Detecting Informal Data Use in Literature

Sara Lafia[1], Elizabeth Moss[1], Andrea Thomer[2], and Libby Hemphill[1,2] (University of Michigan: ICPSR[1]; UMSI[2])

Challenge

Formal data citations using unique identifiers are readily discoverable; however, informal references indicating research data reuse are challenging to detect. *How can computational approaches to detect data use complement human efforts?*

Procedure

- Search for formal (unique identifiers) and informal mentions (study names, aliases) of research data
- Extract terms ("survey", "sample"…) that often accompany data citations and sections of articles (Methods…) where found
- Predict custom entity type (*Data*) at the sentence level
- Evaluate candidate documents for inclusion in the ICPSR Bibliography of Data-Related Publications

Detecting citations: human vs. computational approach

ICPSR Studies
- ***Narrow***: search by project archive
- ***Precise***: Iterate on high quality, custom queries

Retrieve, skim documents
- ***Deep*** coverage
- ***Diverse*** kinds of materials

Indicators

Name aliases, PIs, grant numbers, variables, etc.

Add to ICPSR Bibliography of Data-Related Literature

ICPSR Studies
- ***Broad***: search by identifiers, names, and aliases
- ***Generic***: Apply search criteria to all studies

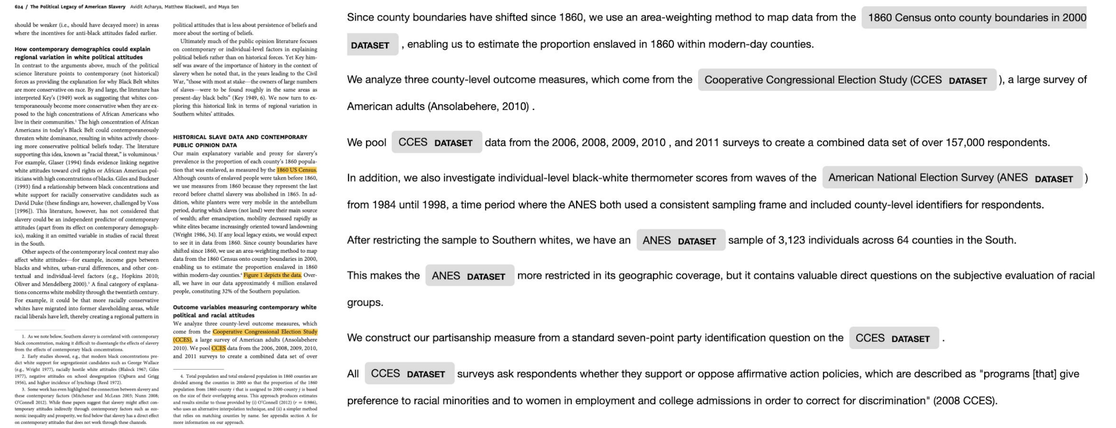
Retrieve, parse documents
- ***Broad*** coverage across search APIs (Dimensions, CrossRef, GROBID)
- ***Narrow*** kinds (born digital, full text)

Indicators

If sentence has keyword(s), is in a paper section, or contains an acronym, keep; else, discard reference

Check sentence for Data entity

Compare entities to ICPSR Study Names, Aliases



634 / The Political Legacy of American Slavery · Avidit Acharya, Matthew Blackwell, and Maya Sen

should be weaker (i.e., should have decayed more) in areas where the incentives for anti-black attitudes faded earlier.

How contemporary demographics could explain regional variation in white political attitudes

In contrast to the arguments above, much of the political science literature points to contemporary (not historical) forces as providing the explanation for why Black Belt whites are more conservative on race. By and large, the literature has interpreted Key's (1949) work as suggesting that whites contemporaneously become more conservative when they are exposed to the high concentrations of African Americans who live in their communities.[1] The high concentration of African Americans in today's Black Belt could contemporaneously threaten white dominance, resulting in whites actively choosing more conservative political beliefs today. The literature supporting this idea, known as "racial threat," is voluminous.[2] For example, Glaser (1994) finds evidence linking negative white attitudes toward civil rights or African American politicians with high concentrations of blacks. Giles and Buckner (1993) find a relationship between black concentrations and white support for racially conservative candidates such as David Duke (these findings are, however, challenged by Voss [1996]). This literature, however, has not considered that slavery could be an independent predictor of contemporary attitudes (apart from its effect on contemporary demographics), making it an omitted variable in studies of racial threat in the South.

Other aspects of the contemporary local context may also affect white attitudes—for example, income gaps between blacks and whites, urban-rural differences, and other contextual and individual-level factors (e.g., Hopkins 2010; Oliver and Mendelberg 2000).[3] A final category of explanations concerns white mobility through the twentieth century. For example, it could be that more racially conservative whites have migrated into former slaveholding areas, while racial liberals have left, thereby creating a regional pattern in political attitudes that is less about persistence of beliefs and more about the sorting of beliefs.

Ultimately much of the public opinion literature focuses on contemporary or individual-level factors in explaining political beliefs rather than on historical forces. Yet Key himself was aware of the importance of history in the context of slavery when he noted that, in the years leading to the Civil War, "those with most at stake—the owners of large numbers of slaves—were to be found roughly in the same areas as present-day black belts" (Key 1949, 6). We now turn to exploring this historical link in terms of regional variation in Southern whites' attitudes.

HISTORICAL SLAVE DATA AND CONTEMPORARY PUBLIC OPINION DATA

Our main explanatory variable and proxy for slavery's prevalence is the proportion of each county's 1860 population that was enslaved, as measured by the 1860 US Census. Although counts of enslaved people were taken before 1860, we use measures from 1860 because they represent the last record before chattel slavery was abolished in 1865. In addition, white planters were very mobile in the antebellum period, during which slaves (not land) were their main source of wealth; after emancipation, mobility decreased rapidly as white elites became increasingly oriented toward landowning (Wright 1986, 34). If any local legacy exists, we would expect to see it in data from 1860. Since county boundaries have shifted since 1860, we use an area-weighting method to map data from the 1860 Census onto county boundaries in 2000, enabling us to estimate the proportion enslaved in 1860 within modern-day counties. Figure 1 depicts the data. Overall, we have in our data approximately 4 million enslaved people, constituting 32% of the Southern population.

Outcome variables measuring contemporary white political and racial attitudes

We analyze three county-level outcome measures, which come from the Cooperative Congressional Election Study (CCES), a large survey of American adults (Ansolabehere 2010). We pool CCES data from the 2006, 2008, 2009, 2010, and 2011 surveys to create a combined data set of over

1. As we note below, Southern slavery is correlated with contemporary black concentration, making it difficult to disentangle the effects of slavery from the effects of contemporary black concentrations.

2. Early studies showed, e.g., that modern black concentrations predict white support for segregationist candidates such as George Wallace (e.g., Wright 1977), racially hostile white attitudes (Blalock 1967; Giles 1977), negative attitudes on school desegregation (Ogburn and Grigg 1984), and higher incidence of lynchings (Reed 1972).

3. Some work has even highlighted the connection between slavery and these contemporary factors (Schuman and McLean 2003; Nunn 2008; O'Connell 2012). While these papers suggest that slavery might affect contemporary attitudes indirectly through contemporary factors such as economic inequality and prosperity, we find below that slavery has a direct effect on contemporary attitudes that does not work through these channels.

4. Total population and total enslaved population in 1860 counties are divided among the counties in 2000 so that the proportion of the 1860 population from 1860 county i that is assigned to 2000 county j is based on the size of their overlapping areas. This approach produces estimates and results similar to those provided by (1) O'Connell (2012) ($r = 0.986$), who uses an alternative interpolation technique, and (2) a simpler method that relies on matching counties by name. See appendix section A for more information on our approach.

Since county boundaries have shifted since 1860, we use an area-weighting method to map data from the 1860 Census onto county boundaries in 2000 DATASET , enabling us to estimate the proportion enslaved in 1860 within modern-day counties.

We analyze three county-level outcome measures, which come from the Cooperative Congressional Election Study (CCES DATASET), a large survey of American adults (Ansolabehere, 2010) .

We pool CCES DATASET data from the 2006, 2008, 2009, 2010 , and 2011 surveys to create a combined data set of over 157,000 respondents.

In addition, we also investigate individual-level black-white thermometer scores from waves of the American National Election Survey (ANES DATASET) from 1984 until 1998, a time period where the ANES both used a consistent sampling frame and included county-level identifiers for respondents.

After restricting the sample to Southern whites, we have an ANES DATASET sample of 3,123 individuals across 64 counties in the South.

This makes the ANES DATASET more restricted in its geographic coverage, but it contains valuable direct questions on the subjective evaluation of racial groups.

We construct our partisanship measure from a standard seven-point party identification question on the CCES DATASET .

All CCES DATASET surveys ask respondents whether they support or oppose affirmative action policies, which are described as "programs [that] give preference to racial minorities and to women in employment and college admissions in order to correct for discrimination" (2008 CCES).

Figure 1. Comparison of human heuristics (left) and a computational approach (right) using a custom Named Entity Recognition model trained to predict passages of text indicating research data applied to the article, **The political legacy of American slavery** (Acharya et al., 2016). doi: 10.1086/686631

Training a computational model

1. Label *Data* entities in 2,056 sentences from 400 papers
2. Analyze distribution of indicator terms by section
3. Train custom spaCy NER pipeline with annotated sentences

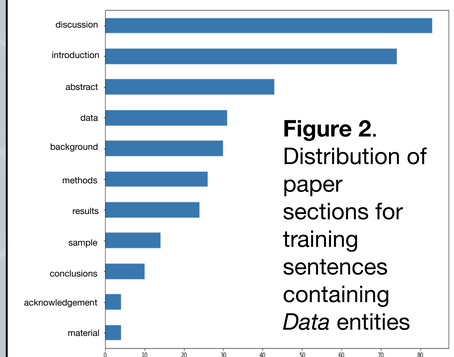


Figure 2. Distribution of paper sections for training sentences containing *Data* entities

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant 1930645.

ICPSR a partner in social science research

SCHOOL OF INFORMATION UNIVERSITY OF MICHIGAN