Mill Establishment Panel Documentation

July 2026

I Establishment Panel Linking

This document describes our creation of a linked panel of milling establishments over time. The Census manuscripts do not have a time-consistent identifier for each establishment, just as in the Censuses of Population (Ferrie, 1996; Feigenbaum, 2015; Ruggles, Fitch and Roberts, 2018; Bailey et al., 2020; Abramitzky et al., 2021; Price et al., 2021), so we generate our own links.

We define a stable manufacturing establishment based on its owner name, industry, and place. If an owner shuts down an establishment and reopens an establishment in a different county, we consider that a new establishment.[1] Similarly, if the owner changes their establishment to no longer be a mill, we consider the mill closed.[2] While we link establishments with partial ownership changes (such as a son taking over from his father), if the establishment's ownership changes entirely, with no clear link between previous and new owners, then we also consider that a new establishment. This is dictated by data availability, and also raises philosophical questions about what is a surviving establishment. Our view is that mill owners at the time were sufficiently involved in the operation of the establishment that entire ownership changes are akin to closing operations and selling capital assets to a new venture.[3]

We hand-link establishments over time, within a county, using data on owner or company names, industry, product types, and (when available) nearest post office. Importantly, we do not use mills' type of power to make the panel identifiers. We hand-linked all lumber and flour mills, across each decade. Two people searched for matches for each mill, and we reconciled any disagreements.

I.A Panel-Linking Procedure

We link mills by hand, from one decade to the next, in combination with a machine-learning linkage model. We employed a team of data associates to compare a mill in one decade to plausible matches in the subsequent decade. We matched mills on name and location, but did not force establishments to be in the same industry in every decade. Because mills rarely switched between lumber and flour, and we consider working in a different manufacturing sector to be part of the outside option in our model, we treat non-mill industry switches as "exits."

To guide the large-scale hand-links, we first matched a few counties and compared every

[1] These cross-county "migrations" appear unusual for millers, based on historical society records.

[2] When we hand-linked the establishments, we allowed for cross-industry links and found very few outside of milling. Around 4% of surviving mills switched between lumber and flour.

[3] We do find evidence of ownership transfers in historical accounts, though most business closures appear to be associated with the mill no longer being operated. We discuss elsewhere the implications of unobserved reselling, and we use the quantitative model to estimate how local technology choices affect the relative purchase prices of steam and water power, which captures if the transition to steam power lowered the purchase price of water power.

mill to every manufacturing establishment in the subsequent decade. We then trained a machine learning algorithm on those matches. For the large-scale hand-linking, we then only considered potential matches with a relatively high linking probability. For the possible matches, we included candidates with over a 9% linking probability. For mills with many potential links, we only sent the top twenty; for mills with few potential links, we sent the top five as long as their linking probabilities were above 5%. In practice, the potential links with a low match probability were rarely hand-chosen as an actual match. For the analysis in the paper, we then retrained the machine-learning model on the full set of matches. Below, we describe our approach in more detail.

I.A.1 Hand-Linking Procedure

Our first step was to create some panel links by hand, linking establishments in 1860 to their 1870 counterparts in 97 counties. We chose relatively small counties, to start, so it was feasible to compare all possible matches in the same county: matching 2,709 establishments in 1860 to 5,518 candidate establishments in 1870.

To make the links, we considered each establishment's name, industry classification (including the self-reported string and our own cleaned industry measures), and the nearest post office. We also had access to the original CMF manuscript images for each establishment to double-check mistakes, either in the original handwriting or its transcription. Each hand-linking sheet was completed by two UChicago students, and assigned to a third person to reconcile any discrepancies. For each 1860 establishment, we sorted all 1870 candidates by Jaro-Winkler (JW) name similarity, and by whether or not their broad industries matched, to increase the likelihood that links were at the top of each block of names.

Broadly, we made two types of matches in the data. "Direct" matches are when the establishment names in both periods are close matches. This is similar to common practice in literature linking men across decades in the Census of Population (Ferrie, 1996; Feigenbaum, 2015; Ruggles, Fitch and Roberts, 2018; Bailey et al., 2020; Abramitzky et al., 2021a,b). However, an important difference between linking men and linking establishments is that many mills *actually* changed their names, especially when adding owners. While additional data would be needed to link women who change their last names, our Census of Manufactures data can tolerate moderate changes in ownership. To account for "ownership transfers," we also match establishments where part of the name is very similar but another part is different in a manner consistent with a partial change in ownership. In practice, this second category includes partnership formation or newer members taking on the family business.[4]

I.A.2 Model Specification

From hand-linking establishments, we noticed there were broadly four categories for how the establishment's name was reported (consistent with guidance from Jeremy Atack). These were not formal rules, but we list the categories below along with our interpretation of their meaning.

1. Establishments with sole proprietorship contain a single owner's name. Names were sometimes initialized, and the names did not consistently follow a first/last name order.

[4]In our replication files, we denote direct matches as "y", ownership transfer matches as "o", and non-matches as "n". We denote direct matches where the industry changed within milling as "s".

2. Establishments owned by families normally appeared as a person's name followed by *& sons* or *& brothers.* Others appeared with two first names separated by an ampersand, followed by a last name.

3. Establishments that were a partnership or expanded partnership reported two or more names of the proprietors; limited partnerships reported one or more people's names followed by *& co.*

4. Establishments that reported names that were impersonal, and often included tokens related to the business and location.

For our mills, in particular, there were two broad types of naming patterns: those with general company names, sometimes including the name of the water power source; and those named after people. Across Census decades, the order of people's names can change. Even for establishments with a single owner, the order of first and last names can change, along with changes in the use of initials.

These features motivate us to build two separate linking models: one matching the whole establishment name, and one matching owners' names with flexibility in their ordering.[5] We use two random forest models to predict establishment pairs, either tracking the company as a whole or tracking individual owners.[6] Both linking models predict establishment pairs to be: a same-owner match, an ownership transfer match, or not a match. We describe this approach in more detail below.

I.A.3 Name Classifier

We built a name classifier to categorize establishments by their naming pattern type, extract the name of the owners, and identify the name order. While owner names are embedded in establishments owned by sole proprietors, families, partners, or expanded partnerships, the names were often initialized and would switch first-last name orders.

We first use a list of company tokens to identify establishments with impersonal names, which includes: names of locations, such as state and county names; and tokens related to their product or business, such as tanning, manufacturing, lumber, etc.

For establishments without those company tokens, we implement the following steps to extract and format the owner names. First, we remove the non-name tokens, such as "& co" or "& sons," and split the establishment names into owners' names. For a family-owned establishment with two first names and one last name, we assign the last name to both owners (e.g., turn "J & D. Taflinger" into "J Taflinger" and "D. Taflinger.") We then standardize common nicknames and abbreviations to their original names (e.g., Wm to William and Geo to George.) We determine the name order using the first and last name frequency in the 1880 Census of Population. When both names can be first or last names, we keep both orders and look for both of them in the next Census decade.

[5]We are grateful to Jeremy Atack for suggesting this approach.

[6]We generated linking models based on several classifier families, including logistic regression, random forests, and extreme gradient boosting (Chen and Guestrin, 2016). After evaluating their performance on the validation data, we settled on a random forest trained using the R library `ranger`. The random forest model provided the most reliable output, with respect to false positive and negative rates, and the empirical distribution of predicted probability does not concentrate on the two ends which leaves room for setting the probability threshold and varying the false positive and false negative errors.

I.A.3.1 Owner Linking Model Our owner-linking model predicts links based on three sets of information: establishment name, industry, and post office. We define several sets of variables for each of the first, middle, and last names: Jaro-Winkler string distance, whether the name is initialized, and whether the initial matches exactly. When there are missing values, which are incompatible with the random forest model, we assign the median value and define an indicator flag for missing. For industry, we use our industry classification based on the raw industry string to create matching indicators for broad and detailed industries. We also create a measure of industry distance based on the industry classification and similarity in their reported kinds of products. For post office, we use the Jaro-Winkler string distance between post office names and an indicator for missing values.

For establishments with multiple owners, the model predicts matches at the establishment-owner level. At the predicting stage, we take the maximum of the predicted probability for each establishment pair (from all owner pairs) to let the output be at the establishment-pair level. This process allows a firm to match when one owner is the same, even if other owners are different, which mimics how humans generally make links.

I.A.3.2 Company Linking Model The company-linking model also predicts links based on establishment name, industry, and post office. However, instead of extracting the owner information from the establishment names, this model uses the full string of establishment names and looks for establishments with similar whole names. We use the Jaro-Winkler string distance for the full names, in addition to string distance after removing business and location tokens and the minimum string distance between those remaining tokens among all token pairs. The remaining name distances measure the name similarity unrelated to the business itself, which removes false matches that only have closer string distances on the full name because of common tokens (e.g., "Eagle Mill" and "James Mill").

I.A.4 Model Prediction Reconciliation and Hand-Linking

We use both models to predict matches, separately, and then take the maximum of the predicted probabilities. For the set of potential matches that we consider when making hand-links, we select the top 20 pairs with a linking probability above 9%. If there are 5 or fewer pairs to send, we send the top 5 pairs with a linking probability above 5%.

We worked with Digital Divide Data (DDD) in Kenya to hand-link the matches, at scale. Our team helped train the DDD associates in person, who also had experience linking individuals across decades in the Census of Population. We then continued to work closely with them remotely, handling the data process ourselves while their managers handled HR.

We sent DDD lists of all potential matches with identifying information: establishment name, industry, post office, and product kinds produced. We did not include the estimated linking probabilities. Two separate members of the DDD team found the best match for each establishment, or indicated no close match, and a third random member reconciled any disagreements between the original two members.

We then iterated on these hand-links using the machine-learning model, asking them to manually check "unlikely" matches or "likely" non-matches. We used the same protocol as for the original data, sending DDD the information about the firm but not the estimated link probability. First, we flagged the following three sets of potential matches for review: (1) links that were made for which the algorithm predicted link probability was below 40%, (2) mills with no links, but for which the algorithm predicted at least one link probability

above 40%, and (3) if DDD and the highest-predicted link were different (and the predicted link probability of the actual match was at least 0.1 lower than the best predicted match). For all mills that met one of these three criteria, we resent all of the candidate matches back to DDD for hand-linking. After iteration, the "unlikely" hand-linked matches were generally found to be reasonable matches (and missed by the machine-learning model) and the predicted "likely" matches were also generally decided to be matches after a second look. The automated linking model performed relatively worse in identifying ownership transfers, compared to the hand-links.

References

Abramitzky, Ran, Leah Boustan, Elisa Jácome, and Santiago Pérez. 2021a. "Intergenerational Mobility of Immigrants in the United States over Two Centuries." *American Economic Review*, 111(2): 580–608.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature*, 59(3): 865–918.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021b. "Automated Linking of Historical Data." *Journal of Economic Literature*, 59(3): 865–918.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How Well do Automated Linking methods perform? Lessons from US Historical Data." *Journal of Economic Literature*, 58(4): 997–1044.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." 785–794.

Feigenbaum, James. 2015. "Automated Census Record Linking: A Machine Learning Approach." Working Paper.

Ferrie, Joseph P. 1996. "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 US Federal Census of Population to the 1860 US Federal Census Manuscript Schedules." *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29(4): 141–156.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. 2021. "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set." *Explorations in Economic History*, 80: 101391.

Ruggles, Steven, Catherine A. Fitch, and Evan Roberts. 2018. "Historical Census Record Linkage." *Annual Review of Sociology*, 44: 19–37.

Wright, Marvin N, and Andreas Ziegler. 2015. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *arXiv preprint arXiv:1508.04409*.