economic Journal: Applied Economics*, 2014.

[3] M. Sebilo, B. Mayer, V. Nicolardot, and A. Mariotti, "Long-term fate of nitrate fertilizer in agricultural soils," *PNAS*, 2013.

[4] S. Stehle and R. Schulz, "Agricultural insecticides threaten surface waters at the global scale," *PNAS*, 2015.

[5] F. Lange, A. L. Olmstead, and P. W. Rhode, "The impact of the boll weevil, 1892-1932," *The Journal of Economic History*, 2009.

[6] K. M. Jetter and J. H. Klotz, "Eradication costs calculated: Red imported fire ants threaten agriculture, wildlife and homes," *california Agriculture*, 2002.

[7] H. Druckenmiller, "Estimating an economic and social value for healthy forests: Evidence from tree mortality in the american west," *Working Paper*, 2020.

[8] M. Crossley, W. E. Snyder, and N. B. Hardy, "Insect–plant relationships predict the speed of insecticide adaptation," *Evolutionary Applications*, 2021.

[9] M. Dias, R. Rocha, and R. R. Soares, "Down the river: Glyphosate use in agriculture and birth outcomes of surrounding populations," *The Review of Economic Studies*, 2023.

[10] C. Taylor, "Cicadian rythm: Insecticides, infant health, and long-term outcomes," *CEEP Working Paper*, 2022.

[11] E. Frank and C. Taylor, "The "golden age" of pesticides? trade-offs of ddt and health in the us," *Working Paper*, 2022.

[12] J. Calzada, M. Gisbert, and B. Moscoso, "The hidden cost of bananas: Pesticide effects on newborns' health," *Working Paper*, 2021.

[13] D. Vieira, A. Franco, D. D. Medici, D. M. Jimenez, P. Wojda, and A. Jones, "Pesticides residues in european agricultural soils: Results from lucas 2018 soil module," *Publications Office of the European Union*, 2023.

[14] L. Beaumelle, L. Tison, N. Eisenhauer, J. Hines, S. Malladi, C. Pelosi, L. Thouvenot, and H. R. Philips, "Pesticide effects on soil fauna communities - a meta-analysis," *Journal of Applied Ecology*, 2023.

[15] K. S. Prendergast, "Beyond ecosystem services as a justification for biodiversity conservation," *Austral Ecology*, 2020.

[16] M. Freitag, N. Hölzel, L. Neuenkamp, F. van der Plas, P. Manning, A. Abrahão, J. Bergmann, R. Boeddinghaus, R. Bolliger, U. Hammer, E. Kandeler, T. Kleinebecker, K.-H. Knorr, S. Marhan, M. Neyret, D. Prati, G. le Provost, H. Saiz, M. van Kleunen, D. Schäfer, and V. H. Klaus, "Increasing plant species richness by seeding has marginal effects on ecosystem functioning in agricultural grasslands," *Journal of Ecology*, 2023.

[17] D. R. Schoomaster, C. R. Zirbel, and J. P. Cronin, "A graphical causal model for resolving species identity effects and biodiversity–ecosystem function correlations," *Ecology*, 2020.

[18] B. Gardner, *American Agriculture in the Twentieth Century.* Harvard University Press, 2002.

[19] K. Kassel, "Farming and farm income," *USDA, Economic Research Service*, 2023.

[20] M. S. Crossley, K. D. Burke, S. D. Schoville, and V. C. Radeloff, "Recent collapse of crop belts and declining diversity of us agriculture since 1840," *Global Change Biology*, 2020.

[21] D. Bloome, J. Feigenbaum, and C. Muller, "Tenancy, marriage, and the boll weevil infestation, 1892-1930," *Demography*, 2017.

[22] "Usda aphis imported fire ant federal quarantine data," *Data*, 2023.

[23] U. J. Sánchez-Reyes, R. W. Jones, T. J. Raszick, R. Ruiz-Arce, and G. A. Sword, "Potential distribution of wild host plants of the boll weevil (anthonomus grandis) in the united states and mexico," *Insects*, 2022.

[24] B. Drees and D. Oi, "Natural enemies of fire ants," *Extension Foundation*, 2017.

[25] F. Eckert, A. Gvirtz, J. Liang, and M. Peters, "A method to construct geographical crosswalks with an application to us counties since 1790," *NBER Working Paper*, 2020.

[26] M. Labouyrie, C. Ballabio, F. Romero, P. Panagos, A. Jones, M. W. Schmid, V. Kiryukov, O. Dulya, L. Tedersoo, H. Bahram, E. Lugato, M. G. A. van der Heijden, and A. Orgiazzi, "Patterns in soil microbial diversity across europe," *Nature Communications*, 2023.

[27] E. Frank, "The economic impacts of ecosystem disruptions: Private and social costs from substituting biological pest control," *Working Paper*, 2022.

[28] A. E. Larsen and F. Noack, "Identifying the landscape drivers of agricultural insecticide use leveraging evidence from 100,000 fields," *PNAS*, 2017.

[29] A. E. Larsen and F. Noack, "Impact of local and landscape complexity on the stability of field-level pest control," *Nature Sustainability*, 2021.

[30] D. S. K. et al, "Crop pests and predators exhibit inconsistentresponses to surrounding landscape composition," *Proceedings of the National Academy of Science*, 2015.

[31] M. Hartfield, K. A. J. White, and K. Kurtenbach, "The role of deer in facilitating the spatial spread of the pathogen borrelia burgdorferi," *Theoretical Ecology*, 2011.

[32] D. J. Civitello, J. Cohen, H. Fatima, and J. R. Rohr, "Biodiversity inhibits parasites: Broad evidence for the dilution effect," *PNAS*, 2015.

[33] J. R. Rohr, D. J. Civitello, F. W. Halliday, P. J. Judson, K. D. Lafferty, C. L. Woods, and E. A. Mordeca, "Towards common ground in the biodiversity– disease debate," *Nature Ecology and Evolution*, 2020.

[34] E. Frank and A. Sudarshan, "The social costs of keystone species collapse: Evidence from the decline of vultures in india," *Working Paper*, 2023.

[35] D. L. Feir, R. Gillezeau, and M. E. Jones, "The slaughter of the bison and reversal of fortunes on the great plains," *Review of Economic Studies*, 2023.

[36] Y. Liang, I. Rudik, and E. Zou, "The environmental effects of economic production: Evidence from ecological observatinos," *NBER Working Paper*, 2023.

[37] C. Santana, "Biodiversity in a chimera, and chimeras aren't real," *Biology and Philosophy*, 2018.

[38] E. Strobl, "Preserving local biodiversity through crop diversification," *American Journal of Agricultural Economics*, 2021.

[39] E. Palmu, J. Ekroos, H. I. Hanson, H. G. Smith, and K. Hedlund, "Landscape-scale crop diversity interacts with local management to determine ground beetle diversity," *Basic and Applied Ecology*, 2014.

[40] M. A. Clemens, E. G. Lewis, and H. M. Postel, "Immigration restrictions as active labor market policy: Evidence from the mexican bracero exclusion," *American Economic Review*, 2018.

[41] J. Lafortune, J. Tessada, and C. González-Velosa, "More hands, more power? estimating the impact of immigration on output and technology choices using early 20th century us agriculture," *Journal of International Economics*, 2015.

[42] M. Haines, P. Fishback, and P. Rhode, "United states agricultural data, 1840-2012 (icpsr 35206)," 2019.

[43] G. Fischer, F. O. Nachtergaele, H. van Velthuizen, F. Chiozza, G. Francheschini, M. Henry, D. Muchoney, and S. Tramberend, "Global agro-ecological zones (gaez v4)-model documentation," 2021.

[44] W. M. Persons, "The crisis of 1920 in the united states: A quantitative survey," *American Economic Review*, 1922.

[45] N. Baum-Snow, "Did highways cause suburbanization?," *Quarterly Journal of Economics*, 2007.

[46] R. Hornbeck and S. Naidu, "When the levee breaks: Black migration and economic development in the american south," *American Economic Review*, 2014.

[47] usda national agricultural statistics service, "Cropland data layer," *Data*.

[48] prism climate group University of Oregon, "Prism historical past," *Data*.

[49] g. s. U S department of interior and us department of agriculture, "Landfire: Landfire existing vegetation type layer," vol. Data, 2013.

A Data and Methods

A.1 Measuring crop diversity

Diversity is a rich concept with many different meanings. In our case, we use local crop diversity as a proxy for the fragmentation of cropland, the competitiveness of local ecological niches, and the presence of natural predators or competitors to our studied pests. First, we construct a **crop richness** index by computing the number of crop species present in each county in each decade.[15] Second, we construct two different measures of evenness, in order to describe the homogeneity of land allocation across these crops. The **Gini-Simpson** index corresponds to the probability of inter-species encounters.[16] In our context, we compute the probability that two hectares of land sampled at random from a given county are covered by the same crop. The **Pielou** index is a concentration measure and captures the commonness or rarity of a species in a county year.[17] All three of these diversity measures are associated to landscape segmentation in different ways. Using a variety of measures allows to compare their relative impact on pest diffusion and to draw a more complete picture of the effect of crop diversity on pest diffusion.

Computing these indices is challenging in our context. Ideally, we would derive the indices using the universe of crops grown in the county of interest. Because we build our measures using historical Census data, we are limited by the number of crops that were recorded consistently over time. We use two different subsets of crops for which we can construct a balanced time series of planted acres. We focus as much as possible on crop categories that are not redefined throughout the census. The first sample is based on the twelve crops with acreage reported consistently from 1880 onwards. Crops included in this sample are rice, cane sugar, wheat, corn, cotton, potato, sweet potato, barley, rye, tobacco, buckwheat, and oats.

We also measure crop diversity using a larger thirty-eight-crop subset, composed of the same crops as previously listed, to which we add several specialty crops.[18] This index is built from 1930 onwards due to a lack of records for some of these crops before the turn of the twentieth century.

Table A.1 presents the correlation between the diversity index obtained from the two subsets of crops described above. As expected the richness index roughly doubles when

[15] Crop richness$_{ct} = \sum_{k \in \mathcal{K}} 1$ where $\mathcal{K}$ is the set of crop grown in county c at time t

[16] Gini-Simpson$_{ct} = 1 - \sum_{k \in \mathcal{K}} p_{kct}^2$ where $\mathcal{K}$ is the set of crop grown in county c at time t and p_k is the number of acres allocated to crop k.

[17] Pielou$_{ct} = -\frac{\sum_{k \in \mathcal{K}} p_{kct} \ln p_{kct}}{\sum_{k \in \mathcal{K}} 1}$ where $\mathcal{K}$ is the set of crop grown in county c at time t and p_k is the number of acres allocated to crop k.

[18] Crops included in this index are soybean, rice, wheat, corn, cotton, sorghum, peas, beans, potatoes, peanuts, sweet potatoes, barley, rye, flax, tobacco, buckwheat, oats, asparagus, beets, broccoli, carrots, cauliflower, celery, collards, cucumber, eggplants, escarole, kale, lettuce romaine, okra, pumpkin, radish, squash, turnip, millet, cantaloup, raspberries, strawberries, watermelon.

Table A.1: Correlations Within Diversity Indices Across Crop Sets

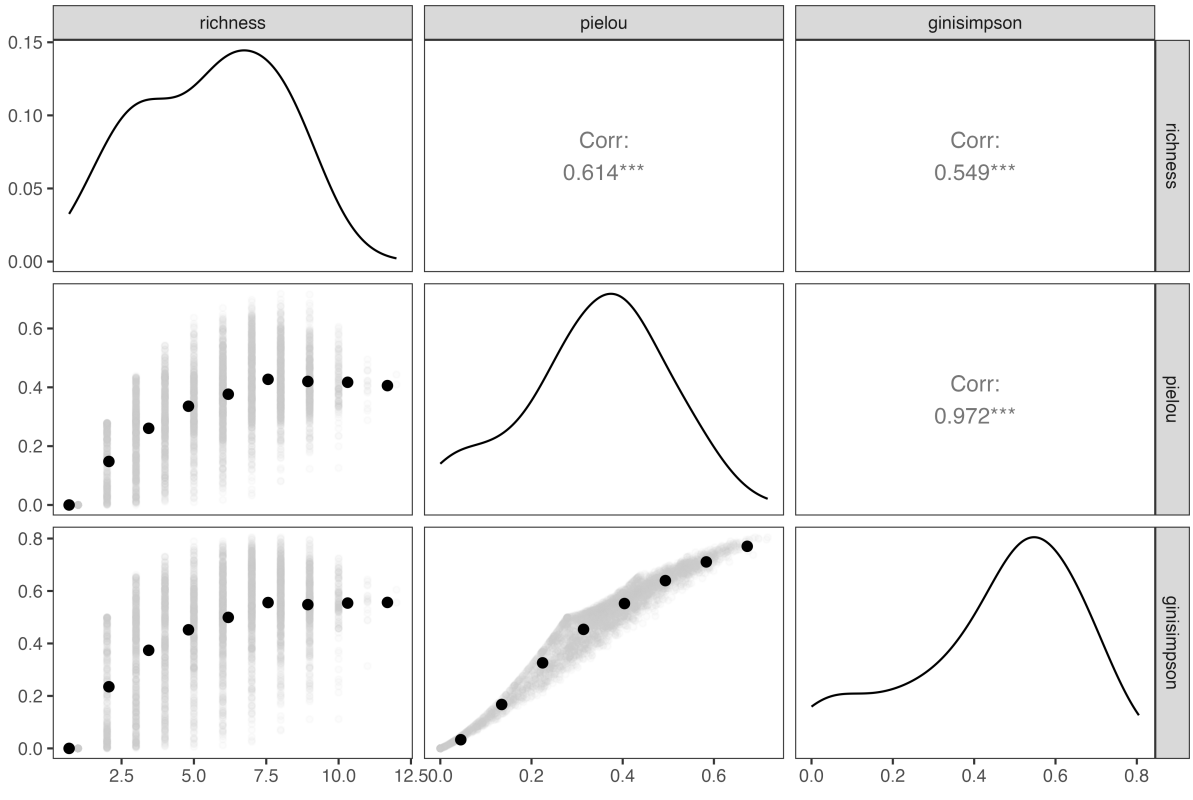
Measures:	Richness	Pielou	Gini-Simpson
Correlation 12-crop	1.989***	0.8714***	1.169***
vs 38-crop samples	(0.0024)	(0.0011)	(0.0016)

Notes: Correlations between county-level diversity indices obtained using the twelve-crop sample versus the thirty-eight-crop sample. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

going from the 12-crop sample to the 38-crop sample. By contrast, the Pielou and Gini-Simpson indices exhibit a correlation close to one between the two samples: 0.87 and 1.17 respectively. Because these are highly correlated and given that the 12-crop sample provides us with the longest time series and the most county-decade observations, we run our analysis with the smaller set of crops.

Figure A.1 shows the distribution and correlations between our three diversity measures. All measures are positively correlated: a county with a high crop richness is more likely to also exhibit a high evenness: the correlation between richness and the Pielou index is 0.67. We also find a correlation of 0.97 between the Pielou and the Gini-Simpson indices. We thus decide to run our analysis using only richness and the Gini-Simpson index.

Figure A.1: Correlation across Diversity Indices



Notes: The diversity indices are computed using the twelve-crop sample for the period 1890-1997.

A.2 Linking the US Census of agriculture over time

US county boundaries underwent many changes over the 1840-2021 period, making the linking of Census data over time challenging. In order to account for these changes, one needs to use a stable map of US counties and reallocate the data as collected in each Census wave to that stable map.

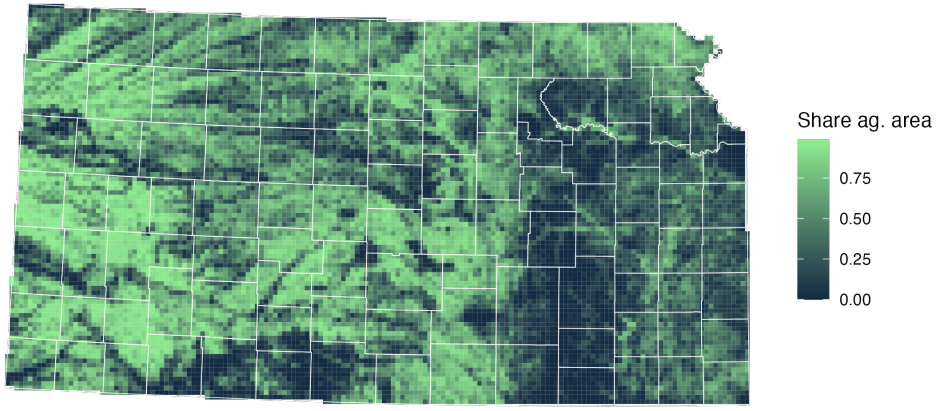
Two traditional methods exist to perform this reallocation exercise. The first method entails identifying the smallest time-invariant geographic units - which could be much larger than a county. This method is assumption-free, and therefore quite useful, but leads to significant aggregation if one wishes to look at census data over more than a few decades.[19]

The second method distributes the data to a selected target map assuming homogenous distribution of the measured variables over the old geographic units. In what follows, we call this method the *area-based method*. This algorithm has been the preferred one for ecological and economic work based on the US Census or the US Agricultural Census [46, 20]. Recently, Eckert et al (2021) [25] published readily available datasets of aggregation weights for applying this method to all US 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.2 shows the 2021 distribution of agricultural coverage in Kansas. White outlines correspond to the boundaries of the 1997 US counties. Agricultural coverage is far from being homogeneous within each shape. This makes the *area-based linking method* hypothesis 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, as well as to create heteroskedasticity in the errors.

We propose a *machine-learning-based method* to ease 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 [47], weather data from PRISM [48] and topographic data from the Landfire dataset [49]. 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 to be smaller than a county, which is the case for the historical PRISM series which 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.3 provides a stylized example comparing the outcomes of the *area-based* and the *machine-learning-based* linking methods.

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

Figure A.2: 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 US counties in 2021. Agricultural land is not homogeneously distributed within counties.

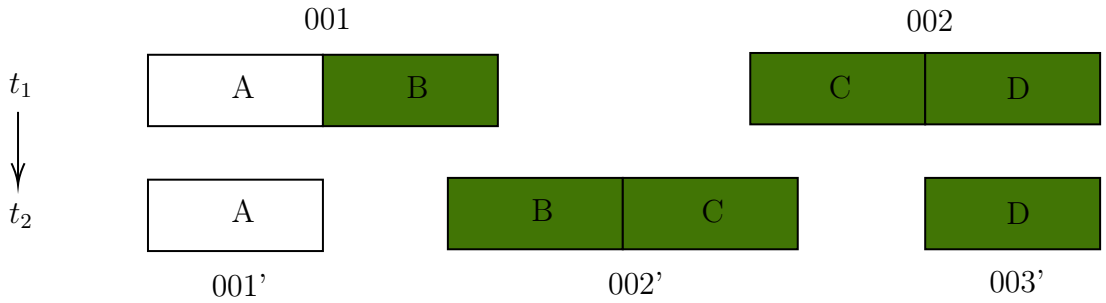
A.2.1 Machine learning based agricultural Census linking method

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 4X4km grid for the contiguous US, and elevation and slope data from the Landfire dataset [49]. An XGboost model is used to predict the presence of agriculture over each pixel. Out-of-sample, we reach an accuracy of 85.5% for 2021. In table A.2, we provide a discretized confusion matrix, testing 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 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 is not as accurate to precisely predict the acreage in medium-range cells, with coverage ranging between 25% and 75%.

Second, we use this model to predict cropland presence for each year of the agricultural Census between 1880 and 1997. Because PRISM only goes back to 1895, we use the

Figure A.3: 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.2: 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 US would lead to an RMSE of 0.87.

1895-1924 average weather for census years prior to 1900.

Finally, we choose the 1997 US county map as our target map and intersect it with every historical county map. For stock variables[20], 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.

A.2.2 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 US mostly fluctuates in the early years of the Census (end of the nineteenth and beginning of the twentieth century). As such, the biases introduced by the intersected area method will be higher for data gathered in the earlier years of the census.

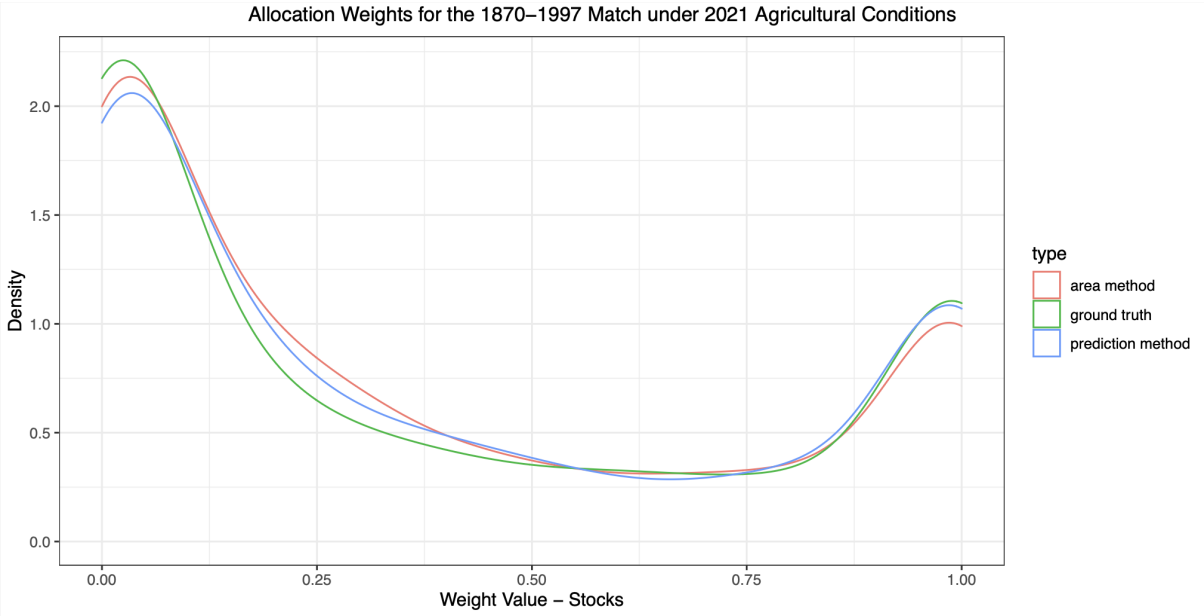
To test our algorithm against the *area-based linking method*, we choose to build an ex-

[20] 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.

ample using the 1870 map of US 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.4 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.

Figure A.4: Comparing the distribution of agricultural fraction across linking methods



Notes: We compute true 1870/2021 county-level agricultural fractions, with the ones predicted off of 1870 counties using either the area-based method or our machine-learning-based one. We note that our distribution is almost always closer to the true distribution.

Second, we check the performance of the *machine-learning-based* algorithm in correctly 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 base crosswalk. Results are presented in Table A.3. Our method outperforms the area-based one for corn and wheat, and slightly underperforms 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 then seems that soy is grown more homogeneously in counties that cultivate soy, relative to wheat and corn.

Table A.3: Crop-Specific Matching Improvement

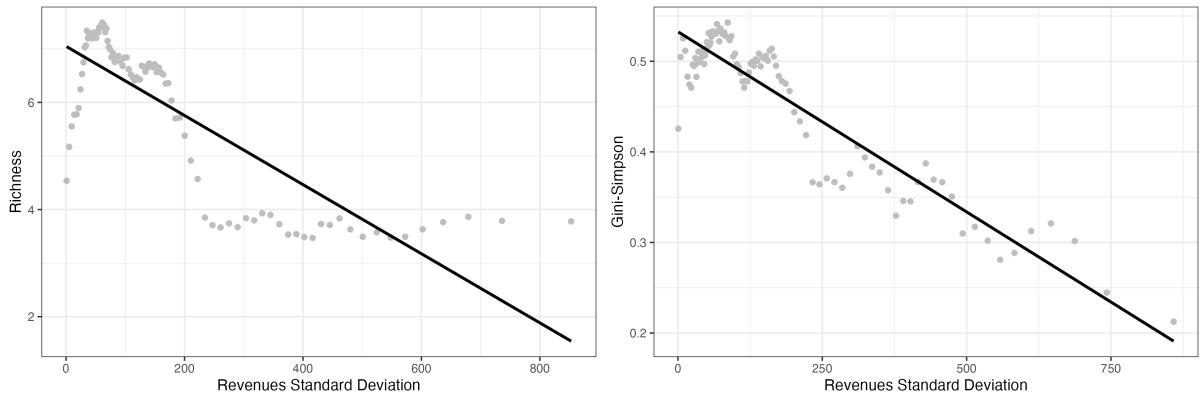
	Area-based	ML-based	RMSE gain
Corn	0.09	0.01	**88%**
Wheat	0.10	0.04	**60%**
Soy	0.03	0.04	-33%

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

A.3 Instrument variable approach

Figure A.5 shows the correlation between the standard deviation in expected crop revenues and county-level measures of crop diversity. As such, these are indicative of the variation used in the first stage of the two-stage instrumental variable approach.

Figure A.5: Standard deviation of expected crop revenues and crop diversity



Notes: Crop diversity is measured at the county level over the period 1890-1997.

B Robustness

Railway expansion and the diffusion of the boll weevil We run the same instrument regression design as in 2, controlling for the evolving shape of the railway network across 1890-1930, a period that also corresponds to its development. Each country's relation to the railway network is measured by the distance between its centroid and the nearest railway segment. The distance to the railway network has no statistically significant relation to a county's contamination status, in both the OLS and the instrument variable regressions, for the two measures of diversity. As discussed previously, the literature expects that the boll weevil did not travel alongside agricultural commodities. We then expect the rail to be a bad control in our context, taking away some of the exogenous variation in county-level crop diversity. We note, however, that in the case of the richness index, the IV regression becomes significant with a positive sign when we include counties' varying distances to the rail network.

Highway expansion and the diffusion of the imported fire ant We run the same instrumental regression design as in Table 2, additionally controlling for the evolution of the interstate network - the major change to the United States transportation network over the second half of the twentieth century. Each county's relation to the interstate is measured by the distance between its centroid and the nearest interstate segment, and is computed using data from Baum-Snow (2007) [45]. The effect of distance to the interstate highway on imported fire ant diffusion is large, negative, and statistically significant. This indicates that the closer counties are to the interstate, the more likely they are to be contaminated by the imported fire ants. As such, a decrease of 100km in a county's distance to the interstate is associated with an average increased probability of contamination of 3.56%. This coefficient remains stable when we use either richness or the Gini-Simpson index as a measure of local crop diversity. While this control is not meant to have a causal interpretation, this negative significant sign likely relates to the role of transportation networks in favoring the spread of the imported fire ant. Results remain largely unchanged. The coefficient of interest increases slightly in absolute value, going from -.160 to -.202 for richness, and from -.826 to -1.04 for the Gini-Simpson index.

Table B.4: Crop Diversity and Pest Contamination: Transportation Networks Controls

Panel A	Boll Weevil (0/1)			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Richness	-0.0417***	0.9251*		
	(0.0068)	(0.5421)		
Gini-Simpson			-0.4498***	-5.679**
			(0.0702)	(2.744)
Observations	4,043	4,043	4,043	4,043
Counties	1,142	1,142	1,142	1,142
R^2	1.0000	0.99999	1.0000	1.0000
F-test (1st stage)		3.9095		8.3411
County fixed effects	✓	✓	✓	✓
Decade fixed effects	✓	✓	✓	✓
Distance to Rail	✓	✓	✓	✓
Panel B	Imported Fire Ants (0/1)			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Richness	0.0060***	-0.2021***		
	(0.0015)	(0.0467)		
Gini-Simpson			0.1541***	-1.042***
			(0.0126)	(0.2392)
Observations	25,052	25,052	25,052	25,052
Counties	1,988	1,988	1,988	1,988
R^2	0.62684	0.32485	0.62943	0.45832
F-test (1st stage)		39.462		86.711
County fixed effects	✓	✓	✓	✓
Decade fixed effects	✓	✓	✓	✓
Distance to Highway	✓	✓	✓	✓

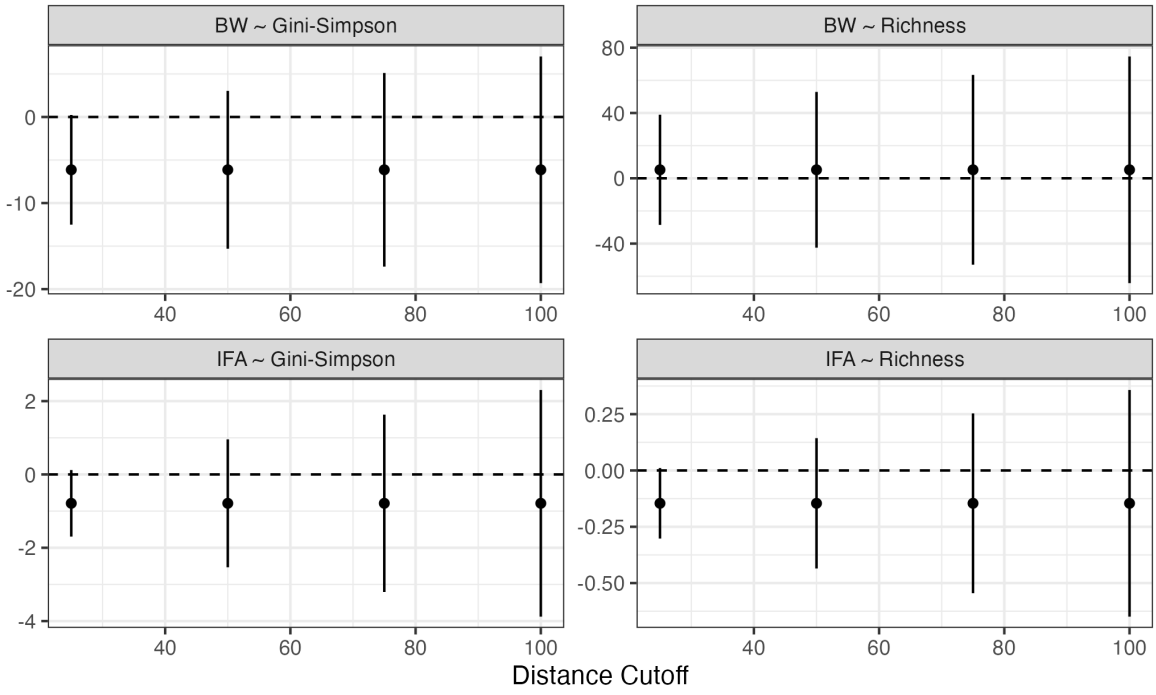
Notes: This table contains the second stage result of our instrument variable regressions, for respectively 1890-1930 for the Boll-Weevil, and 1940-1997 for the imported fire ant. As in the main result section (Table 2), we instrument county level crop diversity with standard deviation in expected crop revenue. The regressions control for county-level total agricultural area as well as county and decade fixed effects. Additionally, we control for county-level distance to resp. the railway (1890-1930), and the highway (1940-1997). Standard errors are heteroskedasticity-robust. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table B.5: Crop Diversity and Pest Contamination: Dropping California & Arizona

	Imported Fire Ants (0/1)			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Richness	0.0074***	-0.1459*		
	(0.0026)	(0.0769)		
Gini-Simpson			0.1659***	-0.7876*
			(0.0224)	(0.4490)
Observations	25,006	25,006	25,006	25,006
Counties	1,988	1,988	1,988	1,988
R^2	0.59854	0.43495	0.60145	0.49292
F-test (1st stage)		45.991		92.233
County fixed effects	✓	✓	✓	✓
Decade fixed effects	✓	✓	✓	✓

Notes: This table contains the second stage result of our instrument variable regressions, for the Imported Fire Ant. As in the main result section (Table 2), we instrument county level crop diversity with standard deviation in expected crop revenue. The regressions control for county-level total agricultural area as well as county and decade fixed effects. Additionally, we remove counties located in California and Arizona, as they were likely contaminated through commodity trade. Standard errors are heteroskedasticity-robust. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Figure B.6: Crop Diversity and Pest Contamination: Alternative Standard Errors



Notes: This figure presents the robustness

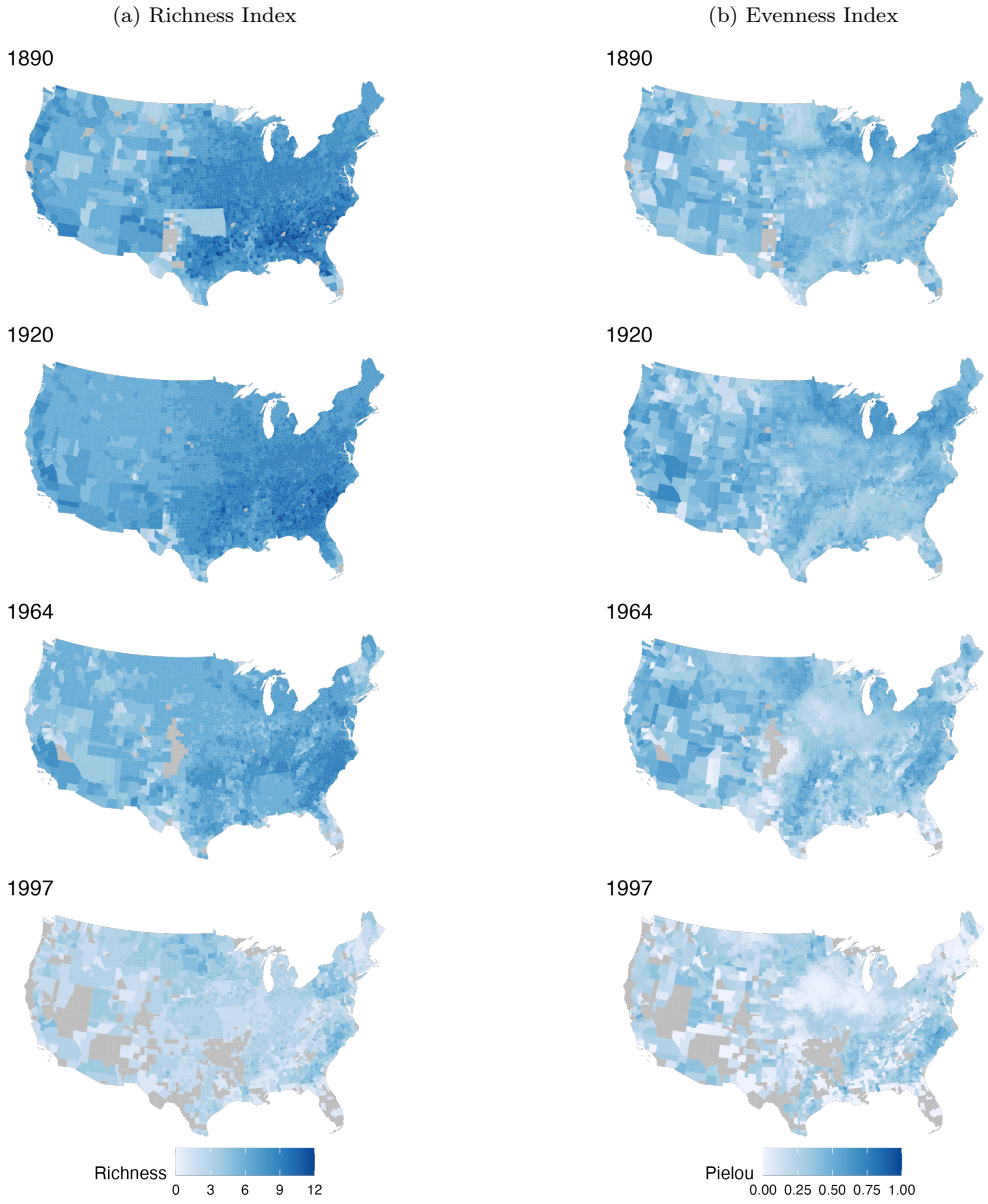
C Additional Results - Crop diversity trends

Figure C.9 shows the trends in crop evenness over time. In Panel (a) evenness is computed on the sample of 12 crops used throughout this analysis. In panel (b), we compute the Gini-Simpson index on the county-specific subset of crops (out of the twelve previously used) that are constantly cultivated in the considered county over the period 1890-2020. As such, each county has its crop mix for which evenness is computed. Comparing the two panels allows us to get a sense of the role of the extensive margin (crops coming in and out of a county's crop mix over time) versus the intensive margin (holding the crop mix constant, variation in their respective land allocations) in explaining the trends in evenness. The trend is much flatter in panel (b). This indicates that the drop in observed evenness is in large part driven by counties stopping production for some crops they would previously have grown.

In Figure C.10, we test the robustness of our measure to the homogenization of agriculture driven specifically by corn. Over the twentieth century, corn and soy played a large role in the development of intensive large-scale agricultural systems in the US. Soy is not accounted for in our 12-crop set, so the evolution of its cropping pattern does not impact our index. Corn is, however, and comparing the two graphs, we can account for its role in driving crop evenness down over time. The two graphs do not show large differences, and as such it seems that corn in itself, and the joint corn-soy intensive agricultural system, is not the only driver of cropland homogenization.

In Figure C.11, we show the evolution of the count of US counties growing each crop in each year of the census. We see that the crops that are most dropped from the counties' crop mixes are the following ones: oats, rye, potatoes and sweet potatoes, cotton, tobacco, and buckwheat. These are thus the ones driving the extensive margin changes discussed previously.

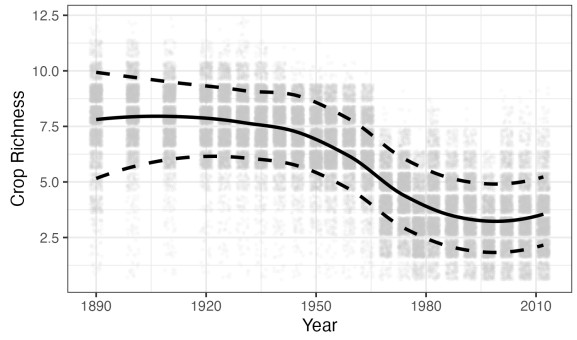
Figure C.7: Changes in Local Crop Diversity over time



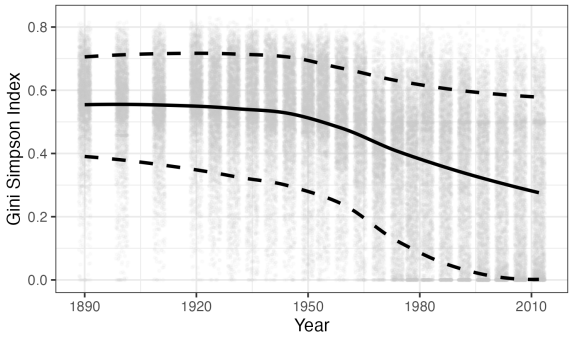
Notes: The diversity indices are computed using the twelve-crop sample.

Figure C.8: Trends in Local Crop Diversity

(a) Richness Index



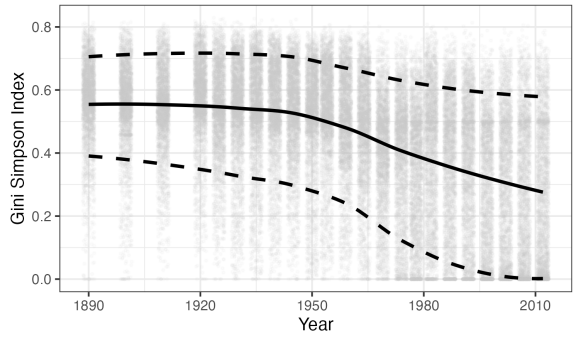
(b) Gini Simpson Index



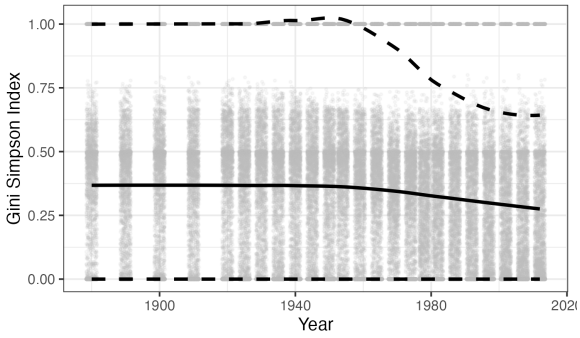
Notes: The diversity indices are computed using the twelve-crop sample.

Figure C.9: Trends in Evenness: Extensive and Intensive Margins

(a) Extensive and Intensive Margins



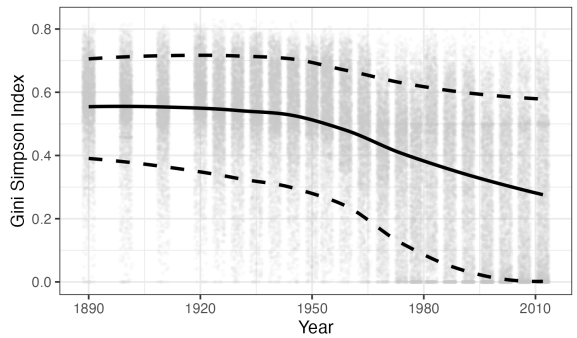
(b) Intensive Margin



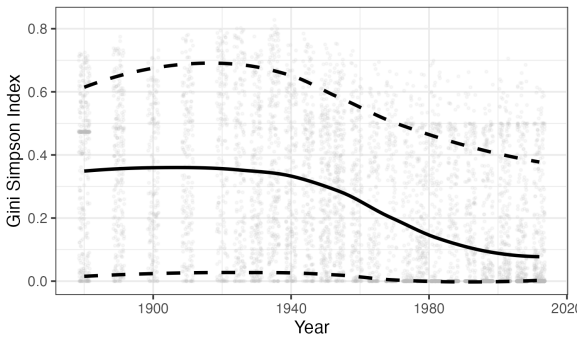
Notes: Panel (a) presents the trend of the Gini Simpson index computed using the sample of 12 crops used throughout the paper. The index in Panel (b) is computed over the maximum county-level stable crop set, or the set of crops for which there exists at least one county with positive acreage in each decade over the period.

Figure C.10: Trends in Evenness: Role of Corn

(a) Initial Crop Set



(b) Excluding Corn



Notes: Panel (a) presents the trend of the Gini Simpson index computed using the sample of 12 crops used throughout the paper. The index in Panel (b) is computed over the small set of crops used throughout the paper, excluding corn. Panel (b) highlights the fact that the loss of diversity over the period 1890-2010 in the US is not only driven by the development of large farms specialized in corn.

Figure C.11: Number of Counties growing each Crop