NBER WORKING PAPER SERIES

USING DIGITIZED NEWSPAPERS TO REFINE HISTORICAL MEASURES: THE CASE OF THE BOLL WEEVIL

Andreas Ferrara
Joung Yeob Ha
Randall Walsh

Working Paper 29808
http://www.nber.org/papers/w29808

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2022

Using Digitized Newspapers to Refine Historical Measures: The Case of the Boll Weevil
Andreas Ferrara, Joung Yeob Ha, and Randall Walsh
NBER Working Paper No. 29808
February 2022
JEL No. N01

ABSTRACT

This paper shows how to remove attenuation bias in regression analyses due to measurement error in historical data for a given variable of interest by using a secondary measure which can be easily generated from digitized newspapers. We provide three methods for using this secondary variable to deal with non-classical measurement error in a binary treatment: set identification, bias reduction via sample restriction, and a parametric bias correction. We demonstrate the usefulness of our methods by replicating two recent studies on the effect of the boll weevil. Relative to the initial analysis, our results yield markedly larger coefficient estimates.

Andreas Ferrara
Department of Economics
University of Pittsburgh
Pittsburgh, PA 15260
a.ferrara@pitt.edu

Joung Yeob Ha
Department of Economics
University of Pittsburgh
Pittsburgh, PA 15260
joh106@pitt.edu

Randall Walsh
Department of Economics
University of Pittsburgh
4901 WW Posvar Hall
230 S. Bouquet St.
Pittsburgh, PA 15260
and NBER
walshr@pitt.edu

1 Introduction

The use of digitized newspaper data by economic historians has become relatively commonplace in recent years. Here we propose the use of such data to overcome measurement error, a problem that is pervasive in the statistical analysis of historical data. Given that regression coefficients of mismeasured variables are attenuated (Aigner, 1973), left unaddressed, measurement error can lead promising research to be abandoned. A solution to such attenuation bias for continuous variables with classical measurement error is to use an instrumental variables approach leveraging a second, also mis-measured, data source as the instrument. In the absence of other endogeneity concerns,[1] as long as the measurement error in the two variables is uncorrelated, instrumenting for one mis-measured variable, X_1, with data from a second mis-measured source, X_2, recovers the true parameter (see Chalfin and McCrary, 2018). The main limitation of this approach is that it is difficult to find a second variable that is i) measured with error which is arguably uncorrelated with the error in X_1, and ii) reasonably cheap to collect. Since economic historians often spend a significant amount of time and effort on original data collection, it is usually costly enough to just have X_1.

In this paper, we show how a second measure, X_2, can often be generated at low cost from textual data available via digitized newspapers,[2] and how it can be used to resolve measurement error in the case where X_1 is continuous or binary. The distinction between continuous and binary variables is important because using X_2 as instrument for X_1 to recover the true parameter only applies to cases of classical measurement error, which requires X_1 to be continuous (Bingley and Martinello, 2017).[3] If X_1 is binary *and* mismeasured, any IV estimate will be inflated by the inverse of the misclassification rate in X_1. This is true even when the instrument is generated by an otherwise perfectly valid natural experiment.

[1] Many current papers in economic history seek to establish causal relationships for which instrumental variables from natural experiments are commonly used and because other endogeneity concerns, such as omitted variables, are a potential problem (see Dippel and Leonard, 2021). These instruments also resolve classical measurement error, however, when treatment variables are not continuous, measurement error is non-classical by construction and the approach fails, an issue that we discuss more fully below.

[2] Examples of available online repositories for digitized historical newspaper data are Chronicling America and Newspapers.com.

[3] Classical measurement error requires that there is no correlation between the true value and the error. Suppose a binary treatment is misclassified, then the error has a perfect negative correlation with the true value by construction because if $X^* = 1$, then $u = -1$ and, vice versa, $u = 1$ if $X^* = 0$.

We provide three potential solutions when X_1 is binary. First, the treatment effect can be *set identified*. The OLS estimate using X_1 as treatment provides a lower bound while the IV estimate using X_2 as instrument for X_1 provides the upper bound such that $\widehat{\beta}_{\text{OLS}} < \beta < \widehat{\beta}_{\text{IV}}$. Second, we show that restricting the analysis to an *agreement sample* where $X_1 = X_2$ can substantially reduce the OLS bias. The probability that both variables are jointly misclassified is the product of the two variables' misclassification rates and therefore measurement error in the agreement sample tends to be much lower.[4] Third, we provide a *parametric bias correction* procedure that can recover the true parameter of interest as a nonlinear combination of the OLS and IV coefficients. All three procedures are fast and efficient, and given that newspaper data can be scraped in a reasonable amount of time, we hope to provide researchers who work with historical data with low-cost tools for dealing with measurement error. We demonstrate our three procedures by replicating two recent papers that study the economic impact of the spread of the boll weevil across the U.S. South in the late 19th and early 20th century, one by Clay, Schmick and Troesken (2019) and one by Ager, Brueckner and Herz (2017).

To date, the sole source of data used by analysts to measure the timing of boll weevil arrival at the county-level stems from a U.S. Department of Agriculture (USDA) map by Hunter and Coad (1923) which documents the arrival date of the pest across Southern counties. While the map itself is mostly accurate, it does contain errors.[5] Further, it does not necessarily measure what economists are typically interested in, namely the timing of the economic damage caused by the arrival of the boll weevil. As an example, if the weevil arrived late in the summer, it would typically hibernate soon after arrival and thus the actual economic damage would not occur until the following year. The arrival date from the USDA map therefore is a mis-measured proxy for the date of the actual economic impact. And, as we document below, this mis-measurement can markedly attenuate estimated effect sizes.

To produce a second measure for the arrival of the boll weevil, we collect data from Newspapers.com by jointly searching the database for pages containing "boll weevil" and each county's name in all newspapers in the county's state for each year between 1882 and 1932. Our arrival

[4] For instance, suppose X_1 and X_2 have misclassification rates of 30% and 20%, respectively, where one minus the misclassification rate determines the OLS bias. The attenuation bias in the agreement sample will be $0.3 \times 0.2 = 0.06$.

[5] In some instances, the map reports inconsistent arrival dates. The map shows the arrival date with date borders which occasionally overlap in contradictory ways. See Figure 1 for examples of such overlaps.

measure is then the peak salience of the weevil in the news as measured by the maximum five years moving average of boll weevil related pages.[6] We argue that errors in this newspaper-based measure are likely to be uncorrelated with errors in the USDA map, which was generated by trained USDA entomologists who reported back to the federal agency, whereas local newspaper reporters mainly wrote about salient issues in their home counties. Using an event study design, we also show that the newspaper-based salience peaks a year after the official USDA arrival date on average.

Our replications of Clay et al. (2019) and Ager et al. (2017) show that using our newspaper-based arrival measure can reduce measurement error and strengthen the results in both papers. In particular, our theory suggests a ranked pattern between the three proposed solutions, where $\widehat{\beta}_{\text{OLS}} < \widehat{\beta}_{X_1=X_2} < \beta = \widehat{\beta}_{\text{bias-corrected}} < \widehat{\beta}_{\text{IV}}$. While we do not observe the true coefficient, the estimated coefficients largely follow the prescribed pattern in both replication exercises. Namely, the parametric bias correction provides a larger coefficient than the OLS coefficient from the $X_1 = X_2$ agreement sample, and both of these lie within the lower and upper bounds provided by the OLS and IV estimates, respectively. We find evidence that measurement error led to lower coefficient estimates in both studies, a finding which is robust across alternative specifications of our newspaper-based arrival date. However, the difference with the coefficients produced by our procedures was only statistically significant for Ager et al. (2017). We discuss the frequency of the time dimension as potential reason for this finding, as Clay et al. (2019) use annual data while Ager et al. (2017) use data over five year intervals.

We also provide a broader discussion of when such data generation from newspaper articles is a promising avenue to resolve measurement error and when it is not, as well as of the value of newspapers to generate novel data for research in economic history in general. Even though our newspaper-based measure of the boll weevil arrival was generated in a fast and unrefined way, using this noisier measure still produces smaller but significant effects that are comparable to those in Clay et al. (2019) and Ager et al. (2017). Even in the absence of the USDA map, we therefore could have conducted their studies solely based on the newspaper data.

Our paper highlights the usefulness of digitized newspapers to generate additional data to ad-

[6] We use a moving average to additionally smooth out noise in the newspaper data and provide sensitivity checks to show that other transformations, such as using a three or seven years moving average or the raw data, give similar results.

dress measurement error. We extend the secondary measure IV framework in Chalfin and McCrary (2018) to the case where treatment is binary and when instrumenting ordinarily does not resolve measurement error (Bingley and Martinello, 2017). While researchers tend to ignore measurement error when some conventional level of statistical significance is achieved, we hope to draw some attention to the issue when the treatment variable is not continuous given that larger IV estimates compared to OLS are frequently motivated with measurement error in the treatment variable. We also contribute to a recent literature that uses digitized newspapers to generate novel data for research in economic history. This includes measures of media competition and partisan influence (Gentzkow, Shapiro and Sinkinson, 2014; Gentzkow, Petek, Shapiro and Sinkinson, 2015), racial and anti-group sentiment (Ferrara and Fishback, 2020; Ottinger and Winkler, 2021; Bazzi, Ferrara, Fiszbein, Pearson and Testa, 2021), the spread of news relating to racial violence (Albright, Cook, Feigenbaum, Kincaide, Long and Nunn, 2021; Calderon, Fouka and Tabellini, 2021), the 1918 influenza (Beach, Clay and Saavedra, 2020), fertility restrictions (Beach and Hanlon, 2021), advertisements for the movie "Birth of a Nation" (Esposito, Rotesi, Saia and Theonig, 2014), the price and types of available cotton seeds (Rhode, 2021), among others.

The remainder of the paper is organized as follows. Section 2 introduces the historical setting of the boll weevil infestation of the U.S. South between 1892 and 1922, and reviews previous literature on the topic to set the stage for our later application. In this context, we then discuss potential measurement issues in the widely used USDA map and describe the collection of our newspaper based boll weevil arrival measure. Using an event study approach, we show the difference between the USDA arrival date and salience in local newspapers to highlight why the map may not be the ideal source for what economists typically aim to measure when studying the effect of the boll weevil. Section 3 provides the econometric framework for how our newspaper-based arrival measure can be used to resolve measurement error in the USDA map arrival measure by introducing three approaches based on set identification, noise reduction by using an agreement sample where both arrival dates give the same answer, as well as a parametric bias correction. Section 4 replicates the studies by Clay et al. (2019) and Ager et al. (2017) to demonstrate how easily collected newspaper data can be used to address measurement error and to confirm the theoretical results from the previous section. We also discuss when our approach is suitable and when it is not, as well as the utility of digitized newspapers to generate novel data. The final section concludes.

2 Background and Measurement of the Boll Weevil Infestation

2.1 The Spread of the Boll Weevil and Uses of the USDA Map

The boll weevil spread across the U.S. South starting in 1892 near Brownsville, Texas. The beetle, which gained its name because of its diet consisting mainly of cotton bolls and flowers, had infested all Southern cotton growing regions by 1922. Given that cotton at the time was still the main cash crop in Southern agriculture (Wright, 2013), the arrival of the pest had a substantial impact on the areas it infested. Consequently, the USDA traced the arrival of the weevil on a map in an annual report by Hunter and Coad (1923). A portion of this map is shown in Figure 1. During peak infestation in 1921, cotton acreage had declined by 31% (Ager et al., 2017) and the USDA estimated the average economic loss per year to be 200 to 300 million USD between 1916 and 1920 (Hunter and Coad, 1923).[7] Given this large economic shock, a well developed literature has studied the various impacts of the boll weevil on different aspects of the Southern economy.

Lange, Olmstead and Rhode (2009) show the large negative impact of the pest on cotton production, land value, and yields in the South together with anticipatory behavior by farmers. The drop in productivity also altered the structure of Southern agriculture with a reduced number of tenant farmers, farm wages, and female labor force participation (Ager et al., 2017). Ager, Herz and Brueckner (2020) provide evidence that the lower returns to agriculture reduced fertility due to the opportunity cost of children and the decreased value of child labor. Also Black Southerners tended to marry later after the pest arrived for the same reasons (Bloome, Feigenbaum and Muller, 2017). This fertility transition and the decline in the value of child labor in agriculture have also been linked to increased educational attainment (Baker, 2015; Baker, Blanchette and Eriksson, 2020). Another unintended consequence of the reduction in cotton production was increased food production. Clay et al. (2019) show that this significantly contributed to the reduction in pellagra deaths. In a later paper, the authors also find that the boll weevil spread reduced the racial income gap in the South (Clay, Schmick and Troesken, 2020). Similar to the population movements discussed in Lange et al. (2009), Feigenbaum, Mazumder and Smith (2020) show that the decline in cotton reliance also resulted in less violence against Black Southerners who saw an increased ability to move and vote with their feet against overtly discriminatory behavior.

[7] The damage corresponds to \$3.2-\$4.8 billion in 2021 dollars.

Most of the above papers either assign the arrival date for a county whenever the USDA map first arrival year line crosses that county's area, or the arrival date is selected for the year line which contains most of the county's area. What should be noted is that the solid lines in the map technically show the farthest extent of the boll weevil in any territory. This measure does not necessarily correlate with the exact timing of damage caused by the insect. Mature boll weevils hibernate during the winter and infest the cotton field after the crop season in the subsequent year. Lange et al. (2009) explicitly mention this caveat in their paper: "First contact usually occurred during the August seasonal migration, too late to build up significant populations or do much damage in that year. Maximum damage occurred after the local weevil population became established and multiplied. Thus, the classic USDA maps detailing the spread of the weevil present a somewhat misleading picture of the area ravaged by the insect" (p. 689).

2.2 Measuring the Boll Weevil's Arrival From Newspaper Data

Newspapers were the primary source of information in the late 19th and early 20th centuries and mainly operated locally in the county where the paper was based (Gentzkow et al., 2014). Newspapers published articles about the boll weevil's arrival as well as damages in cotton production caused by the insects. An example of such reporting is shown in Appendix Figure A.1. Digitized newspaper data are a potential source to generate information on the arrival and damage extent caused by the pest independent of the USDA map. We use Newspapers.com as our primary data source of digitized historical newspapers. To the best of our knowledge, this is the largest newspaper archive available online.[8]

For each county, in order to construct our newspaper-based boll weevil arrival and salience measure, we take all of the available newspapers from said county's state and identify by year the number of newspaper pages that include both the words "boll weevil" and said county's name.[9] We are forced to use all newspapers from an individual county's state because no newspaper archive has information on the universe of newspaper pages. Thus, as described, our search not only

[8] Chronicling America is another digital archive for historical newspapers that is commonly used by researchers (e.g. Wang, 2019; Ferrara and Fishback, 2020). However, it has fewer volumes than Newspapers.com and does not contain many digitized newspapers that cover our sample period. As of 11/21/2021, 691,037,256 newspaper pages are available in Newspapers.com while 18,773,412 pages are available in Chronicling America.

[9] In principle, one could search each article for a specific arrival date mentioned in the page for each county. However, this would be time consuming and therefore costly. We instead use this simple search procedure to minimize the cost for researchers and later show that this quickly obtained raw measure of boll weevil activity is still a good proxy for the insect's arrival and salience in a county.

considers pages in the county of interest but in all counties that are in the same state.[10] So, even if Autauga County in Alabama has no available newspaper pages for the search period but "Autauga County" and "boll weevil" are mentioned in a newspaper based in Barbour County, Alabama, we obtain data for Autauga County. Some counties may feature more prominently in the news than others, which is why we need to adjust these counts for the overall number of pages that mention the county. Thus, we apply the same search logic to generate the numerator in our boll weevil measure, which we compute as

$$\%BW_{ct} = \frac{\text{No. of in-state newspaper pages mentioning ``boll weevil'' and a county's name}_{ct}}{\text{No. of in-state newspaper pages mentioning a county's name}_{ct}} \quad (1)$$

where $\%BW_{ct}$ captures the salience of the boll weevil for county c in year t in the news. Our sample includes 911 infested counties from 13 Southern states between 1882 and 1932,[11] which is ten years before and after the time periods covered by the USDA map.

How does our salience measure relate to the official arrival date in the USDA map? To answer this question formally, we use an event study design and estimate the following equation,

$$\%BW_{ct} = \pi_c + \gamma_{st} + \sum_{\ell=-10}^{-2} \beta_\ell \cdot D(t - BW_c^{USDA} = \ell) + \sum_{\ell=0}^{10} \beta_\ell \cdot D(t - BW_c^{USDA} = \ell) + \varepsilon_{ct} \quad (2)$$

where $\%BW_{ct}$ is our newspaper-based salience measure for county c in year t. $D(t - BW_c^{USDA} = \ell)$ is an event indicator relative to the arrival of the boll weevil from the USDA map for the ten years before and after the official arrival date. The year before the arrival from the USDA map, $\ell = -1$, is omitted and serves as the baseline period. The county fixed effects π_c capture time invariant unobservable county characteristics and aggregate time trends that affect counties jointly in each state are captured by state-by-year fixed effects γ_{st}. Standard errors are clustered at the county level. Given the recent literature on issues related to event study designs, we use the estimator developed by Sun and Abraham (2020).

Our main interest is in the lag coefficients β_ℓ for $\ell \geq 0$. If salience in the news correlates highly with the USDA arrival date, then we should observe an immediate jump at the treatment

[10] For an example see Appendix Figure A.2.

[11] The sample includes Arkansas, Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, Missouri, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, and Virginia.

date $\ell = 0$, followed by an either constant or slowly decaying coefficient pattern. Conversely, if the weevil tends to arrive later in the summer and hibernates, the more salient economic damage would occur in the following year which implies that the main effect on salience in the news should occur after $\ell = 0$. The pattern of the coefficients should not only be informative about the decay in salience after arrival but also reveals potential anticipatory behavior if the lead coefficients are significant for $\ell < -1$.

Figure 2 plots the dynamic treatment effects for the twenty year event window around each county's boll weevil arrival date in the USDA map on our newspaper-based salience measure. The figure shows the coefficients from estimating equation (2) via two-way fixed effects (TWFE) and with the estimator developed by Sun and Abraham (2020). We find that the salience measure significantly increases in counties after the boll weevil's arrival based on the USDA map. More importantly, the effect is largest one year after the arrival date in the USDA map. This confirms the narrative that salience in the news and arrival are somewhat but not perfectly correlated due to the pests' hibernation if it arrives later in the summer (see Harned, 1910). While the post-arrival coefficients slowly decay, they are still statistically significant even ten years after the arrival of the weevil. We find no evidence for anticipatory reporting in the four years prior to the USDA map's arrival date. For earlier periods, there are significant coefficients in the TWFE results. We find no pre-trends using the estimator by Sun and Abraham (2020).

2.2.1 Prediction of the Boll Weevil Infestation using Historical Newspapers

Figure 2 illustrates that the USDA map itself is mostly accurate, but it does not necessarily measure what economists are typically interested in, namely the economic impact of the boll weevil. However, the main purpose of the newspaper data was to generate a second variable that predicts the arrival of the boll weevil. One possible such measure would be to simply take the maximum of $\%BW_{ct}$. To generate a more stable prediction that is less prone to outliers or noise in the newspaper data, we first smooth out noise by applying a five years moving average

$$MA(5)_{ct} = \frac{1}{5} \sum_{k=-2}^{2} \%BW_{c,t+k}$$

and then assign the maximum as predicted year of infestation

$$\text{Predicted year of infestation}_c = \max_{t \in [1882,1932]} (MA(5)_{ct}) \tag{3}$$

For robustness, we later test alternative specifications such as the three and seven years moving averages, as well as the maximum salience measure $\%BW_{c,t}$ within a 10-year window around the USDA map. While our preferred specification is MA(5), the results in Section 4 are robust across alternative specifications. More details are discussed below.

To illustrate how our approach based on newspapers can predict its effective infestation, consider the following example for Marion County in Mississippi. The USDA map recorded that the boll weevil arrived Marion in 1909. However, the damage by the insect was not severe. Harned (1910), the head of the department and entomologist for the Mississippi Agricultural Experiment Station, investigated the infestation in Mississippi during 1907 and 1909. For Marion County he found that, "boll weevils probably spread entirely over this county during September, 1909, although not in large enough number to do serious damage" (p. 22). For each year between 1882 and 1932, we first calculate the salience of the boll weevil of Marion County using pages mentioning "boll weevil" and "Marion County". We calculate $\text{MA}(5)_{Marion,t}$ for each year, and define the effective infestation of Marion County by choosing the year with the maximum $\text{MA}(5)_{Marion,t}$. Our newspaper-based approach predicts that the effective infestation was 1910 in Marion County, which is one year after the boll weevil's arrival in 1909 according to the USDA map. This analysis is shown in Figure 3. The dashed line and solid line indicate Marion County's $\%BW_{ct}$ and $\text{MA}(5)_{ct}$ over time, respectively. While our salience measure based on newspapers is noisy (dashed line), the five years moving average smooths out this noise (solid line). Peak salience in the news appears to be a reasonable approximation for the arrival of the pest.

Lastly, we provide a comparison between our predicted arrival date (equation (3)) and that provided by the USDA map. Figure 4 plots the difference in the two arrival dates for the 911 counties in our sample. A positive difference means that the predicted year based on newspapers is later than the arrival of the boll weevil as presented in the USDA map. While the difference is typically small, less than 4 years for more than half of sample counties (54.88%), we find that the difference is extreme for a small number of counties. This result is likely due to the noise in the newspaper data such as cases where the search words appear in separate articles even though they show in the same newspaper page.[12] It should be kept in mind that our measure is in some ways purposefully noisy simply to reduce the cost of collecting the data. More refined versions are possible by applying a visual inspection of the newspaper data, which would increase the cost of data collection.

Another reason for some of the extreme values in the difference is due to some newly constructed counties. An example is shown in Appendix Figure A.4. Dixie County in Florida was created in 1921 from the southern portion of Lafayette County. While the boll weevil arrived Dixie County in 1916 according to the USDA map, our newspaper-based measure predicts its effective infestation as 1932. This is because our prediction based on newspapers mentioning "Dixie County". Since Dixie County did not exist before 1921, the prediction is only based newspapers after 1921, which shown in Panel (a) of Appendix Figure A.4. One possible solution is to aggregate those counties (as "multi-counties" in Lange et al. (2009) and Ager et al. (2017)) or assign the predicted year from its original county.[13]

3 Resolving Bias from Measurement Error using Secondary Measures

3.1 Classical Measurement Error

How can the second measure for the boll weevil arrival from newspaper data be used to correct for measurement error in the USDA map arrival date? First, consider the case where the data is used as continuous exposure measure, such as years since arrival of the pest, for instance. Suppose a researcher wants to estimate the following linear equation by OLS, which is assumed to be unconfounded with a clear direction of causality but where years since arrival of the boll weevil, X_1, is continuous and measured with error

$$y = \alpha + \beta X_1 + \epsilon \quad \text{and} \quad X_1 = X^* + u$$

[12] Appendix Figure A.3 shows that the search word "boll weevil" appears in one article and "Marion County" appears in another article.

[13] For available crosswalks to standardize county boundaries over time see Ferrara, Testa and Zhou (2021).

with $Cov(X^*, u) = 0$, β is the true parameter, and X^* is the true measure (i.e. measured without error). The estimated coefficient will then suffer from the typical attenuation bias,

$$\widehat{\beta}_{OLS,X_1} = \beta \frac{Var\left(X^*\right)}{Var\left(X^*\right) + Var\left(u\right)}$$

where we denote the estimator and treatment variable of interest in the subscripts of $\widehat{\beta}_{OLS,X_1}$, respectively. Now suppose there is a second variable that seeks to capture X^* as well but that is also mismeasured, $X_2 = X^* + e$, and for which the same conditions apply as for X_1. We can then use X_2 as instrument for X_1 to solve the measurement error problem (see Chalfin and McCrary, 2018). The IV estimate will be

$$\widehat{\beta}_{IV,X_1} = \beta \frac{Var\left(X^*\right)}{Var\left(X^*\right) + Cov\left(u, e\right)} \tag{4}$$

In the absence of any other endogeneity problems and if the two measurement errors are uncorrelated such that $Cov(u, e) = 0$, the IV estimate will recover the true parameter. As with the exclusion restriction, one would then have to make an argument as to why the two errors should be uncorrelated or that this correlation is close to zero. In the case of the boll weevil, a possible argument would be that the USDA map was compiled by trained entomologist who primarily reported back to the agency, whereas the newspapers were written by journalists who reacted to local developments in their county. If journalists were basing their stories, and in particular the timing of their articles, on the USDA map, then this assumption fails in which case $Cov(u, e) > 0$ and the estimated IV coefficient in (4) would be biased downward.

Since applied economists tend to think hard about the exclusion restriction, we would like to highlight that this condition is satisfied in our case by assuming away endogeneity concerns other than measurement error. If X_2 affects y through channels other than X_1, such other channels must necessarily be in ϵ. Since X_2 and X_1 seek to measure the same quantity, this essentially also implies a correlation between X_1 and the error term as well. This is something that our approaches in this paper cannot solve. At best, X_2 can remove biases relating to measurement error but not those stemming from omitted variables or reverse causality, for instance.

3.2 Nonlassical Measurement Error

Oftentimes the arrival or presence of the boll weevil, however, is coded as a binary variable (e.g. Clay et al., 2019, 2020; Ager et al., 2017). In this case, the IV coefficient will no longer be unbiased because when the treatment variable is discrete or binary, then measurement error is no longer classical by construction (Bingley and Martinello, 2017).[14] Suppose that now X_1 is binary. When regressing y on X_1, the estimated OLS coefficient is still attenuated with

$$\widehat{\beta}_{OLS,X_1} = \beta\,(1-\theta)$$

where θ is misclassification rate in X_1 (Aigner, 1973). If $\theta = 0$, then there is no measurement error whereas $\theta = 1$ means that X_1 is entirely randomly misclassified such that it is uncorrelated with X^* and therefore contains no usable information. Now suppose that X_2 is also binary and misclassified, but with an error γ that is uncorrelated with θ, and $\gamma < \theta$. If we then regress y on X_2, the estimated coefficient will also be biased, $\widehat{\beta}_{OLS,X_2} = \beta(1-\gamma)$, however, this attenuation bias will be smaller than for X_1 since $\beta(1-\gamma) > \beta(1-\theta)$ in absolute terms.

If we now try to instrument X_1 with X_2, or vice versa X_2 with X_1, as in the classical measurement error case considered before, this will not recover the true parameter of interest. Instead, the estimated coefficient for those two cases will be

$$\widehat{\beta}_{IV,X_1} = \beta\frac{1}{(1-\theta)} \quad \text{and} \quad \widehat{\beta}_{IV,X_2} = \beta\frac{1}{(1-\gamma)}$$

depending on which variable was used as the treatment and the instrument. This outcome is the inverse of the respective OLS bias terms.[15] Unlike OLS, which suffers from attenuation bias, the IV estimate will be inflated instead with $\beta\frac{1}{(1-\gamma)} < \beta\frac{1}{(1-\theta)}$.[16] Neither OLS nor IV yield an unbiased estimate, however, we now offer three potential approaches for identifying the treatment effect or for at least minimizing the attenuation coming from the misclassification.

[14] A key assumption of classical measurement error is $Cov(X^*, u) = 0$, i.e. the error is uncorrelated with the true variable. Now suppose X^* is binary. If for a given observation $X^* = 1$, then the error can only be $u = -1$. Conversely, if $X^* = 0$ then $u = 1$, meaning that there is a perfect negative correlation between the true variable and the error.

[15] See Bingley and Martinello (2017) as well as Dupraz and Ferrara (2021) for measurement error in linked Census data. For a derivation see the Appendix.

[16] Notice that this requires $\theta \neq 1$ and $\gamma \neq 1$ as the IV estimator is not even defined otherwise.

Solution 1 - set identification: Even though the true parameter of interest cannot be directly point identified, the above OLS and IV coefficients can be used as lower and upper bounds, respectively, to set identify β it given that

$$\widehat{\beta}_{OLS,X_1} < \widehat{\beta}_{OLS,X_2} < \beta < \widehat{\beta}_{IV,X_2} < \widehat{\beta}_{IV,X_1}$$

While it is not known a priori whether X_1 or X_2 has the higher measurement error, the above inequality suggests the set can be inferred from the relative magnitudes of the OLS and IV coefficients. In the above example, set identification implies that $\beta \in \left(\widehat{\beta}_{OLS,X_2}, \widehat{\beta}_{IV,X_2}\right)$. Without additional assumptions, these bounds are tight and are informative as long as zero is not included in the set. To assess the latter condition, the OLS estimate provides the corresponding test which rejects non-informativeness when $\widehat{\beta}_{OLS,X_2}$ is significantly different from zero.

Solution 2 - agreement sample: If instrumenting as described above is too complicated, e.g. if researchers wish to estimate nonlinear treatment effects or their specification includes interactions of the treatment with other variables, the OLS bias can be further reduced by considering only the part of the sample for which X_1 and X_2 both provide the same value. We call this an agreement sample. The probability that both measures are jointly incorrect is $\theta \times \gamma = \delta$. For example, suppose the error rates are $\theta = 0.3$ and $\gamma = 0.2$, then $\delta = 0.06$ which substantially reduces the OLS bias for $\widehat{\beta}_{OLS,X_1=X_2} = \beta(1-\delta)$, which will be closer to the true parameter.

One concern with the agreement sample is that it potentially generates a selected subsample that is not necessarily representative of the underlying population. If such selection is a concern, one available correction is to apply inverse propensity score reweighting (see Bailey, Cole and Massey, 2019). First, regress the indicator for being included in the agreement sample on a wide set of pre-treatment county characteristics using a Probit regression. Second, obtain the predicted probability from the previous Probit regression $\widehat{p}$ and use the actual share of observations in the agreement sample q to generate weights as $\frac{(1-\widehat{p})}{\widehat{p}} \times \frac{q}{(1-q)}$. Lastly, run the regression of interest, weighting observations in the agreement by the weights created in the previous step. The weights ensure that the estimation sample is more representative of observations in the entire sample.

Solution 3 - parametric bias correction: While neither OLS nor IV on their own identify the true parameter, their estimates can be used jointly to recover β. The bias-corrected (BC) estimate is then

$$\widehat{\beta}_{\text{BC}} = \sqrt{\widehat{\beta}_{OLS,X_1} \times \widehat{\beta}_{IV,X_1}} = \sqrt{\beta(1-\theta) \times \frac{1}{(1-\theta)}\beta} = \sqrt{\beta^2} = \beta \tag{5}$$

Estimation of (5) is straightforward as the product of two coefficients from different equations can be readily estimated in standard statistical software with standard errors being estimated via the delta method or bootstrapping. Taken together, our three possible solutions yield the following relationship,

$$\widehat{\beta}_{OLS,X_1} < \widehat{\beta}_{OLS,X_2} < \widehat{\beta}_{OLS,X_1=X_2} < \beta = \widehat{\beta}_{\text{BC}} < \widehat{\beta}_{IV,X_2} < \widehat{\beta}_{IV,X_1} \tag{6}$$

which is the pattern that we look for in the subsequent replication exercises.

4 Replication of Clay et al. (2019) and Ager et al. (2017)

In this section, we replicate two recent papers that study the boll weevil's impacts on pellagra deaths (Clay et al., 2019) and cotton productivity (Ager et al., 2017). Implementing our suggested approaches to measurement error based on historical newspaper data, we demonstrate the potential for such data to markedly reduce attenuation bias. Our results suggest that the impact of the boll weevil was larger than previously documented. Further, our analysis largely confirms the ranked pattern for the different measurement error approaches as suggested by equation (6) in the previous section. Results are robust across the alternative specifications discussed in Section 2.

4.1 Replication of Clay et al. (2019)

Using annual data between 1915 and 1925 for counties in North and South Carolina, Clay et al. (2019) show that pellagra deaths decreased following the boll weevil infestation. They argue that this outcome can be explained by the resulting diversification in food production. After the boll weevil infestation, the prevailing cotton monoculture was switched to more niacin-rich crops such as corn and sweet potato. This led to the fall of pellagra, which is a disease related to insufficient niacin consumption. Clay et al. (2019) estimate the following regression equation,

$$\ln[\text{pellagra}]_{ct} = \alpha + \theta_1 \text{boll weevil}_{ct} + \theta_2 \left(\text{boll weevil}_{ct} \times \text{intensity}_{c,1909}\right) + \theta_c + \theta_t + \varepsilon_{ct} \qquad (7)$$

where $\ln[\text{pellagra}]_{ct}$ is the log number of pellagra deaths, or the log pellagra death rate in other specifications, and boll weevil_{ct} is an indicator for whether or not the boll weevil has arrived in county c as of time t. They provide results with and without the additional interaction of the boll weevil variable and an intensity measure. The latter is an indicator for whether a county was in the top quartile of either i) the pre-treatment pellagra death rates measured as average for 1915-16 or ii) cotton acres per capita in 1909. County and year fixed effects are captured by θ_c and θ_t, and standard errors are clustered at the county level.

Our Table 1 replicates the corresponding Table 3 in Clay et al. (2019) using the arrival date from the USDA map (X_2) and our predicted arrival from the newspaper data (X_1). We label the treatment variable used by Clay et al. (2019) as X_2 as the results presented in Table 1 suggest that, for their application, the map-based measure contains less measurement error than that based on our newspaper data.[17] Each column corresponds to different specifications in Table 3 of Clay et al. (2019). Columns 1-4 report the impact of the boll weevil on pellagra deaths, and Columns 5-8 repeat the same exercise using the log pellagra death rate as outcome. The table reports estimates of θ_1 in equation (7), and we return to θ_2 below. The first row reports the OLS ($\widehat{\beta}_{OLS,X_1}$) results for our newspaper-based arrival date treatment. These coefficient estimates are statistically significant and of the same sign as those provided by Clay et al. (2019), except for one statistically insignificant coefficient in Column 4 (same sign, p-value = .11). The second row for $\widehat{\beta}_{OLS,X_2}$ is the replication of Table 3 in Clay et al. (2019). The following rows report the coefficient estimates for each specification using the agreement sample, the parametric bias correction, and the IV regressions respectively. Due to the inclusion of the interaction term, in Columns 2 to 4 and Columns 6 to 8, the bias-correction estimate using equation (5) was only produced for the specifications in Columns 1 and 5. However, the agreement sample approach is still valid under the interaction term models. For the IV models, we follow the standard approach of using the interacted instrument to instrument for the interaction itself. While the IV interaction models don't technically fit the

[17] This distinction is based on the relative differences between the OLS and IV estimates as discussed in Section 3.

analysis in Section 3, the basic intuition still holds and we believe that a comparison of the IV coefficients remains informative.

Focusing on the main effect, θ_1, we draw four main conclusions from our results. First, as might be expected, our newspaper-based arrival measure appears to be more noisy than that provided by the map. Nonetheless we achieve similar, though smaller, results compared to those of Clay et al. (2019). Thus, in the absence of the USDA map, Clay et al. (2019) could have successfully conducted their study using information from newspaper data alone - highlighting the usefulness of digitized historical newspapers as a potential data source for economic historians. Second, the relationship between the various coefficient estimates are consistent with the prediction provided in equation (6) of our theoretical section. To illustrate this point more clearly, we visualize Column 1 of Table 1 as a bar chart in Figure 5. Third, for all 8 columns, coefficient estimates from the agreement sample and parametric bias correction models are on the order of 40%-60% larger than the original estimates of Clay et al. (2019), suggesting marked gains from our measurement error corrections. Finally, we note that in the two cases where we can implement our parametric bias correction model these coefficient estimates are quite similar in magnitude to the agreement sample estimates.

The above discussion focused on the estimated main effect, θ_1. To account for the interaction term, θ_2, in Table 2 we report the estimated marginal boll-weevil impact for counties in the top 25th percentile of cotton production (Columns 2, 4, 6, and 8) and pellagra deaths (Columns 3 and 7).[18] These results mimic those from Table 1. In all but one model, Column 4 (p-value = .11), we obtain slightly attenuated but significant results based solely on the newspaper data. In all models, the agreement sample estimates are highly significant and larger in magnitude than those reported by Clay et al. (2019). The pattern of the IV estimates exactly match the predictions from Section 3.

4.2 Replication of Ager et al. (2017)

To further validate our approach, we replicate a second paper - that of Ager et al. (2017). They study the boll weevil's effect on Southern agriculture in terms of output, labor arrangements, and labor market outcomes using data from 13 Southern states between 1889 and 1929 in five and ten

[18] Here we are reporting on the linear combination $\theta_1 + \theta_2$. Thus, in Columns 1 and 5 we just replicate the exact results from Table 1.

year intervals.[19] The authors show that the boll weevil reduced cotton output and productivity, the number of tenant farms, farm wages, and female labor force participation. They estimate the following linear regression model,

$$y_{ct} = \alpha_c + \beta_t + \gamma \text{BollWeevil}_{ct} + \delta \text{BollWeevil}_{ct} \times \text{Cotton}_{c,1889} + \epsilon_{ct} \tag{8}$$

where y_{ct} is a given outcome variable for county c in a given five year period t. As in the previous study, BollWeevil_{ct} is an indicator for whether a county is infested in the current five year period. $\text{Cotton}_{c,1889}$ is the demeaned acreage share of cotton planted in 1889 as measure of cotton intensity. County and time fixed effects are captured by α_c and β_t, and standard errors are again clustered at the county level. Because Ager et al. (2017) estimate models incorporating interaction terms in all specifications, we are not able to implement the bias correction model, β_{BC} and we thus focus attention on the agreement sample results as our preferred model.

Table 3 reports the resulting γ coefficients from estimating equation (8).[20] Ager et al. (2017) find significant main effects in 7 of the 12 models that they estimate. Using only our newspaper data, we also find significant results in each of these 7 models - with our newspaper-based coefficient estimates being larger in magnitude for all but 2 of these models. Further, the newspaper data leads to significant estimates of the main effect in 3 of the 5 models where Ager et al. (2017) find no significant effect. For this reason, we keep the same notation in terms of X_1 and X_2 as in Tables 1 and 2 (with X_1 reflecting the newspaper-based data). In 6 of the 7 models where Ager et al. (2017) find statistically significant main effects the agreement sample point estimates, $\beta_{X_1=X_2}$, are larger in magnitude than those based solely on either the map data or the newspaper data - the exception being the estimated effect on corn yield in Column 7. Notice that in all seven of these models the overall pattern of the OLS and IV estimates match the predictions of equation (6). The only exception is Column 7 where the agreement sample estimate is slightly below that of the map-based OLS estimate.

To account for the continuous interaction terms in Ager et al. (2017) in Table 4 we present estimated marginal effects at the 75th percentile of cotton production.[21] The overall pattern of results is

[19] These are 1889, 1899, 1909, 1919, 1924, and 1929.

[20] Because Ager et al. (2017) demean the cotton production data before constructing their interaction measures, γ represents that marginal effect at the mean level of cotton production.

[21] The table summarizes the linear combination $\gamma + .165 * \delta$.

similar to that of Table 3. The newspaper-based treatment yields significant OLS results in 8 of the 9 cases where the map-based data gives significant results. In 5 of these cases the newspaper-based data leads to larger OLS estimates. The newspaper data also leads to significant OLS results in the 3 models where the map-based data did not find significant results. Focusing on the agreement sample, this specification yields estimated marginal effects that are larger in magnitude than either newspaper-based or map-based OLS estimates in 10 of the 12 models. Within the eight models where both data sets have predictive power, agreement sample estimates are on average 37 percent larger than the original estimates of Ager et al. (2017). While not quite as uniform in nature as with the other three sets of results, the IV based marginal estimates in Table 4 generally follow the pattern predicted in equation (6).

4.3 Sensitivity Analysis

As a sensitivity analysis we replicate our main analysis (Tables 1 and 3) under the alternative variable definitions discussed above in Section 2. In particular, we test different approaches to constructing the newspaper-based boll weevil arrival measure: including the three and seven years moving averages, as well as the maximum of the raw salience measure in equation (1) in a 10 year window around the USDA map arrival date. Because we feel the most important issue is the ability to substantially reduce measurement error, for parsimony, we focus our attention in the sensitivity analysis on the estimated coefficients based on the agreement sample.

The agreement sample coefficients using the different measures of the newspaper-based boll weevil arrival for the replication of Clay et al. (2019) are plotted in Figure 6. The same exercise for the replication of Ager et al. (2017) is plotted in Figure 7. Each bar presents the estimated coefficient using a given measure with error bars reporting their corresponding 95% confidence intervals. The red crosses indicate the coefficient value in the original study that was replicated.

Figure 6 shows coefficient estimates for each alternative measure of the newspaper-based boll weevil arrival for Columns 1-8 of the corresponding Table 1. We observe that the estimates using the agreement sample, $\widehat{\beta}_{OLS,X_1=X_2}$ (the green bars), are robust to different measures of the newspaper-based boll weevil arrival date. Moreover, the point estimates based on these measures are bigger than $\widehat{\beta}_{OLS,X_2}$ (the red crosses) regardless of specification. In addition to the agreement sample estimates, columns C1* and C5* of Figure 6 present the bias-corrected estimates, $\widehat{\beta}_{BC}$, using alternative specifications corresponding to Columns 1 and 5 of Table 1. While these coefficients are less precisely estimated in some specifications, the magnitude of each point estimate remains bigger than that of the OLS estimates from the USDA map.

Figure 7 repeats the exercise in Figure 6 using the agreement sample for Ager et al. (2017). We report $\widehat{\beta}_{OLS,X_1=X_2}$ as well as the OLS estimates based on the arrival from the USDA map, which is $\widehat{\beta}_{OLS,X_2}$ in Table 3. Again, Figure 7 shows very similar $\widehat{\beta}_{OLS,X_1=X_2}$ estimates, regardless of different specifications and outcome variables. As in the main replication exercise, $\widehat{\beta}_{OLS,X_1=X_2}$ is bigger than $\widehat{\beta}_{OLS,X_2}$ (the red crosses). Here, not only are the agreement sample estimates uniformly larger in magnitude than the original estimates of Ager et al. (2017), but in all but one case the entire 95 percent confidence interval lies above the original point estimates.

4.4 Discussion

These two replications have shown that newspaper data can be gainfully used for bias reduction in statistical analyses using historical data. Both cases have demonstrated that the predictions based on the inequality in equation (6) hold up in applied examples. The gains in bias reduction appear to have been larger in the replication of Ager et al. (2017) as compared to the replication of Clay et al. (2019). While we cannot offer a definitive explanation for this outcome, a possible reason seems to be the difference in the frequency of the time dimension. The study by Clay et al. (2019) uses annual data, a much higher frequency than the five year intervals in Ager et al. (2017), which potentially mitigated some of the measurement error bias. Nonetheless, results in both papers held up in our replications and could be strengthened using our methods.

The results above also highlight the value of data extracted from digitized newspapers in general. Our newspaper-based boll weevil arrival measure was generated in a fast and low-cost way. Compared to the USDA measure used by Clay et al. (2019), it appears to be more noisy which is to be expected. It would certainly be possible to refine the measure but doing so would increase the time and cost of collecting the information. What we want to highlight instead is that our very coarse measure still managed to produce very similar results in the two replications, meaning that both studies could have been conducted had the USDA map never existed.

For the purpose of the methods introduced in this paper, it does not matter whether the data from the newspapers or the original variable (here the USDA map arrival date) is noisier as long as the measurement errors in the two variables are uncorrelated. This assumption cannot be directly tested, as with the exclusion restriction in instrumental variable regressions, for instance. To provide an example, we argue that the assumption holds in our setting because newspapers reported any boll weevil related events that were observed by newspaper reporters whereas the USDA map was created by the federal entomologists. Which of the two measures is noisier makes no difference when applying our methods aside from determining the bounds in the set identification solution.

Our approach is particularly suited for measures that can be easily generated or extracted using textual data. Simple n-gram or bag-of-words approaches as in Beach et al. (2020), Ferrara and Fishback (2020), Albright et al. (2021), Beach and Hanlon (2021), Bazzi et al. (2021), or Ottinger and Winkler (2021) are particularly promising. For variables such as prices, this approach is less promising because these can rarely be extracted in a low-cost manner as they oftentimes require more careful extraction, possibly by hand. It is also impractical for variables that would typically not be reported in the news or for which the non-random nature of the availability of digitized newspapers might be a concern. For instance, measures relating to corruption or trade might be more difficult to find in newspapers. Nevertheless, newspapers can be a great data source when one wishes to study large scale events because those are more likely to be covered in newspapers. Our boll weevil infestation example fits into this category as Lange et al. (2009) described "the boll weevil is America's most celebrated agricultural pest" (p. 685). Another example along this line studied in previous literature is the 1918 Influenza Pandemic (Beach et al., 2020). Newspapers are also useful for the sensational events that had extensive media coverage such as the tragic Tulsa race massacre in 1921 (Albright et al., 2021) or the famous Bradlaugh-Besant trial of 1877 (Beach and Hanlon, 2021).

5 Conclusion

Measurement error in historical data is often a source of bias in statistical analyses which leads to attenuation bias in the relationships that researchers seek to identify. When measurement error is classical, it is known that this attenuation bias can be removed via an instrumental variable approach. A potential instrument is a second measure of the same variable with errors, as long as the errors in two variables are uncorrelated (Chalfin and McCrary, 2018). Generating such a second measure tends to be expensive and therefore measurement error tends to be ignored as long as some conventional level of statistical significance is achieved.

In this paper, we introduce the idea of cheaply generating such a second measure from digitized newspapers, which can be scraped or downloaded at low costs. We show how a newspaper-based secondary measure can be used to deal with measurement error when the variable of interest is either continuous or binary. The latter case is more challenging since measurement error in a binary variable is non-classical by construction and therefore an instrumental variables approach alone does not remove the associated bias (Bingley and Martinello, 2017). Instead, we propose three alternative methods for dealing with measurement error in this setting based on i) set identification, ii) using an agreement sample where both the primary and secondary measure give the same answer, and iii) a parametric bias correction that can be obtained as nonlinear combination of the OLS and IV coefficients. Our theory predicts that OLS and IV provide the lower and upper bounds of the identified set that includes the true parameter, and that the coefficients from the agreement sample and the parametric bias correction should lie in between these bounds. Also, the bias corrected estimate should still be larger in magnitude than the OLS coefficient from the agreement sample.

To test this prediction as well as to showcase our methods, we replicate two recent papers by Clay et al. (2019) and Ager et al. (2017) on the impact of the boll weevil infestation in the U.S. South between 1892 and 1922. Like most studies on the boll weevil, the main treatment is measured from a map of the pest by Hunter and Coad (1923), which arguable is measured with error because of crossing lines and given that the arrival dates are an imperfect measure of the economic impact of the beetle. To produce a second measure for the boll weevil arrival from digitized newspaper data, we scrape Newspapers.com and search for pages that mention "boll weevil" and each county's name from all newspapers in the county's state. This approach maximizes the chance to find articles related to the arrival of the weevil in that county. In both replications, we find larger coefficients than in the original studies which show the usefulness of our approach to dealing with measurement error and also reaffirm the main results of the two papers. In both cases we also find the patterns prescribed in the theoretical section, where plain OLS yields the smallest coefficient, followed by the agreement sample, and the parametric bias correction.

The main contribution of the paper is to provide an easy way to generate a secondary measure for a given mismeasured variable of interest and to show how this secondary measure can be used to remove attenuation bias resulting from measurement error. We extend the framework in Chalfin and McCrary (2018) for classical measurement error to the case where a variable is binary. The emphasis is on the newspaper data being easily available, which substantially reduces the cost of generating a secondary measure for bias correction purposes that is usually the main prohibiting factor for researchers to apply such methods. We also contribute to a recent literature that has highlighted the usefulness of historical newspapers to generate novel data for the purpose of research in economic history.

References

Ager, Philipp, Benedikt Herz, and Markus Brueckner, “Structural Change and the Fertility Transition,” *Review of Economics and Statistics*, 2020, *102* (4), 806–822.

__ , Markus Brueckner, and Benedikt Herz, “The Boll Weevil Plague and its Effect on the Southern Agricultural Sector, 1889–1929,” *Explorations in Economic History*, 2017, *65*, 94–105.

Aigner, Dennis J., “Regression with a Binary Independent Variable subject to Errors of Observation,” *Journal of Econometrics*, 1973, *1* (1), 49–59.

Albright, Alex, Jeremy A. Cook, James J. Feigenbaum, Laura Kincaide, Jason Long, and Nathan Nunn, “After the Burning: The Economic Effects of the 1921 Tulsa Race Massacre,” *NBER Working Paper No. 28985*, 2021.

Bailey, Martha, Connor Cole, and Catherine Massey, “Simple strategies for improving inference with linked data: a case study of the 1850–1930 IPUMS linked representative historical samples,” *Historical Methods*, 2019, *53* (2), 80–93.

Baker, Richard B., “From the Field to the Classroom: The Boll Weevil’s Impact on Education in Rural Georgia,” *Journal of Economic History*, 2015, *75* (4), 1128–1160.

__ , John Blanchette, and Katherine Eriksson, “Long-Run Impacts of Agricultural Shocks on Educational Attainment: Evidence from the Boll Weevil,” *Journal of Economic History*, 2020, *80* (1), 136–174.

Bazzi, Samuel, Andreas Ferrara, Martin Fiszbein, Thomas P. Pearson, and Patrick A. Testa, “The Other Great Migration: Southern Whites and the New Right,” *NBER Working Paper No. 29506*, 2021.

Beach, Brian and Walker W. Hanlon, “Culture and the Historical Fertility Transition,” *working paper*, 2021.

__ , Karen Clay, and Martin H. Saavedra, “The 1918 Influenza Pandemic and Its Lessons for Covid-19,” *NBER Working Paper No. 27673*, 2020.

Bingley, Paul and Alessandro Martinello, “Measurement Error in Income and Schooling and the Bias of Linear Estimators,” *Journal of Labor Economics*, 2017, *35* (4), 1117–1148.

Bloome, Deirdre, James Feigenbaum, and Christopher Muller, “Tenancy, Marriage, and the Boll Weevil Infestation, 1892–1930,” *Demography*, 2017, *54* (3), 1029–1049.

Calderon, Alvaro, Vasiliki Fouka, and Marco Tabellini, “Racial Diversity and Racial Policy Preferences: The Great Migration and Civil Rights,” *NBER Working Paper No. 28965*, 2021.

Chalfin, Aaron and Justin McCrary, “Are U.S. Cities Underpoliced? Theory and Evidence,” *Review of Economics and Statistics*, 2018, *100* (1), 167–186.

Clay, Karen, Ethan Schmick, and Werner Troesken, “The Rise and Fall of Pellagra in the American South,” *Journal of Economic History*, 2019, *79* (1), 32–62.

__ , __ , and __ , “The Boll Weevil’s Impact on Racial Income Gaps in the Early Twentieth Century,” *NBER Working Paper No. 27101*, 2020.

Dippel, Christian and Bryan Leonard, “Not-so-Natural Experiments in History,” *Journal of Historical Political Economy*, 2021, *1* (1), 1–30.

Dupraz, Yannick and Andreas Ferrara, “Fatherless: The Long-Term Effects of Losing a Father in the US Civil War,” *CAGE Working Paper No. 538*, 2021.

Esposito, Elena, Tiziano Rotesi, Alessandro Saia, and Mathias Theonig, “Reconciliation Narratives: The Birth of a Nation after the US Civil War,” *CEPR Working Paper No. 15938*, 2014.

Feigenbaum, James J., Soumyajit Mazumder, and Cory B. Smith, "When Coercive Economies Fail: The Political Economy of the US South After the Boll Weevil," *NBER Working Paper No. 27161*, 2020.

Ferrara, Andreas and Price V. Fishback, "Discrimination, Migration, and Economic Outcomes: Evidence from World War I," *NBER Working Paper No. 26936*, 2020.

__, Patrick A. Testa, and Liyang Zhou, "New area- and population-based geographic crosswalks for U.S. counties and congressional districts, 1790-2020," *CAGE Working Paper No. 588*, 2021.

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson, "Competition and Ideological Diversity: Historical Evidence from US Newspapers," *American Economic Review*, 2014, *104* (10), 3073–3114.

__, Nathan Petek, Jesse M. Shapiro, and Michael Sinkinson, "Do Newspapers Serve the State? Incumbent Party Influence on the US Press, 1869-1928," *Journal of the European Economic Association*, 2015, *13* (1), 29–61.

Harned, Robey W., *Boll Weevil in Mississippi, 1909*, Mississippi Agricultural Experiment Station, 1910.

Hunter, Walter D. and Bert R. Coad, *The Boll-Weevil Problem*, U.S. Department of Agriculture, Washington D.C., 1923.

Lange, Fabian, Alan L. Olmstead, and Paul W. Rhode, "The Impact of the Boll Weevil, 1892–1932," *Journal of Economic History*, 2009, *69* (3), 685–718.

Meyer, Bruce D. and Nikolas Mittag, "Misclassification in Binary Choice Models," *Journal of Econometrics*, 2017, *200* (2), 295–311.

Ottinger, Sebastian and Max Winkler, "The Political Economy of Propaganda: Evidence from U.S. Newspapers," *mimeo*, 2021.

Rhode, Paul W., "Biological Innovation without Intellectual Property Rights: Cottonseed Markets in the Antebellum American South," *Journal of Economic History*, 2021, *81* (1), 198–238.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2020, *forthcoming.*

Wang, Tianyi, "The Electric Telegraph, News Coverage and Political Participation," *mimeo*, 2019.

Wright, Gavin, *Sharing the Prize: The Economics of the Civil Rights Revolution in the American South*, Belknap Press, Cambridge, MA, 2013.

Tables

Table 1: Replication of Clay et al. (2019) - Main Effects

	Log Pellagra Deaths				Log Pellagra Death Rate			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\widehat{\beta}_{OLS,X_1}$	-0.183***	-0.142*	-0.150**	-0.122	-0.151***	-0.113**	-0.144**	-0.125**
	(0.068)	(0.072)	(0.075)	(0.077)	(0.052)	(0.055)	(0.058)	(0.058)
$\widehat{\beta}_{OLS,X_2}$	-0.283***	-0.197***	-0.237***	-0.202***	-0.235***	-0.161***	-0.212***	-0.185***
	(0.059)	(0.065)	(0.065)	(0.063)	(0.046)	(0.050)	(0.050)	(0.047)
$\widehat{\beta}_{OLS,X_1=X_2}$	-0.396***	-0.310***	-0.333***	-0.278***	-0.326***	-0.251***	-0.295***	-0.256***
	(0.093)	(0.097)	(0.099)	(0.101)	(0.074)	(0.076)	(0.078)	(0.078)
$\widehat{\beta}_{BC}$	-0.410***				-0.340***			
	(0.101)				(0.080)			
$\widehat{\beta}_{IV,X_2}$	-0.595**	-0.460**	-0.427**	-0.346*	-0.493***	-0.371**	-0.401**	-0.346**
	(0.231)	(0.216)	(0.207)	(0.206)	(0.173)	(0.164)	(0.159)	(0.158)
$\widehat{\beta}_{IV,X_1}$	-1.073***	-1.058***	-1.092***	-1.094***	-0.892***	-0.879***	-0.893***	-0.893***
	(0.260)	(0.275)	(0.259)	(0.269)	(0.208)	(0.221)	(0.202)	(0.207)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High pellagra		Yes				Yes		
BW × High cotton			Yes	Yes			Yes	Yes
Controls				Yes				Yes
Obs.	1,312	1,312	1,312	1,312	1,312	1,312	1,312	1,312
Counties	141	141	141	141	141	141	141	141
Obs. ($X_1 = X_2$)	1,051	1,051	1,051	1,051	1,051	1,051	1,051	1,051

Note: Replication of equation (1) in Clay et al. (2019) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). Columns 1 and 5 report OLS and IV regressions of deaths by pellagra on an indicator for whether the boll weevil has arrived in county c. The coefficients β_{BC} are estimated using equation (5) and the delta method. The rest of columns report OLS and IV regressions of deaths by pellagra on an boll weevil indicator and its interaction term with an indicator for whether county c was in top 25% cotton production in 1909 (Columns 2, 4, 6, and 8) or a dummy variable equal to one if county c was in top 25% pellagra death rates in 1915-1916 (Columns 3 and 7). The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e. an agreement sample). In IV regressions, X_1 is instrumented with X_2 and vice versa. The sample is 141 counties in North Carolina and South Carolina between 1915 and 1925. All regressions include county and year fixed effects. Controls include county c's malaria death rate in 1915 and the share of urban population in 1910 both interacted with a full set of year dummies. Standard errors are clustered at the county level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Replication of Clay et al. (2019) - Marginal Effects at the 75th Percentile

	Log Pellagra Deaths				Log Pellagra Death Rate			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\widehat{\beta}_{OLS,X_1}$	-0.183***	-0.278***	-0.259***	-0.253***	-0.151***	-0.241***	-0.169***	-0.169***
	(0.068)	(0.096)	(0.083)	(0.084)	(0.052)	(0.070)	(0.059)	(0.058)
$\widehat{\beta}_{OLS,X_2}$	-0.283***	-0.531***	-0.442***	-0.469***	-0.235***	-0.452***	-0.314***	-0.335***
	(0.059)	(0.086)	(0.083)	(0.080)	(0.046)	(0.067)	(0.063)	(0.060)
$\widehat{\beta}_{OLS,X_1=X_2}$	-0.396***	-0.652***	-0.603***	-0.595***	-0.326***	-0.551***	-0.429***	-0.428***
	(0.093)	(0.120)	(0.110)	(0.110)	(0.074)	(0.094)	(0.084)	(0.082)
$\widehat{\beta}_{BC}$	-0.410***				-0.340***			
	(0.101)				(0.080)			
$\widehat{\beta}_{IV,X_2}$	-0.595**	-0.806***	-0.817***	-0.750***	-0.493***	-0.682***	-0.613***	-0.579***
	(0.231)	(0.280)	(0.269)	(0.268)	(0.173)	(0.205)	(0.199)	(0.196)
$\widehat{\beta}_{IV,X_1}$	-1.073***	-1.476***	-1.221***	-1.271***	-0.892***	-1.247***	-0.900***	-0.938***
	(0.260)	(0.280)	(0.246)	(0.247)	(0.208)	(0.227)	(0.188)	(0.187)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High pellagra		Yes				Yes		
BW × High cotton			Yes	Yes			Yes	Yes
Controls				Yes				Yes
Obs.	1,312	1,312	1,312	1,312	1,312	1,312	1,312	1,312
Counties	141	141	141	141	141	141	141	141
Obs. ($X_1 = X_2$)	1,051	1,051	1,051	1,051	1,051	1,051	1,051	1,051

Note: Replication of equation (1) in Clay et al. (2019) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). Columns 1 and 5 report OLS and IV regressions of deaths by pellagra on an indicator for whether the boll weevil has arrived in county c. The coefficients β_{BC} are estimated using equation (5) and the delta method. The rest of columns report OLS and IV regressions of deaths by pellagra on an boll weevil indicator and its interaction term with an indicator for whether county c was in top 25% cotton production in 1909 (Columns 2, 4, 6, and 8) or a dummy variable equal to one if county c was in top 25% pellagra death rates in 1915-1916 (Columns 3 and 7). The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e. an agreement sample). In IV regressions, X_1 is instrumented with X_2 and vice versa. The sample is 141 counties in North Carolina and South Carolina between 1915 and 1925. All regressions include county and year fixed effects. Controls include county c's malaria death rate in 1915 and the share of urban population in 1910 both interacted with a full set of year dummies. Standard errors are clustered at the county level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Replication of Ager et al. (2017) - Main Effects

	Log Cotton Production				Log Corn Production				Log Other Outcomes			
	Bales	Acres	Yield	Share	Bushels	Acres	Yield	Share	Farm	Farm Value	Pop.	Black Pop.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\widehat{\beta}_{OLS,X_1}$	-0.486***	-0.245***	-0.248***	-0.066***	0.037	0.077***	-0.040**	0.053***	-0.025**	-0.038**	0.024*	-0.028
	(0.057)	(0.052)	(0.022)	(0.006)	(0.029)	(0.019)	(0.017)	(0.006)	(0.011)	(0.015)	(0.013)	(0.034)
$\widehat{\beta}_{OLS,X_2}$	-0.386***	-0.173***	-0.208***	-0.061***	-0.032	0.045**	-0.077***	0.064***	-0.000	0.005	-0.002	0.028
	(0.055)	(0.049)	(0.024)	(0.007)	(0.037)	(0.023)	(0.021)	(0.007)	(0.012)	(0.018)	(0.014)	(0.030)
$\widehat{\beta}_{OLS,X_1=X_2}$	-0.651***	-0.290***	-0.360***	-0.091***	0.055	0.118***	-0.062**	0.087***	-0.028**	-0.023	0.001	-0.028
	(0.060)	(0.054)	(0.025)	(0.008)	(0.041)	(0.023)	(0.026)	(0.007)	(0.013)	(0.017)	(0.016)	(0.042)
$\widehat{\beta}_{IV,X_2}$	-1.069***	-0.532***	-0.551***	-0.141***	0.090	0.177***	-0.087**	0.114***	-0.052**	-0.075**	0.080**	-0.048
	(0.123)	(0.112)	(0.047)	(0.014)	(0.066)	(0.043)	(0.039)	(0.014)	(0.026)	(0.037)	(0.034)	(0.079)
$\widehat{\beta}_{IV,X_1}$	-0.939***	-0.485***	-0.447***	-0.135***	-0.074	0.091**	-0.165***	0.139***	-0.005	0.001	-0.024	0.051
	(0.106)	(0.094)	(0.048)	(0.014)	(0.076)	(0.046)	(0.044)	(0.014)	(0.025)	(0.036)	(0.033)	(0.071)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High cotton	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weather controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	4,323	4,329	4,323	4,440	4,440	4,440	4,440	4,440	4,440	4,440	3,700	3,679
Counties	735	735	735	740	740	740	740	740	740	740	740	739
Obs. ($X_1 = X_2$)	3,927	3,933	3,927	4,032	4,032	4,032	4,032	4,032	4,032	4,032	3,328	3,311

Note: Replication of equation (1) in Ager et al. (2017) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). OLS and IV regressions of agricultural and demographic outcome variables on an indicator for whether the boll weevil has arrived in county c and its interaction term with county c's acreage share of cotton in 1889. The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e. an agreement sample). In the IV regressions, X_1 is instrumented with X_2 and vice versa. The sample includes counties in the U.S. South between 1889 and 1929. All regressions include county and year fixed effects as well as weather controls. Weather controls are January's mean temperature and average summer precipitation from May to July. Standard errors are clustered at the county level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

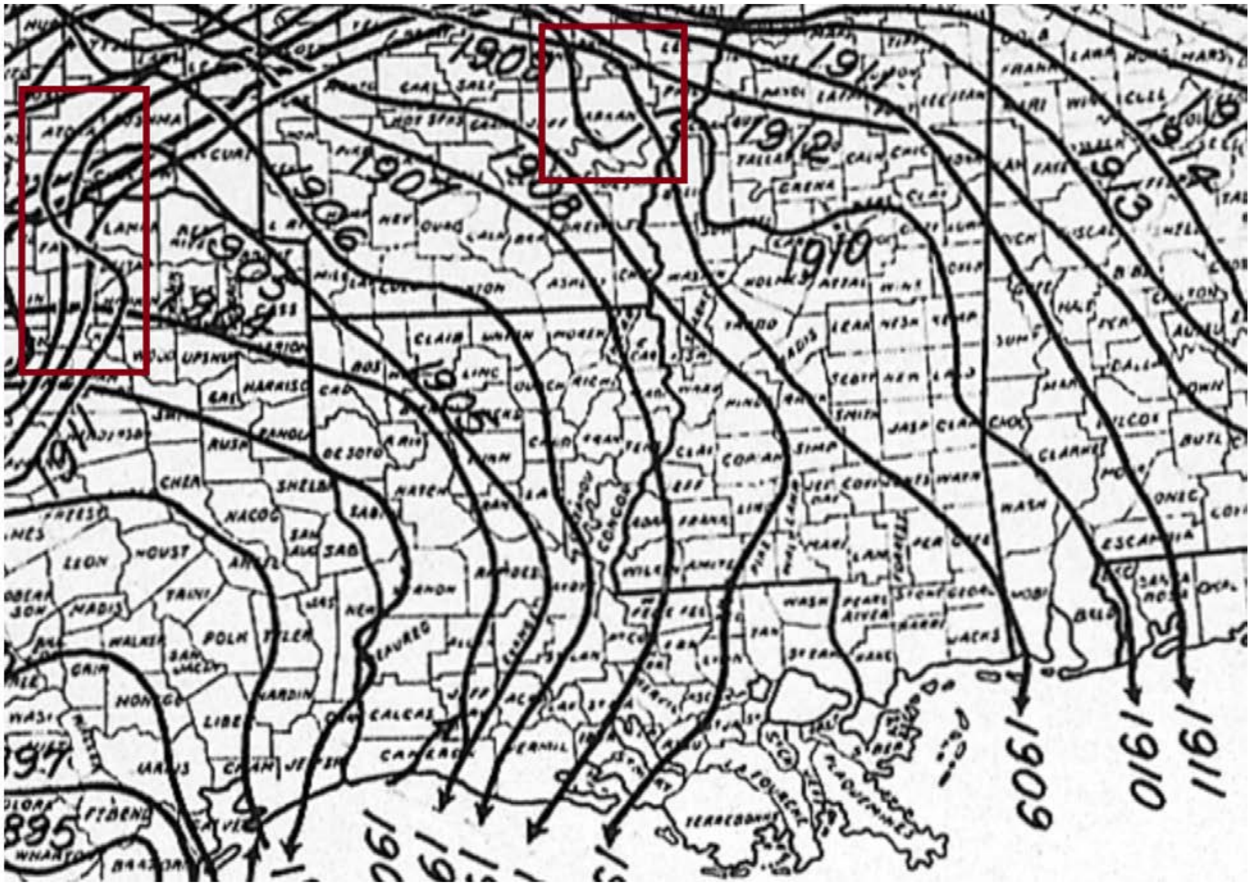
Table 4: Replication of Ager et al. (2017) - Marginal Effects at the 75th Percentile

	Log Cotton Production				Log Corn Production				Log Other Outcomes			
	Bales	Acres	Yield	Share	Bushels	Acres	Yield	Share	Farm	Farm Value	Pop.	Black Pop.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\widehat{\beta}_{OLS,X_1}$	-1.008***	-0.713***	-0.311***	-0.116***	-0.006	0.056***	-0.062***	0.086***	-0.051***	-0.106***	-0.073***	-0.099**
	(0.058)	(0.054)	(0.023)	(0.007)	(0.031)	(0.021)	(0.018)	(0.007)	(0.014)	(0.019)	(0.017)	(0.038)
$\widehat{\beta}_{OLS,X_2}$	-0.918***	-0.638***	-0.287***	-0.108***	-0.072*	0.038	-0.110***	0.098***	-0.026	-0.039*	-0.076***	-0.028
	(0.063)	(0.058)	(0.025)	(0.008)	(0.039)	(0.026)	(0.021)	(0.008)	(0.017)	(0.021)	(0.018)	(0.033)
$\widehat{\beta}_{OLS,X_1=X_2}$	-1.223***	-0.796***	-0.437***	-0.143***	0.011	0.103***	-0.092***	0.122***	-0.057***	-0.086***	-0.094***	-0.104**
	(0.066)	(0.061)	(0.026)	(0.009)	(0.044)	(0.027)	(0.026)	(0.008)	(0.018)	(0.022)	(0.020)	(0.047)
$\widehat{\beta}_{IV,X_2}$	-1.712***	-1.098***	-0.639***	-0.203***	0.040	0.156***	-0.115***	0.156***	-0.084***	-0.158***	-0.034	-0.133
	(0.119)	(0.109)	(0.048)	(0.015)	(0.067)	(0.043)	(0.039)	(0.015)	(0.026)	(0.039)	(0.033)	(0.082)
$\widehat{\beta}_{IV,X_1}$	-1.566***	-1.037***	-0.537***	-0.190***	-0.121	0.082*	-0.202***	0.178***	-0.036	-0.052	-0.119***	-0.022
	(0.113)	(0.100)	(0.048)	(0.016)	(0.077)	(0.048)	(0.044)	(0.015)	(0.030)	(0.040)	(0.036)	(0.071)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High cotton	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weather controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	4,323	4,329	4,323	4,440	4,440	4,440	4,440	4,440	4,440	4,440	3,700	3,679
Counties	735	735	735	740	740	740	740	740	740	740	740	739
Obs. ($X_1 = X_2$)	3,927	3,933	3,927	4,032	4,032	4,032	4,032	4,032	4,032	4,032	3,328	3,311

Note: Replication of equation (1) in Ager et al. (2017) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). OLS and IV regressions of agricultural and demographic outcome variables on an indicator for whether the boll weevil has arrived in county c and its interaction term with county c's acreage share of cotton in 1889. The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e. an agreement sample). In the IV regressions, X_1 is instrumented with X_2 and vice versa. The sample includes counties in the U.S. South between 1889 and 1929. All regressions include county and year fixed effects as well as weather controls. Weather controls are January's mean temperature and average summer precipitation from May to July. Standard errors are clustered at the county level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

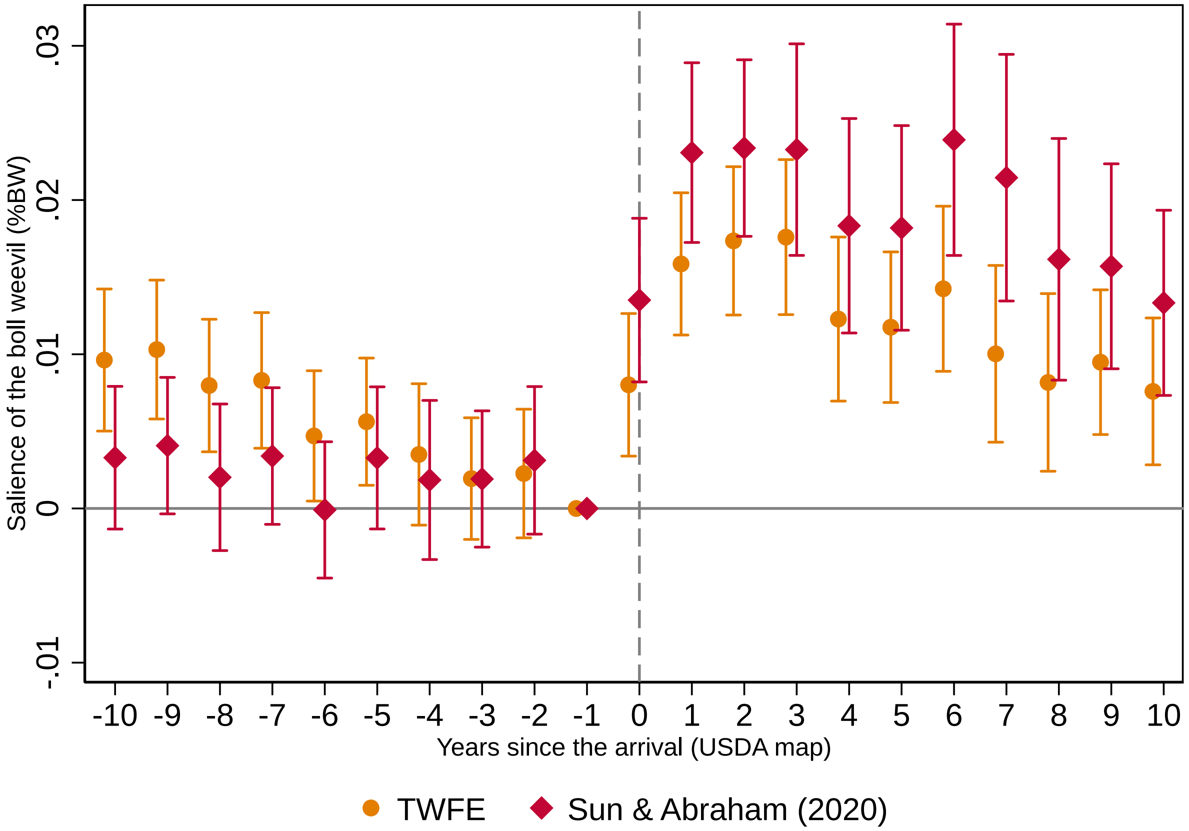
Figures

Figure 1: Errors in the USDA Map for the Arrival of the Boll Weevil



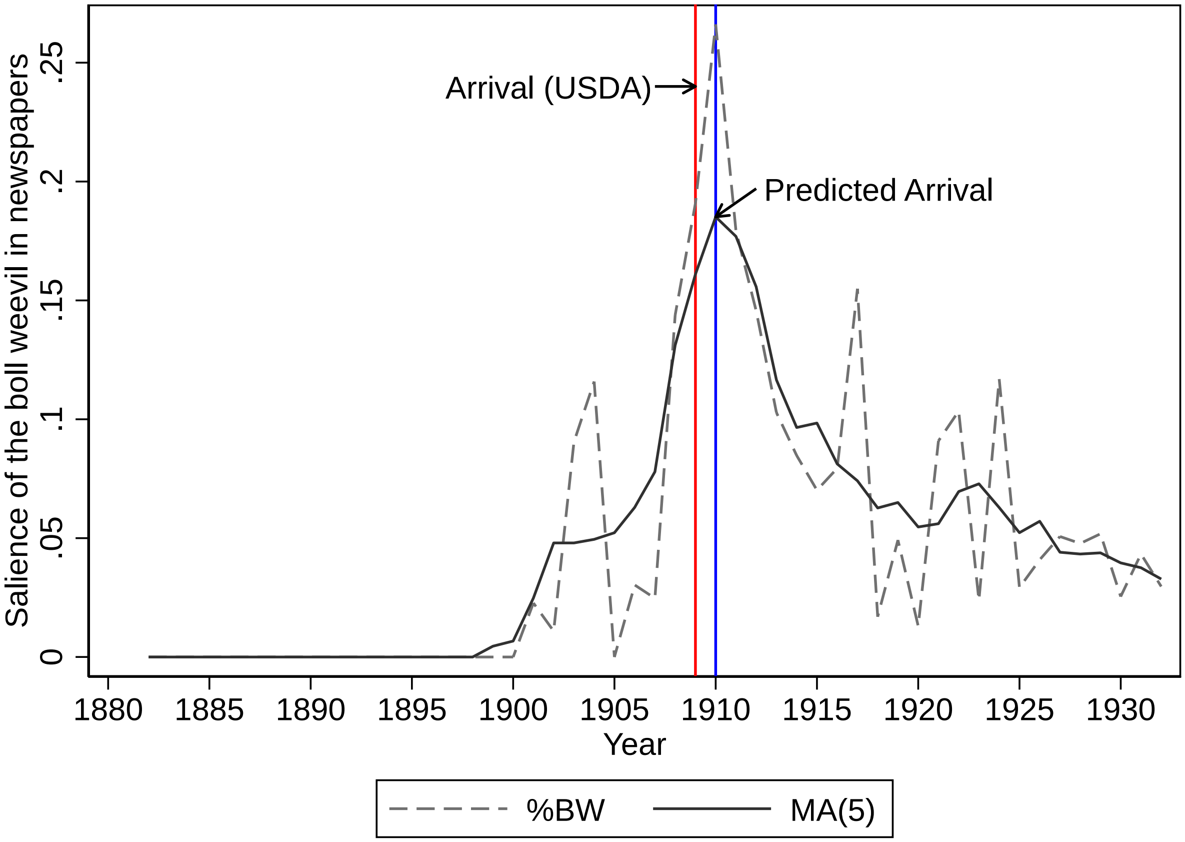
Note: Snipped of the USDA map for the arrival of the boll weevil provided by Hunter and Coad (1923). Each solid line marks the arrival year of the pest. Researchers typically overlay the lines onto a map of Southern counties and determine the arrival date by the line that covers most of the county area. The red boxes highlight areas where date lines cross in contradictory ways.

Figure 2: Event Study Plot - TWFE and Sun and Abraham (2020)



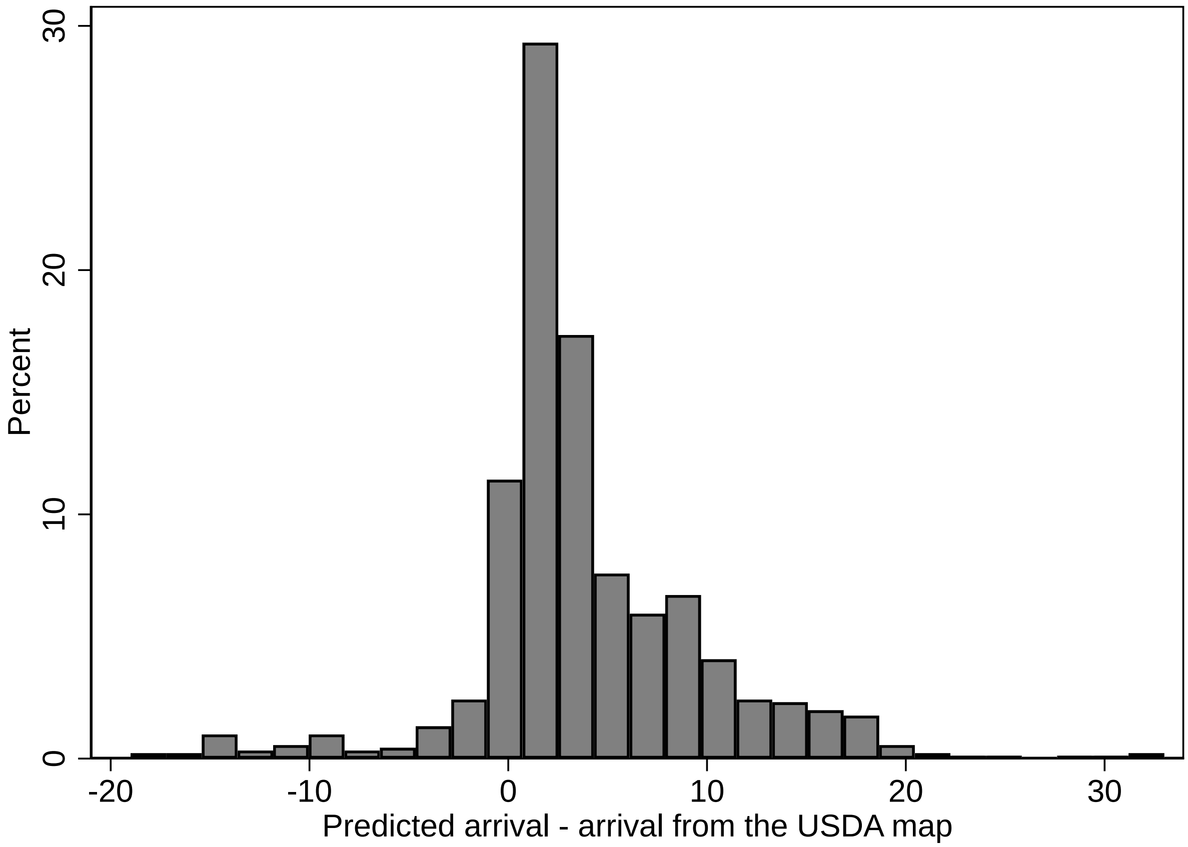
Note: Coefficient plot from an event study regression of %BW on an event indicator relative to the arrival of the boll weevil from the USDA map as well as county and state-by-year fixed effects. Each circle and diamond presents the estimates β_ℓ in equation (2) using OLS and the estimator proposed by Sun and Abraham (2020), respectively. The sample consists of 911 infested counties in 13 Southern states. The omitted baseline period is $\ell = -1$, which is one year before the arrival from the USDA map. The relative time period for the latest-infested counties is omitted as well for the estimates using Sun and Abraham (2020) due to the lack of never-infested counties in our sample. Standard errors are clustered at the county level and 95% confidence intervals are reported around the point estimates.

Figure 3: Salience of the Boll Weevil - Marion County, Mississippi



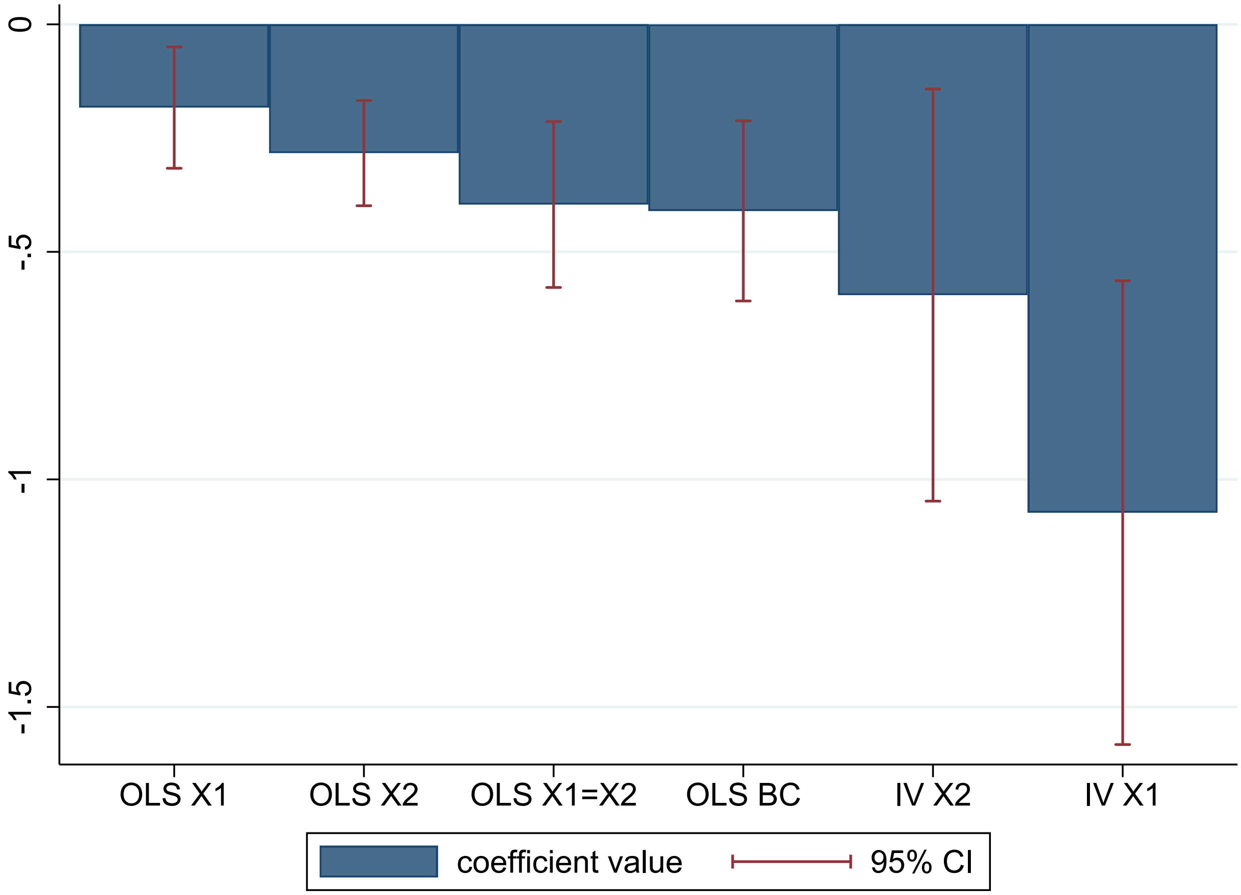
Note: The dashed line is the salience measure of Marion County over time. The salience measure is constructed based every available newspaper outlet in Mississippi between 1882 and 1932. The solid line is 5-year moving averages of the salience measure (MA(5)). The red horizontal line shows the boll weevil's arrival in Marion County from the USDA map. The blue horizontal line indicates the predicted arrival where MA(5) is the highest.

Figure 4: Distribution of Differences between Boll Weevil Measures



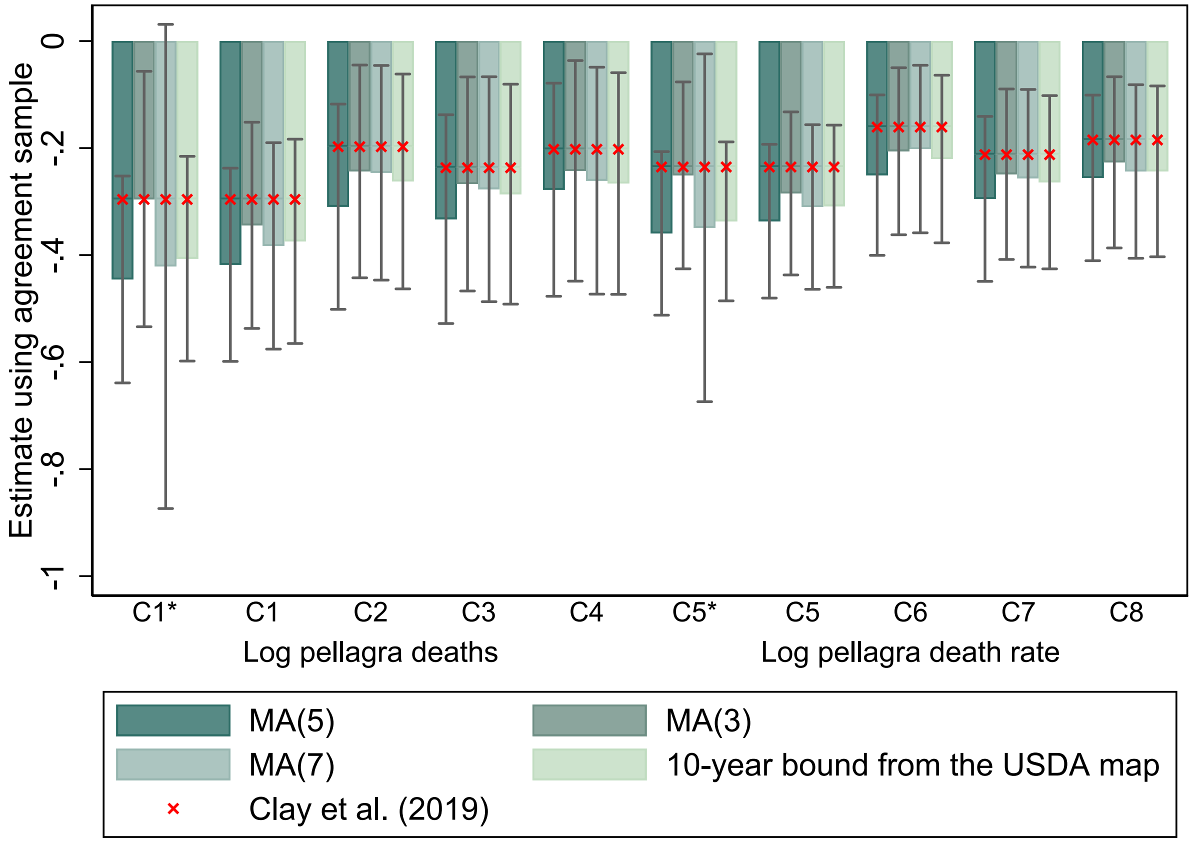
Note: Distribution of the difference between the year of the boll weevil infestation from the USDA map and predicted by newspapers for 911 infested counties in our sample. Each bar indicates the share of counties in our sample for different level of differences. The difference is defined as the predicted year based on our newspaper approach minus the year of arrival from the USDA map. The right-skewed distribution indicates that in general the prediction is later than the arrival from the map.

Figure 5: Visualization of Column 1 in the Replication of Clay et al. (2019)



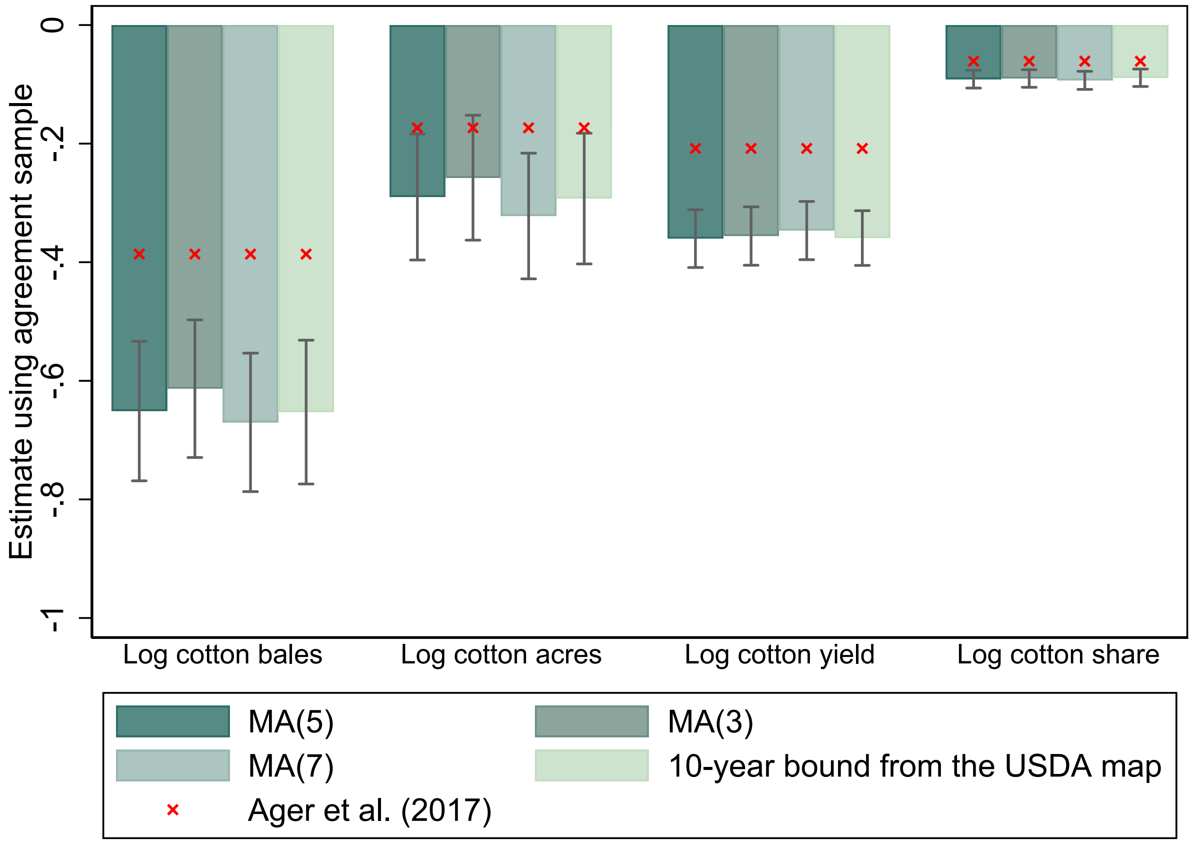
Note: Regression of log pellagra deaths on an indicator for the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1) for the replication of Clay et al. (2019). The figure visualizes the coefficients from our replication exercise in Column 1 of Table 1 to show the ranked pattern according to our theory in Section 3. The variable subscript indicates which variable was used as treatment. The figure confirms the pattern described in the inequality in equation (6).

Figure 6: Sensitivity Analysis for the Replication of Clay et al. (2019)



Note: Each bar in C1 through C8 presents the estimates $\widehat{\beta}_{OLS,X_1=X_2}$ in Table 1 using alternative specifications and 95% confidence intervals are reported. Bars in C1* and C5* show the estimates $\widehat{\beta}_{BC}$ in Table 1 across specifications. Alternative specifications include 3- and 7-year moving averages as well as 10-year bound from the USDA map. Red crosses indicate the OLS estimates based on the arrival from the USDA map (i.e. $\widehat{\beta}_{OLS,X_2}$ in Table 1).

Figure 7: Sensitivity Analysis for the Replication of Ager et al. (2017)

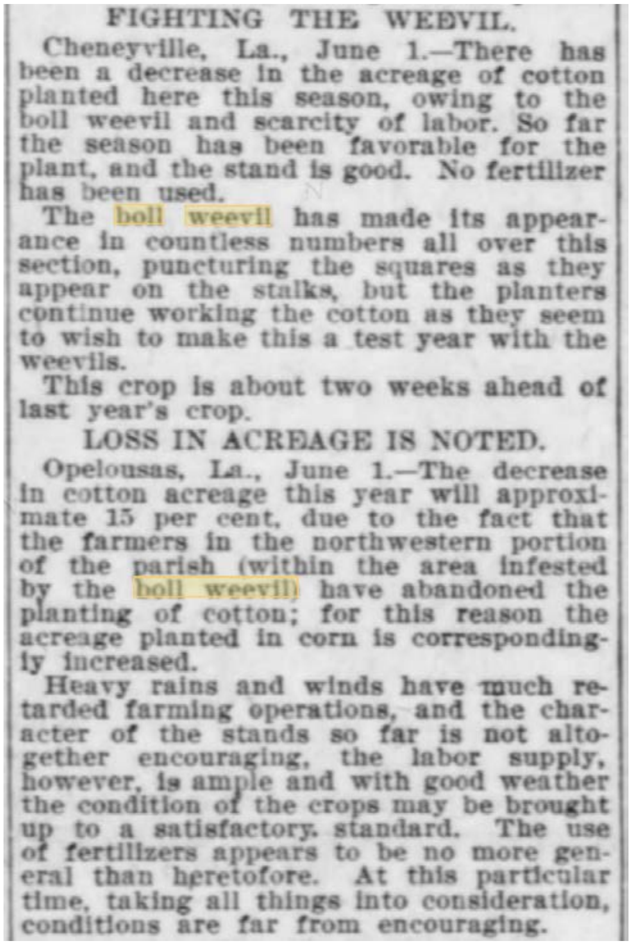


Note: Each bar presents the estimates $\widehat{\beta}_{OLS,X_1=X_2}$ in Table 3 using alternative specifications and 95% confidence intervals are reported. Alternative specifications include 3- and 7-year moving averages as well as 10-year bound from the USDA map. Red crosses indicate the OLS estimates based on the arrival from the USDA map (i.e. $\widehat{\beta}_{OLS,X_2}$ in Table 3).

Online Appendix

Additional Tables and Figures

Figure A.1: Boll Weevil in Newspapers



FIGHTING THE WEEVIL.

Cheneyville, La., June 1.—There has been a decrease in the acreage of cotton planted here this season, owing to the boll weevil and scarcity of labor. So far the season has been favorable for the plant, and the stand is good. No fertilizer has been used.

The boll weevil has made its appearance in countless numbers all over this section, puncturing the squares as they appear on the stalks, but the planters continue working the cotton as they seem to wish to make this a test year with the weevils.

This crop is about two weeks ahead of last year's crop.

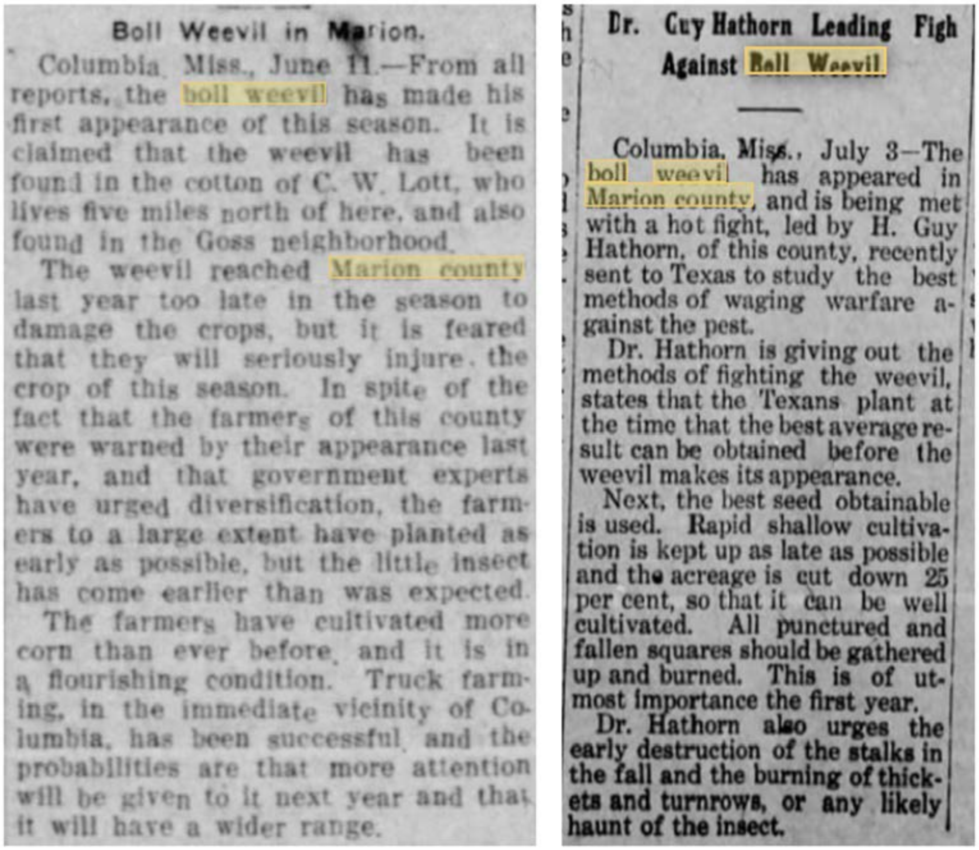
LOSS IN ACREAGE IS NOTED.

Opelousas, La., June 1.—The decrease in cotton acreage this year will approximate 15 per cent, due to the fact that the farmers in the northwestern portion of the parish (within the area infested by the boll weevil) have abandoned the planting of cotton; for this reason the acreage planted in corn is correspondingly increased.

Heavy rains and winds have much retarded farming operations, and the character of the stands so far is not altogether encouraging, the labor supply, however, is ample and with good weather the condition of the crops may be brought up to a satisfactory standard. The use of fertilizers appears to be no more general than heretofore. At this particular time, taking all things into consideration, conditions are far from encouraging.

Note: An example of newspaper articles on the boll weevil infestation published in The Times-Democrat on June 8th, 1908. Newspapers published articles about the boll weevil's arrival as well as damages in cotton production caused by the insects.

Figure A.2: Boll Weevil in Newspapers from Different Counties



Boll Weevil in Marion.

Columbia Miss., June 11.—From all reports, the boll weevil has made his first appearance of this season. It is claimed that the weevil has been found in the cotton of C. W. Lott, who lives five miles north of here, and also found in the Goss neighborhood.

The weevil reached Marion county last year too late in the season to damage the crops, but it is feared that they will seriously injure the crop of this season. In spite of the fact that the farmers of this county were warned by their appearance last year, and that government experts have urged diversification, the farmers to a large extent have planted as early as possible, but the little insect has come earlier than was expected.

The farmers have cultivated more corn than ever before, and it is in a flourishing condition. Truck farming, in the immediate vicinity of Columbia, has been successful and the probabilities are that more attention will be given to it next year and that it will have a wider range.

Dr. Cuy Hathorn Leading Figh Against Boll Weevil

Columbia, Miss., July 3—The boll weevil has appeared in Marion county, and is being met with a hot fight, led by H. Guy Hathorn, of this county, recently sent to Texas to study the best methods of waging warfare against the pest.

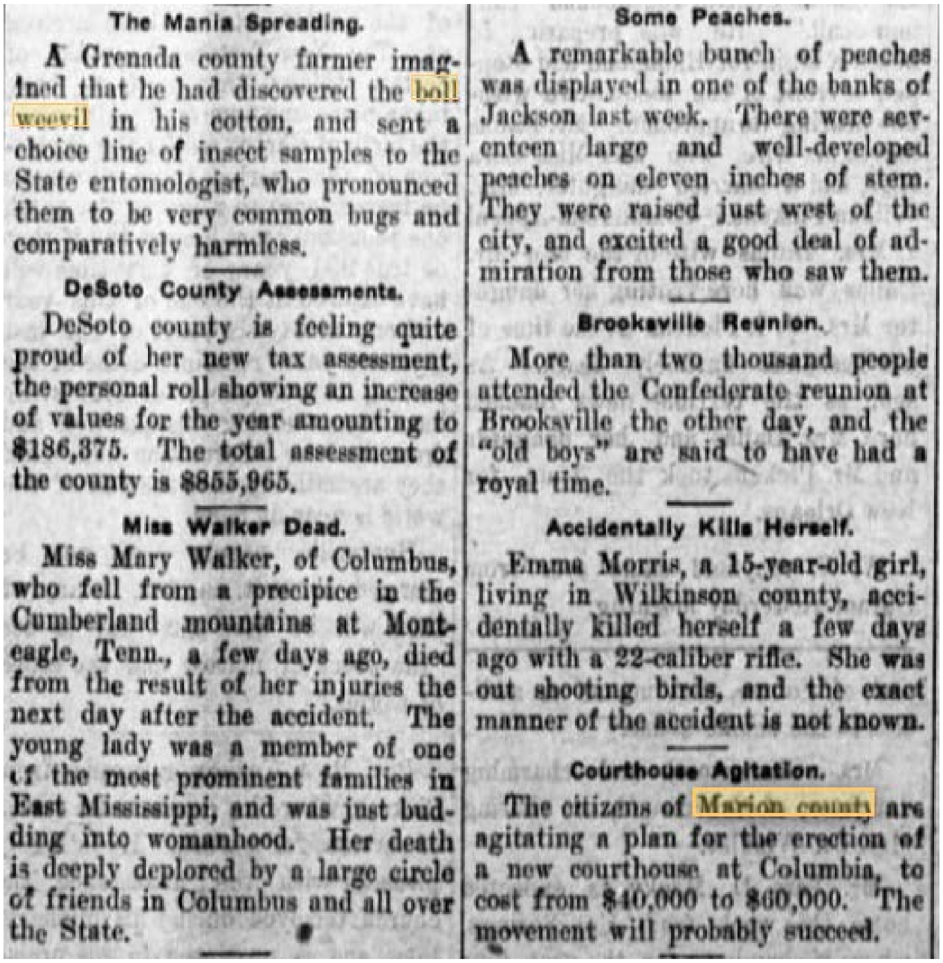
Dr. Hathorn is giving out the methods of fighting the weevil, states that the Texans plant at the time that the best average result can be obtained before the weevil makes its appearance.

Next, the best seed obtainable is used. Rapid shallow cultivation is kept up as late as possible and the acreage is cut down 25 per cent, so that it can be well cultivated. All punctured and fallen squares should be gathered up and burned. This is of utmost importance the first year.

Dr. Hathorn also urges the early destruction of the stalks in the fall and the burning of thickets and turnrows, or any likely haunt of the insect.

Note: Newspapers reported the number of boll weevil cases not only in their own county, but also from other counties or even different states. Figure A.2 shows that the boll weevil infestation in Marion County was reported in Jasper County (left) and Attala County (right). Sources: Jackson Daily News on June 11th, 1910 (left); The Star Ledger on July 8th, 1910 (right).

Figure A.3: Errors in Newspapers



The Mania Spreading.

A Grenada county farmer imagined that he had discovered the boll weevil in his cotton, and sent a choice line of insect samples to the State entomologist, who pronounced them to be very common bugs and comparatively harmless.

DeSoto County Assessments.

DeSoto county is feeling quite proud of her new tax assessment, the personal roll showing an increase of values for the year amounting to $186,375. The total assessment of the county is $855,965.

Miss Walker Dead.

Miss Mary Walker, of Columbus, who fell from a precipice in the Cumberland mountains at Monteagle, Tenn., a few days ago, died from the result of her injuries the next day after the accident. The young lady was a member of one of the most prominent families in East Mississippi, and was just budding into womanhood. Her death is deeply deplored by a large circle of friends in Columbus and all over the State.

Some Peaches.

A remarkable bunch of peaches was displayed in one of the banks of Jackson last week. There were seventeen large and well-developed peaches on eleven inches of stem. They were raised just west of the city, and excited a good deal of admiration from those who saw them.

Brooksville Reunion.

More than two thousand people attended the Confederate reunion at Brooksville the other day, and the "old boys" are said to have had a royal time.

Accidentally Kills Herself.

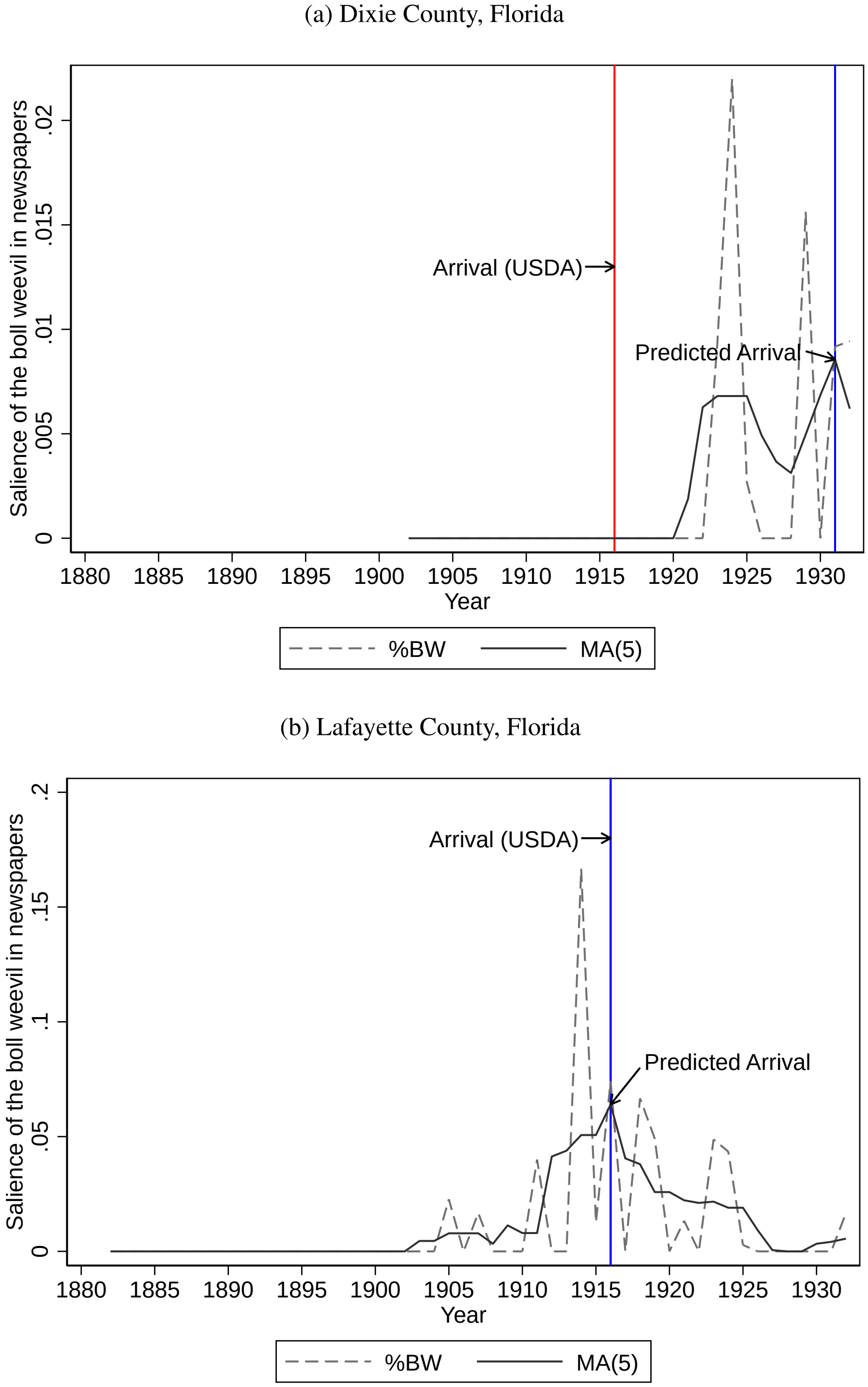
Emma Morris, a 15-year-old girl, living in Wilkinson county, accidentally killed herself a few days ago with a 22-caliber rifle. She was out shooting birds, and the exact manner of the accident is not known.

Courthouse Agitation.

The citizens of Marion county are agitating a plan for the erection of a new courthouse at Columbia, to cost from $40,000 to $60,000. The movement will probably succeed.

Note: An example of possible errors in our approach. The search word "boll weevil" shows in one article and "Marion County" shows in another article in the page 2 of The Lexington Advertiser on July 28th, 1904. This page is still counted when we construct the salience measure of Marion County even though it did not report the boll weevil infestation of the county.

Figure A.4: Salience of the Boll Weevil - Dixie and Lafayette Counties, Florida



(a) Dixie County, Florida

(b) Lafayette County, Florida

Note: The dashed line is the salience measure of Dixie County (in Panel (a)) and Lafayette County (in Panel (b)) over time, respectively. The salience measure is constructed based every available newspaper outlet in Florida between 1882 and 1932. In Panel (a), missing values are shown in early periods because the search word "Dixie County" did not show in newspapers until 1921 (except for errors). Dixie County was created in 1921 from the southern portion of Lafayette County. The solid line is 5-year moving averages of the salience measure (MA(5)) for each county. The red horizontal lines indicate the boll weevil's arrival from the USDA map. The blue horizontal lines show the predicted arrival where MA(5) is the highest.

Measurement Error Derivations

For the continuous case where measurement error in X_1 and X_2 are classical and uncorrelated with each other, the IV estimator is

$$\begin{aligned}\widehat{\beta}_{IV,X1} &= \frac{Cov(y, X_1)}{Cov(y, X_2)} \\ &= \frac{Cov(\alpha + \beta X^* + \epsilon, X^* + u)}{Cov(X^* + u, X^* + e)} \\ &= \beta \frac{Var(X^*)}{Var(X^*) + Cov(u, e)}\end{aligned}$$

which yields the true parameter β if $Cov(u, e) = 0$.

For the binary case, consider the first stage regression

$$X_1 = \pi_0 + \pi_1 X_2 + \eta$$

Meyer and Mittag (2017) show that measurement error in a binary outcome yields a biased right-hand side coefficient. In the absence of measurement error in X_2, the estimated first stage coefficient therefore would be $\widehat{\pi}_1 = \pi_1(1 - \theta)$. However, since also X_2 is mismeasured, the coefficient is additionally attenuated as $\widehat{\pi}_1 = \pi_1(1 - \theta)(1 - \gamma)$. Now consider the reduced form,

$$\begin{aligned}y &= \alpha + \beta X_1 + \epsilon \\ &= \alpha + \beta(\pi_0 + \pi_1 X_2 + \eta) + \epsilon \\ &= \kappa + \psi X_2 + \xi\end{aligned}$$

where $\psi = \beta\pi_1$, $\kappa = \alpha + \beta\pi_0$, and $\xi = \epsilon + \beta\eta$. Also here the measurement error in X_2 is reflected in the attenuation of the reduced form coefficient, $\widehat{\psi} = \psi(1 - \gamma) = \beta\pi_1(1 - \gamma)$. Lastly, the IV coefficient on X_1, using X_2 as instrument, will be the estimated reduced form coefficient divided by the estimated first stage coefficient,

$$\begin{aligned}\widehat{\beta}_{IV,X1} &= \frac{\widehat{\psi}}{\widehat{\pi}} \\ &= \frac{\beta\pi_1(1-\gamma)}{\pi_1(1-\theta)(1-\gamma)} \\ &= \beta\frac{1}{(1-\theta)}\end{aligned}$$

resulting in an inflated estimate of the true parameter unless $\theta = 1$, in which case the IV estimator is undefined.