CMF Data Documentation

November 2025

Contents

I Census of Manufactures Data

These data reflect all known surviving establishment-level manuscripts. We expand upon previous samples of the Census of Manufactures microdata done by Atack, Bateman and Weiss (1980). The data from this complete set of surviving images are extensive, but do not cover particular counties and industries in some decades, and so aggregating these data by geography or industry could give a wildly misleading impression of total manufacturing activity by place and industry. We discuss the data coverage in detail, on this website and in the documentation that describes the data cleaning process in detail. Our general approach to data collection and data cleaning has been to digitize all establishments, and reflect the underlying manuscripts, but there will be errors in reflecting the economics of individual establishments both due to digitization errors and original collection errors by the Census enumerators. We have not imposed particular economic structure on the data, such as what might seem plausible from the perspective of a production function, but individual users of the data should take care when considering the presence of outliers and other features of the data that might drive spurious results (or imposing further data cleaning assumptions that themselves might drive spurious results).

So, in general, users should beware to understand the intricacies of the data coverage and variables.

II Prepared CMF Data Structure

An observation in the prepared data is one establishment, and is uniquely identified by the variables file_name and firm_number, and schedule in the 1880 special schedules. The following table shows the number of variables and observations in each cleaned dataset:

Dataset File Name	Observations	Variables
1850	124873	329
1860	121168	443
1870	203353	753
1880 General Schedule	170811	102
1880 SS1	1981	204
1880 SS2	677	271
1880 SS3	2161	85
1880 SS4	3083	121
1880 SS5	24240	120
1880 SS6	4856	82
1880 SS7	23251	136
1880 SS8	3508	96
1880 SS9	4632	104
1880 SS10	128	70
1880 SS11	2	82
1880 SS12	1381	97

II.A Variable Availability

II.B Manuscript Variables

The following table lists out variables original to the manuscripts, and their availability by year. Checks indicate that the variables are available, stars indicate that specified variables can be constructed from available data, and dashes mark variables that are unknowable. Since there are many manuscript variables unique to 1880, those are omitted in this table.

Variable	**1850**	**1860**	**1870**	**1880**
Firm Name	✓	✓	✓	✓
Industry	✓	✓	✓	✓
Capital	✓	✓	✓	✓
Materials Quantity	✓	✓	✓	-
Materials Kind	✓	✓	✓	-
Materials Unit of Measure	✓	✓	✓	-
Materials Value	✓	✓	✓	⋆
Number of Male Hands	✓	✓	✓	⋆
Number of Female Hands	✓	✓	✓	⋆
Number of Children Hands	-	-	✓	⋆
Mean Male Wage	✓	✓	-	-
Mean Female Wage	✓	✓	-	-
Total Wages	⋆	⋆	✓	✓
Months Active	-	-	✓	⋆
Production Quantity	✓	✓	✓	-
Production Kinds	✓	✓	✓	-
Production Values	✓	✓	✓	⋆
Production Unit of Measure	✓	✓	✓	-
Power Kind	✓	✓	✓	⋆
Machine Description	⋆	⋆	✓	⋆
Number of Machines	-	-	✓	⋆
Horsepower Measure	-	-	✓	⋆
State	✓	✓	✓	✓
County	✓	✓	✓	✓
Township	✓	✓	✓	✓
Closest Post Office	✓	✓	✓	✓

II.B.1 Constructed Variables

The following table lists out the constructed variables available in the prepared data.

Variable	1850	1860	1870	1880
Broadest Industry Category	✓	✓	✓	✓
Cleaned Industry	✓	✓	✓	✓
Cleaned Material Kind	✓	✓	✓	-
Cleaned Number of Children	-	-	✓	⋆
Cleaned Product Kind	✓	✓	✓	⋆
Cleaned Total Wages	-	-	✓	⋆
Detailed Industry Category	✓	✓	✓	✓
Establishment Number	✓	✓	✓	✓
FIPS Code	✓	✓	✓	✓
Granular Industry Category	✓	✓	✓	✓
Image File Name	✓	✓	✓	✓
Is Maker	✓	✓	✓	✓
Is Manufacturer	✓	✓	✓	⋆
Is Shop	✓	✓	✓	⋆
Is a Factory	✓	✓	✓	✓
Leontief Industry Category	✓	✓	✓	✓
Machine Category	✓	✓	✓	⋆
Machine Kind	✓	✓	✓	⋆
Machine Unit	✓	✓	✓	⋆
Material Kind Attribute	✓	✓	✓	-
Material Units	✓	✓	✓	-
Material Unsure	✓	✓	✓	-
Material is Miscellaneous	✓	✓	✓	-
Material is Note	✓	✓	✓	-
Material is Service	✓	✓	✓	-
Product Kind Attribute	✓	✓	✓	-
Product is Miscellaneous	✓	✓	✓	⋆
Product is Note	✓	✓	✓	-
Product is Service	✓	✓	✓	⋆
Product is Units	✓	✓	✓	-
Product is Unsure	✓	✓	✓	⋆
Uses Hand Power	✓	✓	✓	⋆
Uses Horse Power	✓	✓	✓	⋆
Uses Steam Power	✓	✓	✓	⋆
Uses Water Power	✓	✓	✓	⋆
Uses Wind Power	✓	✓	✓	⋆

II.B.2 Unique 1880 General Schedule variables

This table shows the variables unique to 1880, and if they are constructable or have analogous counterparts in other years. All of these variables unique to 1880 are directly from the manuscript; they have not been constructed by us.

Variable	**1850**	**1860**	**1870**	**1880**
Total Materials Value	⋆	⋆	⋆	✓
Adult Male Workers	⋆	⋆	⋆	✓
Adult Female Workers	⋆	⋆	⋆	✓
Children Workers	-	-	⋆	✓
Maximum Workers	⋆	⋆	⋆	✓
Skilled Daily Wage	-	-	-	✓
Unskilled Daily Wage	-	-	-	✓
Hours May-November	-	-	-	✓
Hours November-May	-	-	-	✓
Months 1/2 Time	-	-	-	✓
Months 2/3 Time	-	-	-	✓
Months 3/4 Time	-	-	-	✓
Months Full-Time	-	-	⋆	✓
Months Idle	-	-	⋆	✓
Total Production Values	⋆	⋆	⋆	✓
Number of Steam Boilers	-	-	-	✓
Number of Steam Engines	-	-	-	✓
Horsepower from Steam	-	-	⋆	✓
Height of River Fall	-	-	-	✓
Which River Used	-	-	-	✓
Number of Water Wheels	-	-	-	✓
Breadth of Water Wheels	-	-	-	✓
Horsepower of Water Wheels	-	-	-	✓
Kind of Water Wheel	-	-	-	✓
RPM of Water Wheels	-	-	-	✓

III Coverage

The Census of Manufacturers data is rich and detailed, but it does not cover all manufacturing establishments in the United States. There are two main reasons for this:

1. 500 dollar minimum threshold: Only establishments producing at least $500 worth of goods were included, leaving out very small manufacturers including potential part-time manufacturers and things made at home. These smaller businesses were often included in the Agriculture Census as home manufacturing, but they are not included in these data.

2. Missing manuscripts: Some original manuscript records have been lost over time, particularly when the records were returned from Washington DC to the individual states. Additionally, for 1880, "special agent schedules" covering some industries were all lost.

As a result, aggregating these establishment-level records will differ from the tabulated county-level and county-industry level data. These data also differ due to tabulation errors and potential differences in processing and allocation to industries.

III.A Comparison to County and County-Industry Data

At the time each census was completed, all manuscripts were gathered by the Census Bureau and tabulated at both the county level and the county-industry level. Comparing the data to the tabulations provides insight into the broad coverage of the data:

State	1850	1860	1870	1880	State	1850	1860	1870	1880
AL	✓	✓	✓	✓	MO	✓	✓	✓	✓
AZ	-	-	0%	0%	MT	-	-	✓	✓
AR	✓	✓	✓	✓	NE	-	✓	✓	✓
CA	✓	✓	✓	✓	NV	-	-	✓	✓
CO	-	-	✓	✓	NH	✓	✓	✓	✓
CT	✓	✓	✓	✓	NJ	✓	✓	✓	✓
DE	✓	✓	✓	✓	NM	-	0%	0%	✓
DC	✓	✓	✓	✓	NY	✓	✓	82%	99%
FL	✓	✓	✓	✓	NC	✓	84%	✓	✓
GA	0%	0%	0%	✓	ND & SD	-	-	0%	18%
ID	-	-	✓	✓	OH	✓	26%	74%	68%
IL	✓	✓	46%	✓	OR	✓	✓	✓	✓
IN	✓	✓	✓	✓	PA	✓	✓	✓	✓
IA	✓	✓	✓	✓	RI	✓	✓	✓	✓
KS	-	✓	✓	✓	SC	✓	✓	✓	✓
KY	✓	✓	✓	✓	TN	✓	30%	35%	✓
LA	0%	0%	0%	✓	TX	✓	✓	85%	✓
ME	✓	✓	✓	✓	UT	-	✓	✓	✓
MD	✓	✓	0%	✓	VT	✓	✓	✓	✓
MA	✓	✓	32%	✓	VA	✓	✓	✓	✓
MI	✓	✓	49%	✓	WA	-	✓	✓	✓
MN	✓	✓	✓	✓	WV	-	-	✓	✓
MS	✓	✓	✓	✓	WI	✓	✓	✓	✓

This table shows our coverage of counties. Percents indicate estimates of the share of establishments that we digitized, given the published county-level tabulations. In 1850, the Census records for three counties in California (Contra Costa, San Francisco, and Santa Clara) were lost and never tabulated, and we have complete coverage of the remaining counties in California. Dashes indicate that no survey was conducted, check marks indicate that we have complete coverage.

IV History and Background

IV.A Collecting the Census

The United States began collecting economic data in the third decennial census in 1810, but the process was incomplete and inconsistent through the 1840 census (Bohme, 1987). It was well recognized that the census of 1840 and its predecessors lacked in quality and detail; starting in the 1850 Census, the Census Board was established and consulted statisticians, business leaders, and academics to create better forms for census canvassers to gather census data, including on manufacturing enterprises (Bohme, 1987; Fishbein, 1973). These reports, or *schedules*, in the 1850 census were the following:

1. Free Inhabitants
2. Slave Inhabitants
3. Productions in Agriculture
4. Products of Industry
5. Social Statistics
6. Persons Who Died

Schedule 4 from 1850 provides the first consistent, comprehensive manufacturing microdata from the Census Bureau. Each assistant marshal, the person that actually recorded information, was to go to each business that had a revenue of more than or equal to $500 and ask questions to fill out Schedule 4 (See figure 1). As noted, manufacturing that did not meet this threshold was considered "home manufacturing", and was recorded in the agricultural census. Checking some of these establishments on the agricultural census manuscripts shows that revenues from home manufactures tended to be much less than $500 (Hornbeck and Rotemberg, 2024)

Assistant marshals were paid 15 cents for each establishment canvassed (U.S. Census Office, 1850). Marshals classified an enterprise as a single establishment even when it operated multiple locations within the same Census subdivision, provided that the activities at all sites belonged to the "same concern, and all engaged in the same manufacture" (U.S. Census Office, 1860; Hornbeck et al., 2025). However, some Census entries corresponded to a single owner operating across multiple industries—for example, E. E. Locke & Co., which ran both a distillery and a mill. We split these enterprises into different establishments. (See **??** Innovations that likely assisted in making this effort more effective than previous iterations included providing explanations of the inquiries on the manuscript themselves, and an instruction to promise confidentiality (U.S. Census Office, 1850; Fishbein, 1973)

The 1860 decennial census was similar to the previous decade, slightly changing the schedule order and names, and the Products of Industry were recorded in schedule 5 (U.S. Census Office, 1860). The information recorded on the manuscripts stayed almost exactly the same, except that the closest post office was recorded in addition to the other geographic information (See figure 1). To assist in the task, marshals and assistant marshals were given more detailed instructions and examples than in 1850 (see U.S. Census Office 1860).

As a result of the Civil War and the Fifteenth Amendment to the Constitution of the United States, the "Slave Inhabitants" schedule was removed from the Census, resulting in the schedule numbers being adjusted once again. "Products of Industry" was again Schedule 4 (U.S. Census Office, 1870). Although the overall process of the census did not change, 1870 saw some significant changes to the questions asked by marshals (see figure 1). This includes adding questions about machines used, time in operation, and child workers.

In 1880, manufacturing data was recorded in schedule 3 (U.S. Census Office, 1880). This was the first year that schedule was named "Manufactures." The largest change to the survey itself was that the schedule was split into three different parts:

1. Manufactures (which we call the "general schedule")
2. Special Schedules of Manufactures (which we call the "special schedules")
3. Special Agent Schedules of Manufactures (which we call the "special agent schedules")

The special schedules and special agent schedules were designed to record information for the most important industries. The special schedules are as follows:

1. Agricultural Implements
2. Paper Mills
3. Boots and Shoes
4. Leather (Tanned and Curried)
5. Lumber Mills and Saw-Mills
6. Brick Yards and Tile Works
7. Flour and Grist Mills
8. Cheese, Butter, and Condensed Milk Factories
9. Slaughtering and Meat-Packing
10. Salt Works
11. Small Coal Mines
12. Quarries

The general schedule was similar to the previous "Products of Industry" schedules, although it is more different than the other three are to each other. The 1880 Census only asks about the total value of production and materials (without additionally asking about the names and quantities) and it includes more questions about power used and time in operation. Industries that were not recorded on the special schedules or the special agent schedules were to be recorded in the general schedule.

Special schedules were filled out by the regular assistant marshals when they came upon a business in one of the above industries. These special schedules included sector-specific questions, and each special schedule was different.

Special Agent Schedules were filled out by marshals who had specific expertise in the industry of the manufacture they were interviewing, and they used an extensive list of questions. Industries that are covered by the special agent schedules are: cotton, wool, and worsted goods; silk and silk goods; iron and steel; the coke industry; the glass industry; the mining of metals, coal, and petroleum; distilleries and breweries; shipbuilding; and fisheries (Delle Donne, 1973; Carroll D. Wright, 1900). Although most special and special agent schedule businesses were recorded on these schedules, some establishments in these industries ended up being covered in the general schedule.

IV.B Aggregation and Publication

After the completion of each of the 1850-1880 censuses, the manuscripts from the Census of Manufactures were sent to Washington D.C. for tabulation and publication. Clerks employed by the Census Bureau went through and tabulated the data. As they went, they marked off establishments, which can be seen on the manuscripts (see 2). Some clerks did "X" marks, some circled the corner, and some wrote check marks. If a canvassed establishment did not actually meet the $500 rule, it was not to be counted, which can be seen by the lack of marks on its line (Fishbein, 1973). These data were tabulated and published at the county-level for every year. In 1860-1880, establishments were also assigned industry groups, and county-by-industry tables were made and published in addition to the county tables. These tables have all since been fully digitized (see Haines (2010) and (Hornbeck and Rotemberg, 2024)). Once the work of aggregation was completed, the manuscripts were then sent away back to the states and were scattered in many different locations, including state archives and in the hands of museums (Don, 1973). In the process of returning the manuscripts, and over the years, some have been lost or destroyed. For example, some manuscripts were used as wrapping paper for other manuscripts on their way back to the states and were thus ruined (Atack and Bateman, 1999).

V Collection

Manuscripts and microfilms of the manuscripts were located in many different places, and needed to be gathered and digitized. The following process was used to gather the images prior to transcription.

- **Locate the records**
  - Manuscripts were found in state archives, the National Archives, university libraries, Ancestry.com, and Jeremy Atack's basement.
- **Acquire images or microfilm**
  - Where digital scans already existed, copies were obtained.
  - Where only microfilm was available, reels were borrowed or purchased.
- **Digitize microfilm**

  - The University of Chicago Library scanned the microfilm into high-resolution images.

- **Standardize the images**
  - Most scans captured two manuscript pages per image; many were blank or contained only enumerators' notes.
  - Every image was reviewed to identify pages with usable data.

These steps produced a unified image archive that serves as the foundation for the data extraction and preparation steps.

VI Image Processing and Metadata Collection

VI.A Processing

As mentioned in the collection steps, images of the manuscripts were generally formatted as two pages per image, many of which were blank or contained only notes from the census takers. In some cases, all schedules of the census were received (population, social statistics, agriculture), and we needed to sort through and identify the images relating to the census of manufacturers. Using code, we split these images up into single page units.

VI.B Metadata Collection

Metadata refers to information about each image: what state and year it comes from, what county is listed, whether the image is legible, whether it was mis-cut or duplicated, and so on. Collecting and cleaning this metadata helped to make the images and their data more usable.

1. **Download the Metadata File**
   Each group of images had a corresponding metadata file (a spreadsheet). Research assistants (RAs) began by downloading the relevant file to work on.

2. **Check for Image Issues**
   A key task was to identify whether any images were badly cut, missing information, or otherwise unusable. If a problem was found, the image was either fixed or flagged for recollection.

3. **Standardize the Spreadsheet**
   Every metadata file was organized into the same set of columns (state, year, county, legibility, duplicates, notes, etc.). This ensured consistency across thousands of files.

4. **Work Through Notes Carefully**
   Much of the work came from reading the "Notes" column. This is where RAs recorded quirks—like duplicate images, multiple counties on a single page, or schedules that were crossed out. These notes then had to be translated into standardized entries in the spreadsheet.

5. **Mark Special Cases**
   Several columns captured unusual situations:

- *Duplicates and Variants*: If two images showed the same page, one was marked as the original and the other as a duplicate or variant.
- *Nullified or Crossed Out*: Some pages were explicitly voided and had to be marked as such.
- *Notes-Only Pages*: Occasionally a page contained no manufacturing data at all, only a census taker's note. These were flagged separately.

At the end of this process, every census image was paired with a clean and standardized metadata record. This meant we could track coverage, identify gaps, and eventually link images to digitized data in a systematic way. The essential *file_name* variable in the data is created by this step.

VI.C Validation against published totals

Once metadata were collected, we summed the total number of establishments in our images and compared them with Census tabulations, which were collected by Michael Haines and published as ICPSR 2896 (Haines, 2010). Specifically, we take the metadata and check it against Haines's establishment totals. In cases where our estimates differed by large amounts (or missed entire counties), efforts were made to find more images to complete each state-year. Outside of things documented in the Coverage section, we believe we have fairly complete coverage.

VII Data Entry

Once we obtained a "complete" set of images, we then packaged them for the data entry company Digital Divide Data (DDD), where a team in Kenya double-entered the data and reconciled their output before sending us their final version. The team ranged from 20 to 35 individuals and completed all of the data entry in just over two years. They followed a number of instructions. Specifically, they did not enter establishments if they were crossed out or marked as omitted. These cases usually indicated that an establishment had been entered twice or that an establishment did not belong in the census of manufacturers. Additionally, they entered whatever data seemed most recent. For example, there are sometimes values crossed out and new values written. Since we can not identify when these changes were made (whether at the time of entry, during tabulation, or at some other time), we chose to enter only those values that seemed like the last iteration of any changes.

VIII Manual Steps

The raw transcribed data that we received from DDD was very good. A check of a large random sample showed that the entries were generally very accurate to what was written on the manuscripts. There were some mistakes in transcription, however, by DDD associates. Additionally there were some mistakes made by census marshals. The initial steps taken by our team were to fix these problems before the data were transformed into a usable state. The nature of these issues was such that most of this work had to be performed manually by research assistants after being identified and marked in the data.

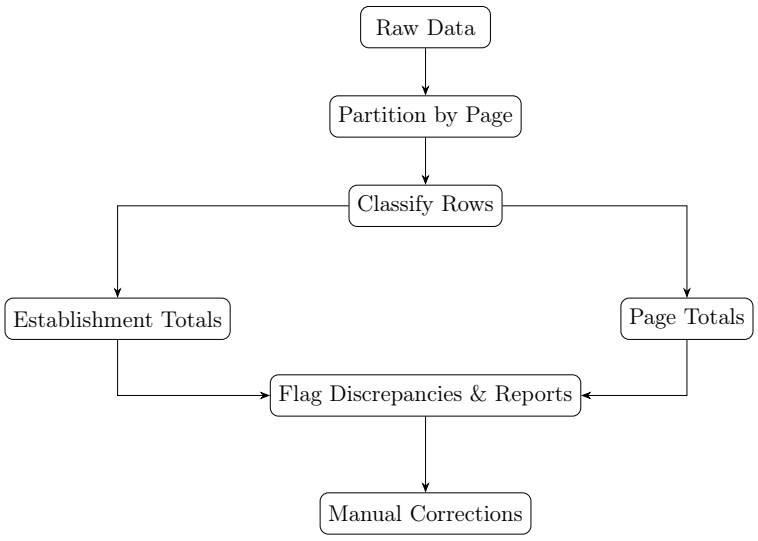
VIII.A Error Checking

Anytime a census taker (or tabulator) entered a total on a page (such as the total production value for the entire page or the summed value of materials value for an individual firm), DDD

entered this data and marked it as a "total." In order to systematically identify possible data entry errors, we calculated a new total based on entered data and compared that value with the total reported in the image. We hired a team of undergraduates at the University of Chicago to examine the entered data and their corresponding images where these totals were mismatched, to look for data entry errors that could be corrected. Commonly the total was merely calculated incorrectly by the census taker or tabulator, or the total reflected a sum of values that were later crossed out and replaced with other values. In these cases, we made no changes. In cases where we identified a data entry error, we corrected it.

VIII.A.1 Process

We wrote code that automatically identified discrepancies between recorded totals and the sums that should have resulted from the underlying data, significantly reducing the manual effort required to clean these historical datasets.



The validation process operated on two levels. Establishment totals represented the sum of values for a single manufacturing firm across multiple product lines or categories. Page totals aggregated all establishment data recorded on individual manuscript pages. Both types of totals were prone to arithmetic errors that compromised data quality and required systematic correction.

This system accommodated the evolving structure of the Census of Manufactures across the 1850-1880 period. As mentioned, the 1850 and 1860 censuses used similar formats with nine core variables covering capital, materials, labor, and production. The 1870 census introduced additional complexity with seven validated variables and new data categories. The 1880 census represented a major expansion with specialized industry schedules containing dozens of variables each. Separate variable definitions and validation logic were maintained for each specialized year and schedule. This approach ensured that validation procedures remained consistent while accommodating the substantial structural differences across industry schedules.

VIII.A.2 File Structure and Partitioning

The digitized data arrived as Excel files containing information from hundreds of manuscript pages. Each file required partitioning into discrete units that corresponded to individual census pages, since totals were calculated within these page boundaries. Unique page identifiers were tracked to determine where one page ended and the next began, creating an index of row ranges for each manuscript page within the larger dataset.

This partitioning step was essential because validation had to occur within the original organizational structure used by census takers. Without proper page boundaries, legitimate totals could not be distinguished from computational errors. The partitioning process handled various data quality issues, including missing page identifiers and inconsistent formatting that commonly appeared in digitized historical materials.

VIII.A.3 Error Classification and Detection Framework

Each data row was categorized according to its role in the census structure. Individual establishment records contained the raw data for manufacturing firms. Establishment total rows contained computed sums for specific firms. Page total rows contained aggregate values for entire manuscript pages. This classification system, stored in a "totals" column, drove the validation logic but itself contained systematic errors that required correction.

Analysis revealed that many establishment totals had been incorrectly labeled as page totals during digitization. This was addressed by examining pages with multiple entries marked as page totals and systematically reclassifying all but the last entry as establishment totals. The remaining entry underwent additional verification to determine its proper classification. This automated correction process handled thousands of mislabeled records that would otherwise have required manual review.

VIII.A.4 Validation Algorithms and Establishment Total Verification

The primary validation method computed expected totals from individual line items and compared these against recorded totals. For most establishments, this straightforward summation approach identified computational errors effectively. However, single-establishment totals also had to be handled, where the recorded total represented internal aggregations not fully detailed in the line items.

The verification process employed a two-stage approach. First, totals were validated using standard summation across all relevant line items. If this failed, whether the discrepancy could be explained by single-establishment totaling was checked, where the total represented aggregated departmental data within a single firm. This secondary check examined firm identification numbers and data patterns to distinguish between computational errors and legitimate single-firm aggregations.

Various data quality challenges were handled, including: mixed data types, missing values, and inconsistent formatting. Automatic type conversion processed numeric data where possible, while robust error handling managed cases where conversion failed. The verification proceeded across all numeric variables simultaneously, enabling comprehensive validation while maintaining detailed logs of results for each variable.

VIII.A.5 Page Total Validation

Page total verification required more complex logic because these values represented sums across multiple establishments while excluding establishment totals to avoid double-counting. All establishment totals on each page were identified, subtracted from the page sum, and the result was compared against the recorded page total.

This process had to account for the hierarchical structure of census data, where page totals aggregated raw establishment data but excluded the computed establishment totals themselves. The validation algorithm tracked which values contributed to page totals and which represented intermediate computations that should be excluded from page-level summations.

VIII.A.6 Error Reporting and Output Generation

Comprehensive error reports were generated that specified the exact location and nature of each computational discrepancy. Error logs included row numbers, variable names, and discrepancy details formatted for efficient manual review. Technical column identifiers were converted into user-friendly variable names, enabling research assistants to quickly locate and address remaining errors.

Error output distinguished between establishment total errors and page total errors, allowing targeted correction efforts. Performance metrics were also tracked including total cells validated, errors detected and corrected automatically, and the reduction in manual labor achieved through automation. These statistics provided quantitative assessment of the validation process's effectiveness.

VIII.A.7 Manual Correction

Once potential errors in the transcribed data had been identified and categorized, research assistants were assigned batches of errors to check. They then checked each of these potential errors in the transcribed data against the manuscripts, making corrections where necessary. These corrections were then compiled and added to the transcribed data.

The validation process was designed to integrate with the broader data cleaning workflow. Output files followed standardized naming conventions that facilitated integration with downstream processing steps. Error reports used consistent formatting that enabled efficient import into spreadsheet applications for manual review and correction.

Entire state-year datasets were processed systematically, handling the geographic and temporal scope of the complete Census of Manufactures collection. Batch processing capabilities enabled efficient validation of large datasets while maintaining detailed logging and error tracking across all processed files. This systematic approach ensured consistent data quality standards across the dataset while providing documentation on the validation activities.

VIII.B Additional Manual Checks and Fixes

After working with the error-checked data, there were a number of common issues that were noticed. Processes were set up to systematically retrieve affected data, which were then checked by research assistants.

VIII.B.1 Firm Line and Firm Number Cleaning

The variables *firm_number*, which corresponds to the establishment identifier within a census manuscript, and *firm_line*, an identifier for the line of data within an establishment, were supposed to be numeric identifiers assigned by DDD as they transcribed the data. These were sometimes mistakenly not assigned, and occasionally contained non-numeric characters.

Many of these issues were discovered while going through the data in other pursuits, so when something was found it was flagged and sent to RAs to check and correct.

This issue was most common to the 1880 special schedules, so simple code was used to detect places to check. This was as simple as bringing up all places that *firm_number* and *firm_line* were either non-numeric, missing, or contained ".". These instances were sent to RAs to fix.

VIII.B.2 Establishment Fixing

There are variables for which each establishment should only have one entry, such as capital value. Sometimes this was not the case with the raw, transcribed data. Simple code was used to mark these places. For 1850-1870, an establishment was tagged if at least one of capital, business name, or wages appeared in an establishment line other than row 1.

A research assistant then went through each of these tagged establishments and checked all of the data entered against the manuscript image. In places where there were mistakes, an RA noted corrections, which were then compiled and applied. If there was not a transcription error, then it was determined that the recorded establishment was, in fact, two establishments, and the data were split to reflect that.

VIII.B.3 Duplicate Establishment Id Checking

In some cases, there were duplicate establishments in the transcribed data. After the establishment-fixing and firm line cleaning steps, each *file_name*, *firm_number*, *firm_line* combination should have been unique. Cases where it was not were marked and checked. If an establishment was accidentally entered twice by DDD, then the duplicate establishment was deleted from the data.

Duplicate establishments were also created in a few more ways. Sometimes, establishments were canvassed twice by marshals, ending up on two separate physical manuscripts. Other times multiple images were made of a single manuscript, resulting in duplicate images and thus duplicate data. In rare cases we also accidentally sent the same image to DDD twice. Once discovered, these duplicates were removed from the data.

IX Data Cleaning

Data cleaning began by extracting the materials, product name, and production kind columns from the raw data and creating "clean" versions of each. To clarify the terminology:

- **Product names** are general labels for establishments (e.g., "cotton mill").
- **Production kinds** are the specific goods produced or sold (e.g., "cotton yarn").
- **Materials** are the inputs used in production (e.g., "raw cotton").

These categories occasionally overlap. The raw *product name*—recorded in the second column of the manuscripts in all years except most special schedules—later becomes the *raw industry* variable, marking its first use in the workflow.

IX.A Compiling Raw Strings

The first step was to compile every unique material, product name, and production kind string across 1850–1880. In 1880, the general schedule and special schedules 8 and 12 did not record materials or provide a separate section for production kinds, so only product names were used. For other special schedules, no additional string cleaning was necessary since they contained fixed, standardized lists of products and materials.

In total, there were roughly 53,000 unique product name strings, 21,000 unique production kind strings, and 18,000 unique materials strings.

IX.B Cleaning and Standardizing Strings

For each of the three variables, a crosswalk was created linking each raw string to a cleaned version. Cleaning was performed both algorithmically—using fuzzy matching and heuristic text normalization—and manually by research assistants.

Typical transformations included:

- **Spelling standardization:** e.g., "bd cloths" → "bedclothes."
- **Separating attributes:** If an attribute was embedded in the raw string, it was split into a new column. For instance, "black ink" was divided into the product "ink" and the attribute "black."
- **Splitting multiple items:** When multiple products or materials appeared in one entry, they were separated into distinct records. For example, "boiler & flue iron" yielded two products: "boilers" and "flue iron."

Product name cleaning proved the most complex. Each product string was independently cleaned by two research assistants and reconciled by a third reviewer. When a raw string already matched a standardized form, it was simply copied over without modification.

IX.C Non-standard Item Flagging

Finally, each unique material and production kind string was reviewed, and certain entries were flagged when they did not represent a straightforward item or if they were not actually a product. In the final data these categories appear as binary variables indicating a category.

Flags for production kinds:

1. Service - If the product that is sold is a service (e.g., "gun repairs")
2. Units - If the product is units of something (e.g., "bags of flour")
3. Miscellaneous - If the product listed is by nature various things (e.g., "mill stuff")
4. Unsure - If the product is not easily classified
5. Note - If the text that is in the production kind column is a note from the census taker

Flags for materials:

1. Service - If the "material" that is used is a service (e.g., "ironing")
2. Units - If the material is units of something (e.g., "whiskey bbls")
3. Miscellaneous - If the material listed is by nature various things (e.g., "smaller articles")
4. Unsure - If the material is not easily classified
5. Note - If the text that is in the materials kind column is a note from the census taker

IX.D Types of Producers

We also categorized some establishments into different types of producers: makers, shops, manufactures, and factories. The categorizations were assigned using string matching to the product name string, so a "Candy Maker" is a "maker" (as opposed to a factory). If a product name does contain one of these words, it is not assigned one of the categories.

Once these cleaned-string crosswalks for the 1850-1880 general schedule were made (and special schedule 8 and special schedule 12 since they record non-standardized products), they were applied to the data:

IX.E Crosswalk implementation

For each census year, the raw product, material, and production strings were first standardized by harmonizing cases and removing excess whitespace. These standardized strings were then merged with the corresponding crosswalk files, which contain the mappings from the raw strings to their cleaned counterparts.

Once the merges were complete, multiple cleaned entries were concatenated into a single string, thus removing blank entries. These were then broken up back into the cleaned variables, guaranteeing that establishments with multiple production kinds or materials have the correct ordering of those items.

The final result was a set of cleaned datasets for each census year. In these, all product, material, and production strings are consistently standardized, attributes are recorded in structured form, and the data are optimized for further analysis. Special treatment is given to the 1880 schedules: the general schedule is cleaned in the same way as earlier decades, while most of the special schedules are passed through unchanged, except for those with product strings, which are cleaned in the same manner as above.

X Reshaping

X.A Wide to Long

The data were originally transcribed into *long* format, where each establishment could occupy multiple rows. This happens when an establishment has multiple materials or products. Later cleaning and analysis required reshaping the data to *wide* format so that each establishment occupies a single row.

X.B Implementation

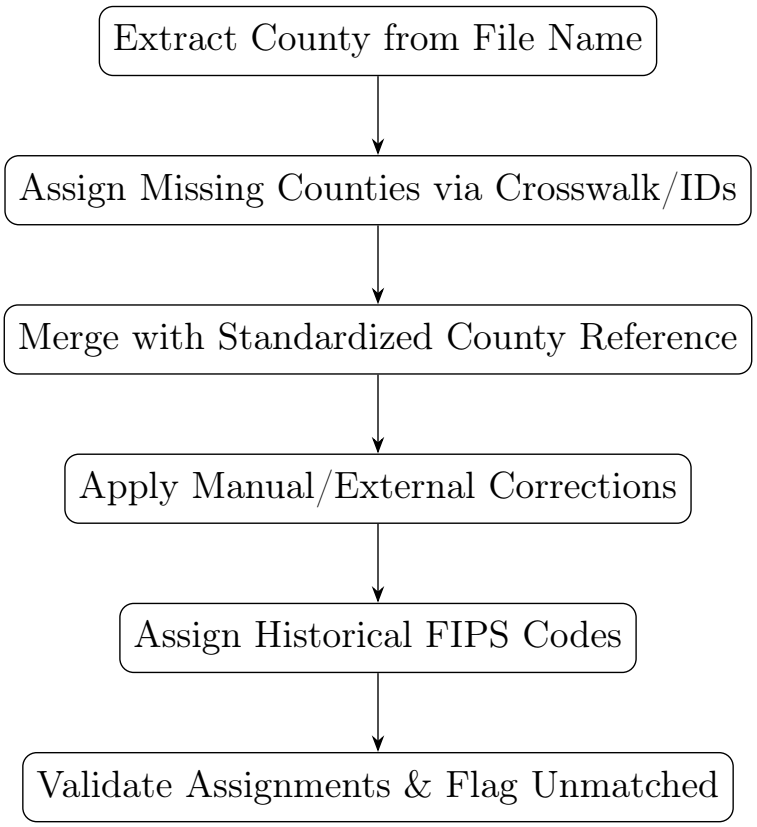
The reshaping process follows these main steps:

1. Remove unnecessary or duplicate rows to retain only unique establishment records.
2. Assign consistent identifiers to establishments with multiple rows.
3. Reshape product-specific and material-specific columns from long to wide format, while retaining single-row information.
4. Remove unused variables, compress storage, and sort for consistency.

This structure consolidates all product, material, and production information for each establishment into a single row, making the dataset easier to analyze. For example, multiple materials are stored in columns such as *materials_kinds1*, *materials_kinds2*, etc.

XI County Assignment

County information was included for each manuscript at the top of the manuscript page. As such, it was not recorded during the transcription of the images by DDD but rather during the metadata check step, where it was usually included in the image file name. We use the file name to connect each establishment to its county. In some occasions the image file name did not include the county, so establishments on those images were assigned counties by hand. Once each file name and thus establishment was assigned to a county, historical fips codes could then be added. These were obtained from Haines via ICPSR (Haines, 2010).



XI.A County Merging

County assignments for establishments were carried out through a combination of automated merges and manual corrections. The general workflow followed several stages.

First, a standardized county reference file was created to ensure consistency in names across different sources. This served as the basis for linking establishment records to counties.

For the census years 1850 through 1870, establishment data were processed in a consistent way. Duplicate records were dropped, establishment identifiers were parsed to generate preliminary county information, and these were merged with the reference file. When county assignments remained missing, information extracted from identifiers was used to fill gaps. Additional external correction files were then incorporated, and the cleaned data with county assignments were saved for each year.

The 1880 data required a more elaborate procedure. Counties were first merged from metadata files, and missing values were supplemented using identifiers. Corrections from manually collected files and external spreadsheets were then integrated, including special adjustments for Nevada counties. These corrected county values replaced preliminary assignments where necessary. Final datasets were saved separately for each schedule type.

In addition to automated procedures, some establishments could not be assigned counties directly. Unmatched cases were flagged, collected manually, and subsequently merged back into the data. Throughout the process, validation checks ensured that establishments were not dropped, that county assignments were present for all records, and that corrections were applied consistently.

XI.B FIPS assignment

For the 1850, 1860, and 1870 census years, county-to-FIPS code assignment followed a consistent methodology. County crosswalk files were prepared by extracting county-level records from ICPSR reference datasets, generating state abbreviations, and standardizing county names to lowercase format. Establishment-level data underwent name standardization procedures to reconcile spelling variations and nomenclature differences between establishment records and reference datasets. County names were adjusted through systematic replacement operations to address common transcription errors, historical name variations, and multi-county designations. Complex cases involving multiple counties listed on single manuscript pages were resolved through firm-specific assignment rules based on establishment names and file identifiers.

The 1880 data processing required an expanded approach due to its organization into multiple schedule types. A comprehensive county name cleaning program was developed to handle systematic data quality issues including missing state information, trailing whitespace, historical county name changes, and transcription errors. Special procedures were implemented for Nevada counties, which were consistently suffixed with "co" in the original data. Multi-county manuscript pages were processed using establishment-specific assignment rules that considered firm names and geographic context.

County crosswalk merging was performed using many-to-one joins between establishment records and reference datasets, with unmatched records flagged for manual review. FIPS codes were standardized to seven-digit string format through systematic formatting operations that addressed floating-point artifacts and ensured consistent representation. District

of Columbia records received special handling to maintain the standard 11001 FIPS code across all years.

The final workflow processed all 1880 schedule types individually, applying the standardized cleaning procedures and FIPS assignment methodology to each dataset. Throughout the process, county name standardization involved converting forward slashes to hyphens for consistency, and validation checks ensured that FIPS assignments were successfully completed for all processable records.

Once completed, these steps ensured that every file name, and thus every establishment, was assigned to an historical FIPS code.

XII Industry Assignment

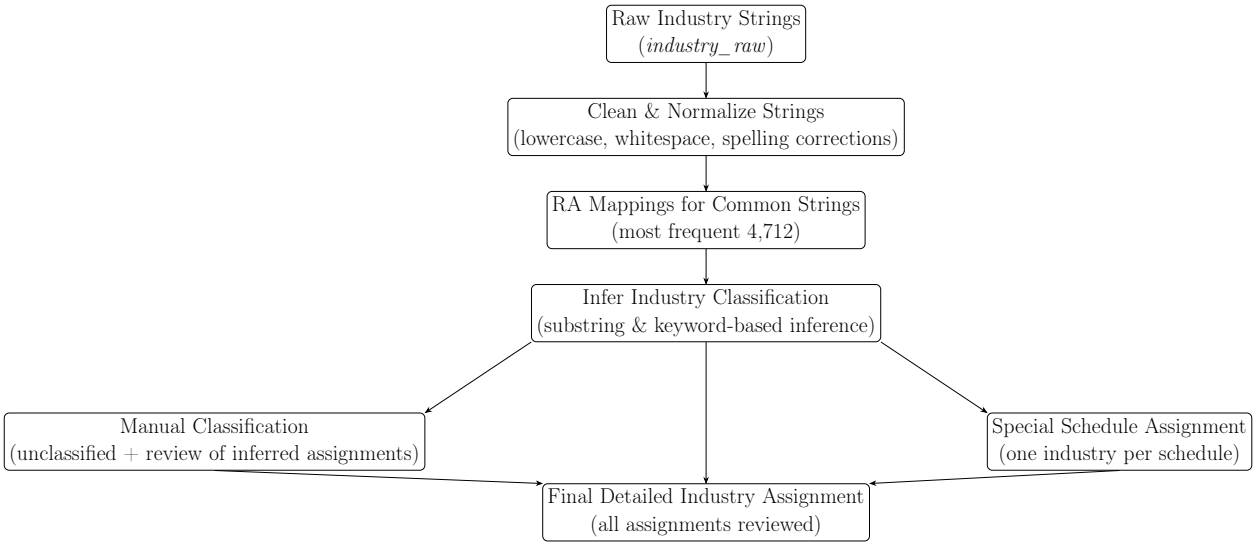
Each establishment's "product name" from the manuscript, which we call *industry_raw* in our final data, was leaned on heavily to categorize each establishment in the 1850-1870 schedule and the 1880 general schedule into an industry. We first established four types of industry categorizations of different generality:

1. Broadest
2. Leontief
3. Granular
4. Detailed

There are 13 broadest industry categories, 31 Leontief industry categories, 212 granular industry categories, and 318 detailed industry categories. Once an establishment is categorized into a detailed category, the other ones follow as there are mappings between detailed and granular, granular and Leontief, and finally Leontief and broadest.

The 31 Leontief categories are the same as the what we use in the county by industry CMF data concordance (Hornbeck and Rotemberg, 2024).

Product name, or *industry_raw*, was first cleaned in the string cleaning step, or section IX. This column was cleaned further in preparation for industry assignment both algorithmically and through crosswalks.



XII.A Raw Industry Cleaning

Raw industry first underwent comprehensive text normalization through conversion to lowercase, whitespace trimming, and removal of extraneous internal spacing. Index variables were generated to maintain record ordering throughout the merging process.

Research assistant-validated raw to clean industry mappings were then integrated through merge operations with external reference files, similar to the string cleaning process. The merge procedure preserved unmatched records while incorporating up to five alternative cleaned industry strings per establishment. Combined industry descriptions were constructed by concatenating multiple cleaned industry possibilities using standardized delimiter patterns, creating comprehensive industry strings that captured the full range of establishment activities.

For establishments lacking pre-validated clean industry strings, automated text cleaning procedures were applied to raw industry descriptions. Standardization operations addressed common formatting inconsistencies by converting conjunction terms and abbreviations to uniform patterns. Frequent misspellings were corrected through systematic string replacement operations targeting historically common transcription errors (e.g. waggon for wagon and cabinett for cabinet).

Four primary establishment categories were defined and flagged through regular expression matching: makers, shops, manufacturers, and factories, building upon previous efforts. The classification process removed type-specific terms from industry descriptions, while preserving the categorical information in separate indicator variables.

Text processing operations eliminated manufacturing-related terminology through hierarchical pattern matching, proceeding from longer to shorter terms to prevent incomplete substitutions. Regular expressions were employed to capture variant spellings and truncated forms of manufacturing terminology. Finally, company designations and partnership indicators were removed through pattern-based substitution operations.

XII.B First Industry Classification

After *industry* was cleaned and prepared, RAs created mappings from the raw industry string to a detailed industry classification. This was done for the most common 4,712 cleaned industry strings; each industry string that appeared 15 or more times was included in this. This mapping was added in, adding industry categorizations for most of the establishments in the data.

XII.C Infer Industry Classification

To extend classification beyond the most common strings, substring and keyword-based inference methods were applied to cleaned industry strings not covered by the RA-created mappings. With some extremely common industries, such as flour and lumber mills, it was possible to assign detailed industries to cleaned industry strings that were not included in the most common 4,712. This was done using sub-strings within the industry strings.

Initial processing for this step involved parsing composite industry descriptions to isolate individual industry terms. Pattern matching procedures were applied sequentially to identify establishments in major industry categories, with flour and grist mills receiving priority classification based on grain-related terminology including wheat, barley, corn, flour, feed, meal, and grist. Establishments with generic operational terms in primary positions were reclassified using secondary industry terms when they matched grain processing patterns.

Lumber industry classification followed similar pattern-based procedures, distinguishing between planed and sawed lumber production through keyword identification. Establishments were classified as lumber producers based on terminology including saw operations, planing activities, lumber production, shingle manufacturing, and board production, with exclusions applied to prevent misclassification of non-lumber activities.

Manual validation procedures supplemented automated classification through integration of external spreadsheets made by RAs with data containing reviewed industry assignments. These additional assignments were made using products of the establishment rather than industry string. The assignments were merged with automatically processed records, with manual classifications taking precedence over automated assignments. Data quality controls ensured consistency in industry terminology and validated the accuracy of automated classification procedures.

XII.D Raw Industry Classification

In the final stage of classification, the raw industry strings for 102,662 establishments were re-examined, including both establishments that had not been classified and those that had been classified with the "infer industry classification" step.

Research Assistants and DDD (Digital Divide Data) associates went through each of the unique raw industry strings (around 50,000) and assigned them to a detailed industry where possible. Through this the vast majority of the 102,000 establishments were categorized, leaving 4,718 establishments in 1850-1880 general schedule without a classification. An important note is that this manual classification took precedence over the automated assignments.

XII.E Special Schedule Assignment

Since each special schedule was tied to a single industry by design, classification was straightforward: all establishments in a given special schedule were directly assigned to the corresponding industry category.

XIII Power Machine Cleaning

To make the information on power used more useful, the *power_kind* variable from 1850 and 1860 was cleaned and categorized. Information about both the type of power and the machine used was extracted. Dummy variables were created indicating what type of power an establishment used and machines were assigned to 274 different machine categories. In 1870, separate questions were asked about the type of power and the machine, so information from both columns were used to assign a power and machine type. The 1880 schedules include extensive information about power used and machines utilized, so it was not necessary to clean the variables in order to assign power types.

This process once again used RA-made crosswalks that mapped the raw information into standardized variables.

XIII.A 1850 and 1860 Power Cleaning

1850 and 1860 information on power was collected in the same way, so we used the same process to clean and categorize both years. We began by importing external reference files that map raw text strings to standardized categories of power types and machines. The raw power source entries were normalized, matched against the reference mappings, and reshaped so that each establishment–power source combination is represented consistently. Records were then consolidated so that each establishment is assigned binary indicators for whether it used steam, water, wind, horse, or hand power.

We also extracted and organized machine information. Establishments identified as having machinery were processed so that each machine type and its associated details are stored in a structured, comparable format. The majority of establishments cannot be identified in 1850 and 1860 with a specific machine.

The cleaned power and machine data were then merged back into the main 1860 establishment dataset. Additional adjustments are made for industries such as flour and lumber mills, where the presence of terms in the industry description provides extra evidence of steam or water power.

The result is harmonized indicators for power sources and systematically structured information on machinery in the 1850 and 1860 data.

XIII.B 1870 Power and Machine Cleaning

Power and machinery information in 1870 was cleaned using a process similar to that of earlier years but with additional refinements. We began by importing external reference files that mapped raw text strings to standardized categories of power types and machines. The raw power source entries were normalized, matched against the reference mappings, and reshaped so that each establishment–power source combination was represented consistently. Records were then consolidated so that each establishment was assigned binary indicators for whether it used steam, water, wind, horse, or hand power.

Machine information was also extracted and organized. Establishments reporting machinery were processed so that each machine type, its unit count, and its associated category were stored in a structured format. The cleaning process in 1870 was designed to prevent mismatches between machine types and categories, ensuring that each machine was correctly classified.

The cleaned power and machine data were then merged back into the main 1870 establishment dataset. As in earlier years, additional adjustments were made for industries such as flour and lumber mills, where references in the industry description provided extra evidence of steam or water power.

The result is harmonized indicators for power sources and systematically structured information on machinery in 1870 data.

XIV Standardize 1880

Materials and products in the 1880 special schedules are recorded in a structured, inflexible way. For most of the schedules, only a predetermined set of items were recorded, with materials and products that were not part of that set getting lumped into "other" categories. For example, brick makers were only able to record specific input information on cords of wood, and meat packers were only able to record product information for the beef, pork, and mutton products they produced.

In order to make analysis of product and materials usage easier across all years, variables were added to the special schedules that mimicked product and materials kinds and quantities variables found in the 1850-1860 data. For example, this is how we processed the information for meat packers' products:

Original Special Schedule Variable	**Standardized Variable**
Beef used	`materials_kind1` / `materials_qty1`
Sheep used	`materials_kind2` / `materials_qty2`
Pigs used	`materials_kind3` / `materials_qty3`

This was done to all of the data for the special schedules, for products and materials kinds and quantities.

XV Numeric Cleaning

Numeric Variables (Capital, Materials value, Labor, Product values, etc.) in the CMF had to be cleaned for two reasons. The first reason is that there were many instances of census marshals putting non-numeric information in numeric columns. Sometimes this was an extra note that the marshal included, for example one wrote in capital "6000 Real' to distinguish between real capital and personal capital invested (although the manuscript makes it clear that a distinction does not need to be made). In another page, a marshal wrote "10 boys" for male workers.

The second reason for numeric cleaning is mistakes in transcription by DDD. This itself comes in two flavors:

1. Non-numeric data from other places in the sheet are accidentally included in numeric columns

(a) e.g. transcribing "raw turpentine" in the male workers column

2. Numeric data were erroneously entered, usually because handwriting was unclear or because cents were mistakenly included in whole-dollar values.

   (a) e.g. 8400 being input as 84000 or 7645.30 being written as 764530

Unlike the string variables, where the raw version of the variable is kept in the data, adjustments made in this step were made directly to the numeric variable. There are no "raw" versions of numeric variables in the final data.

XV.A Non-numeric extraction

The census marshal mistakes/inconsistency and the first type of DDD error are the most easy to identify and correct. We simply brought up every instance of a non-numeric character in a numeric column and RAs made crosswalks fixing these issues where possible. Although this was attempted for every single case, there were many that were not able to be resolved because of illegibility or other reasons. When DDD was unable to read what a census marshal wrote, they put a "?" mark to distinguish between truly missing data and illegible data. Cases with "?" marks were checked and fixed where possible. There are still observations with "?" and other non-numeric data, and the choice of what to do is up to the user.

XV.B Extreme and Improbable Value Checking

Figuring out when a numeric variable was entered incorrectly is a much more difficult task than correcting non-numeric mistakes since there is no trace of a mistake in the entered data itself, besides in the totals columns discussed in section VIII.A. Thus, we needed to use other information recorded about the establishment to detect errors. There were three ways in which this was accomplished.

XV.B.1 Extreme Values Checking

With observation, we noticed that many DDD errors happened with large values since it was easy to lose track of numbers/difficult to read numbers that were squeezed into boxes by census marshals. We brought up the top ten establishments for capital, total materials value, labor, and production value for every year. We checked these top establishments and made fixes where necessary by checking the manuscripts

XV.B.2 Mahalanobis Distance Checks

In order to flag establishments that had potential errors that were not easily identified by looking at extreme values or ratios, we used an establishment's mahalanobis distance (the distance of a point to a distribution, taking into account correlations between variables) using the numeric variables and ratios of those numeric variables.

We first determined ratios to include, e.g. capital and output, materials value and output etc. Once these were chosen, we then calculated mahalanobis distance within a Leontief industry. This was done as we expect different industries to have different levels of variables and different ratios we might expect (for example iron works might have higher capital investment than many other industries in total, and establishments in millinery might have a much higher labor cost to capital ratio). We calculated the mean vector, covariance

matrix and inverse covariance matrix for every industry and then used those to calculate each establishments' mahalanobis distance within their industry.

Mechanically, using raw mahalanobis distance will typically flag the largest establishments. So instead for each establishment we calculated each variable and ratio's normalized contribution of the distance. The Mahalanobis distance of an observation i from the mean vector $\boldsymbol{\mu}$, with covariance matrix $\boldsymbol{\Sigma}$, is defined as:

$$D_i = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}. \tag{1}$$

We can write the squared Mahalanobis distance as a sum of contributions from each variable:

$$D_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = \sum_{j=1}^{p} c_{ij}, \tag{2}$$

where c_{ij} denotes the contribution of variable j to the squared distance.

We can **normalize** these contributions by dividing by the total squared distance:

$$\tilde{c}_{ij} = \frac{c_{ij}}{D_i^2}. \tag{3}$$

By construction, the normalized contributions sum to 1:

$$\sum_{j=1}^{p} \tilde{c}_{ij} = \sum_{j=1}^{p} \frac{c_{ij}}{D_i^2} = \frac{\sum_{j=1}^{p} c_{ij}}{D_i^2} = \frac{D_i^2}{D_i^2} = 1. \tag{4}$$

Since these normalized contributions sum to one, we have a proportional measure of each variable's influence on the Mahalanobis distance. This not only allowed us to pinpoint which variable, or set of variables was causing most of the distance, this allowed us to make comparisons between establishments. For each variable and ratio, we took the top 0.1 percent of establishments in terms of mahalanobis distance share and checked them. Errors were compiled and applied to the data.

References

Atack, Jeremy, and Fred Bateman. 1999. "Nineteenth-Century US Industrial Development through the Eyes of the Census of Manufactures: A New Resource for Historical Research." *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 32(4): 177–188.

Atack, Jeremy, Fred Bateman, and Thomas Weiss. 1980. "The Regional Diffusion and Adoption of the Steam Engine in American Manufacturing." *The Journal of Economic History*, 40(2): 281–308.

Bohme, Frederick G. 1987. "U.S. economic censuses, 1810 to the present." *Government Information Quarterly*, 4(3): 221–243. Special Issue Symposium on the Economic Censuses, United States Bureau of the Census.

Carroll D. Wright, William O. Hunt. 1900. "The History and Growth of the United States Census." Census Bureau.

Delle Donne, Carmen. 1973. *Federal Census Schedules, 1850-1880: Primary Sources for Historical Research.* Washington, DC: National Archives and Records Service.

Don, Delle. 1973. "Federal Census Schedules 1850-1880: Primary Sources for Historical Research." National Archives.

Fishbein, Meyer H. 1973. "The Censuses of Manufactures 1810-1890." *Reference Information Paper, National Archives.*

Haines, Michael R. 2010. "Historical, Demographic, Economic, and Social Data: The United States, 1790–2002." *ICPSR Volume 2896.*

Hornbeck, Richard, Anders Humlum, Shanon Hsuan-Ming Hsu, and Martin Rotemberg. 2025. "Gaining Steam: Technology Diffusion with Recurring Lock-in." Resubmitted to the *Journal of Political Economy.*

Hornbeck, Richard, and Martin Rotemberg. 2024. "Growth Off the Rails: Aggregate Productivity Growth in Distorted Economies." *Journal of Political Economy*, 132(11): 3547–3602.

U.S. Census Office. 1850. *Statistical View of the United States.* Washington, DC: U.S. Government Printing Office.

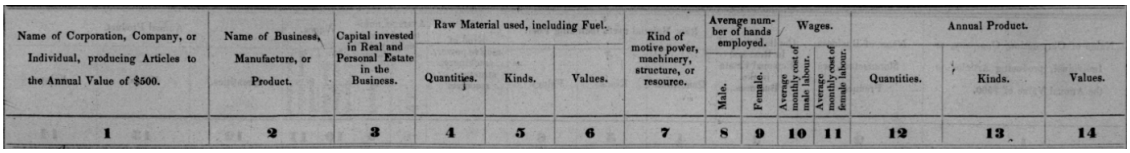
U.S. Census Office. 1860. *Instructions to U.S. Marshals.* Washington, DC: U.S. Government Printing Office.

U.S. Census Office. 1870. *Instructions to Assistant Marshals.* Washington, DC: U.S. Government Printing Office.

U.S. Census Office. 1880. *Instructions to Marshals and Assistants.* Washington, DC: U.S. Government Printing Office.

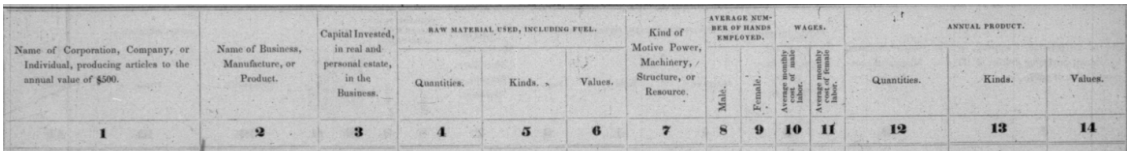
Figure 1. Example Census Images: Questions by Year

Panel A. 1850



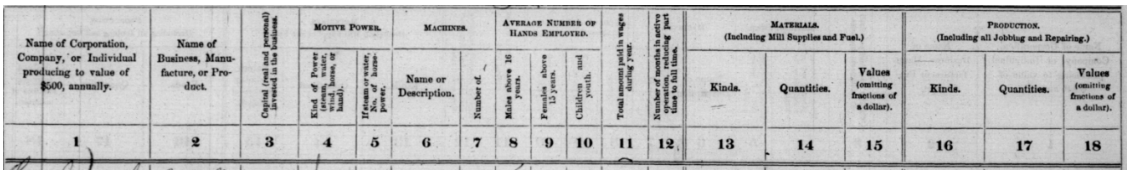
Name of Corporation, Company, or Individual, producing Articles to the Annual Value of $500.	Name of Business, Manufacture, or Product.	Capital invested in Real and Personal Estate in the Business.	Raw Material used, including Fuel.			Kind of motive power, machinery, structure, or resource.	Average number of hands employed.		Wages.		Annual Product.		
			Quantities.	Kinds.	Values.		Male.	Female.	Average monthly cost of male labour.	Average monthly cost of female labour.	Quantities.	Kinds.	Values.
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Panel B. 1860



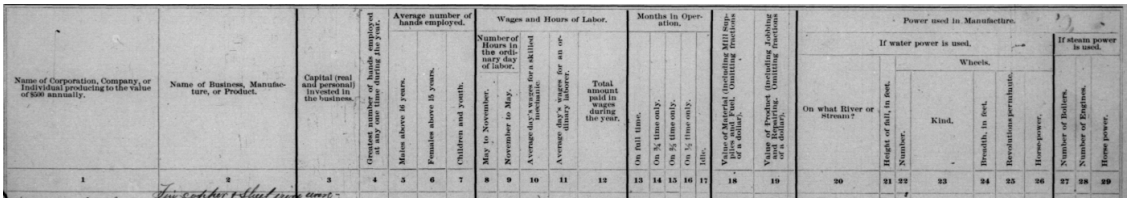
Name of Corporation, Company, or Individual, producing articles to the annual value of $500.	Name of Business, Manufacture, or Product.	Capital Invested, in real and personal estate, in the Business.	RAW MATERIAL USED, INCLUDING FUEL.			Kind of Motive Power, Machinery, Structure, or Resource.	AVERAGE NUMBER OF HANDS EMPLOYED.		WAGES.		ANNUAL PRODUCT.		
			Quantities.	Kinds.	Values.		Males.	Females.	Average monthly cost of male labor.	Average monthly cost of female labor.	Quantities.	Kinds.	Values.
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Panel C. 1870



Name of Corporation, Company, or Individual producing to value of $500, annually.	Name of Business, Manufacture, or Product.	Capital (real and personal) invested in the business.	MOTIVE POWER.		MACHINES.		AVERAGE NUMBER OF HANDS EMPLOYED.			Total amount paid in wages during year.	Number of months in active operation during past year, time to full time.	MATERIALS. (Including Mill Supplies and Fuel.)			PRODUCTION. (Including all Jobbing and Repairing.)		
			Kind of Power, (steam, water, wind, horse, or hand).	If steam or water, No. of horse-power.	Name or Description.	Number of.	Males above 16 years.	Females above 15 years.	Children and youth.			Kinds.	Quantities.	Values (omitting fractions of a dollar).	Kinds.	Quantities.	Values (omitting fractions of a dollar).
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Panel D. 1880

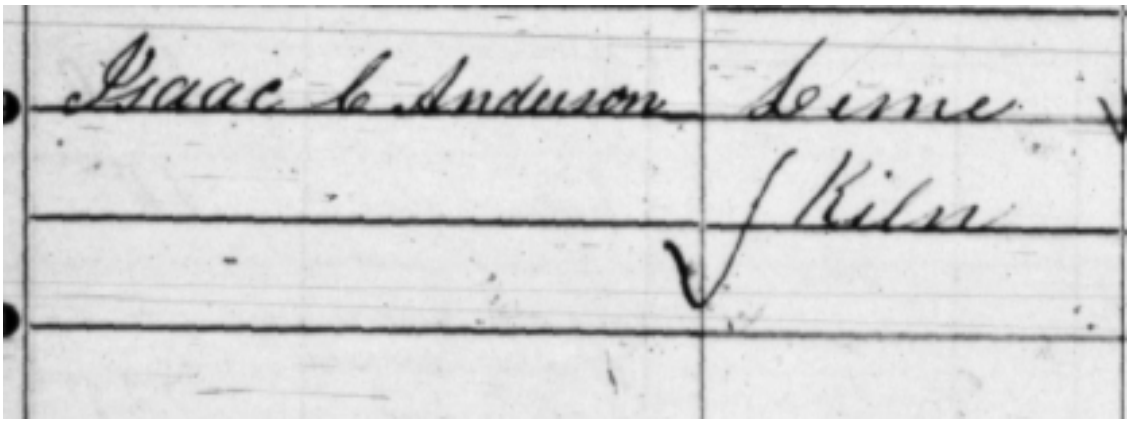


Name of Corporation, Company, or Individual producing to the value of $500 annually.	Name of Business, Manufacture, or Product.	Capital (real and personal) invested in the business.	Greatest number of hands employed at any one time during the year.	Average number of hands employed.			Wages and Hours of Labor.					Months in Operation.					Value of Material (including Mill Supplies and Fuel) consumed, (omitting fractions of a dollar).	Value of Product (including Jobbing and Repairing) (omitting fractions of a dollar).	Power used in Manufacture.								
							Number of Hours in the ordinary day of labor.		Average day's wages for a skilled laborer.	Average day's wages for an ordinary laborer.	Total amount paid in wages during the year.	On full time.	On ¾ time only.	On ½ time only.	On ¼ time only.	Idle.			If water power is used.						If steam power is used.		
																			On what River or Stream?	Height of fall, in feet.	Number.	Kind.	Breadth, in feet.	Revolutions per minute.	Horse-power.	Number of Boilers.	Number of Engines.	Horse power.
																					Wheels.						
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29

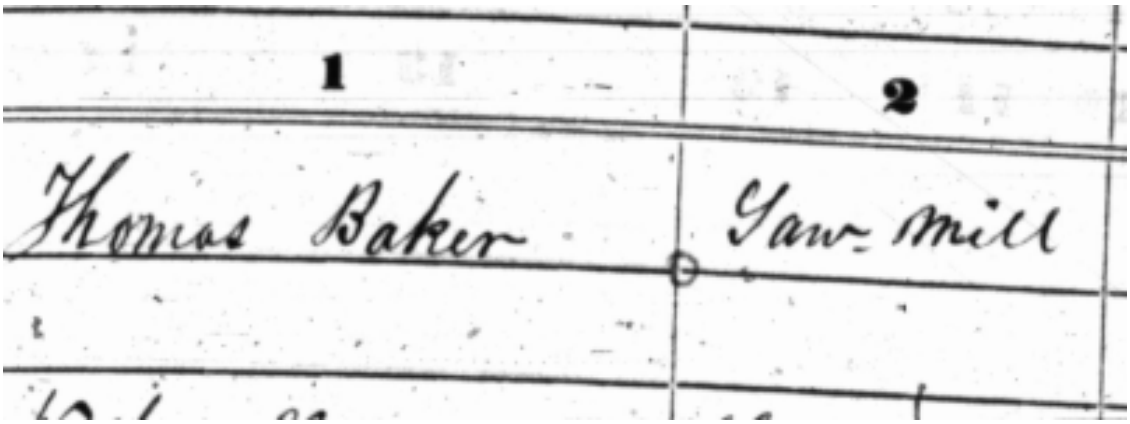
Notes: These images show the columns and questions that census marshals used when they interviewed establishments in 1850-1870 and in the 1880 general schedule.

Figure 2. Example Census Images: Questions by Year

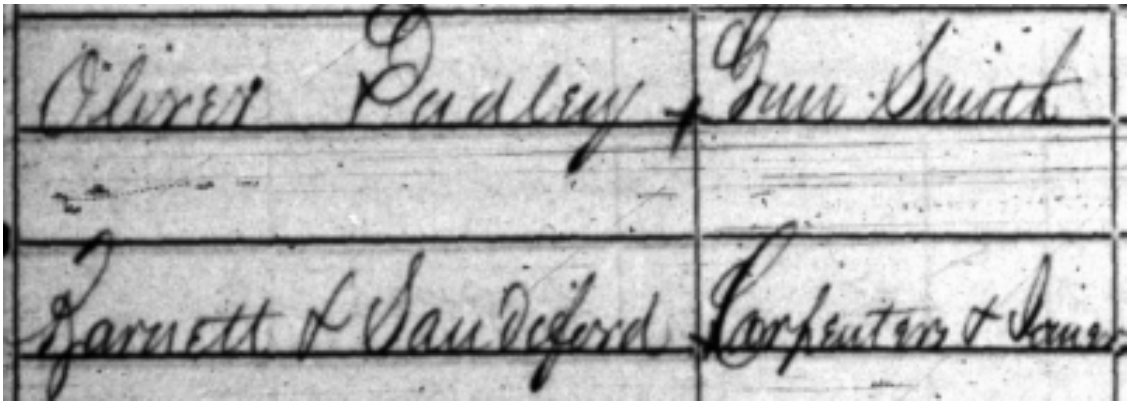
Panel A. Check Mark



Panel B. Circle



Panel C. X-mark



Notes: These marks show various ways in which census tabulators marked off establishments that had been added to the eventually published county and county by industry totals.