Linking the US Censuses to Create a Record-Based Tree

Joseph Price

BYU Record Linking Lab (rll.byu.edu)

Record-Based Tree

A family tree is a way to link family members together. Some family trees are created by a single individual (such as the ones on Ancestry) and others a wiki-style tree in which individuals contribute to a single common tree (such as FamilySearch and WikiTree). The conclusions on these trees rely on the information from sources and all of these platforms provide a way for individuals to attach sources to support their conclusion. All of these require humans to contribute all of the conclusions through their own research. This research is often assisted by machine algorithms to find records and additional information but all of the information is entered by hand.

A record-based-tree can be created through an automated approach that links individuals across multiple records and then reconstructs a family tree based on the relationships that we observe in those records. We are creating a record-based tree using the US census records from 1900 to 1940. There were 188 million unique people that appear in at least one federal census record during this time and they will each have a unique row in our Census Tree. They will also be connected to their parents, siblings, spouses, and children if we observe these relationships in at least one of the census. For example, a person might appear with their parents and siblings in the 1900 census but appear with their wife and children in the 1920 census.

Linkage Process

The key innovation of our approach is that we use linkages creates by individuals doing family history research as training data and then use machine learning to predict all of the other linkages. The image below provides an example of the process for a case with three people.

ID	Ark1900	Ark1910	Ark1920
PID1	A	B	
PID2		C	D
ID1	E		
ID2	F		
ID3		G	
ID4			H
ID5			I

Example: 3 people, 3 censuses
Training data = 2 matches
7 rows, 3 people = 4 matches to find

ID	Ark1900	Ark1910	Ark1920
PID1	A	B	H
PID2		C	D
ID1	E		
ID2	F		
ID3		G	
ID5			I

ID	Ark1900	Ark1910	Ark1920
PID1	A	B	H
PID2	E	C	D
ID2	F	G	I

Final product: 3 rows, 3 people

Our goal in this simple example is to link three people across three records. The first two people (PID1 and PID2) are already on the Family Tree and each have two census records attached. PID1 is attached to record A and B and PID2 is linked to record C and D. We use these linked records as training data to identify the other matches. Each time we find a match, one of our rows is eliminated. We start with 7 rows and 3 people, so this means that we need to find 4 matches to get to our final fully linked dataset.

In practice, our training set includes about 40 million true linked pairs and we use that to identify 376 million matches across the 1900-1940 census. The average person in our sample appears in about 3.5 records. We also link each person to each of their family members by using the family relationship codes that appear in each census records. This allows us to connect individual's across time and reconstitute family networks.

Sparse Data Linking

One of the advantages of the Census Tree is that it can be used as an intermediate platform to link any other type of record together. Archives often have rich content that doesn't include a lot of biographic information. A photograph might include a name and a place with the ability to approximate the person's age. Letters, journals, or other personal records might include names, places, and relationships but no information about birth place or age. The Census Tree makes it possible to conduct sparse data linking. It provides information on nearly every single person in the US and where they are living over time and who they are related to. This makes it possible to identify the candidate set of people that might be the person in the record. In many cases, every little bit of additional information available in the record can narrow the field down to a single match.

We have done three applications of this type of sparse data matching: names of soldiers who died in World War I, patents, and high school yearbooks. In each case, all we had to match on was a name, a town, and an approximate age. The Census Tree allowed us to link a surprising fraction of these records to at least one census record (usually the one closest to the time of the event). Once a record is linked to the Census Tree, then it can be linked to all other records that are linked to that person on the Census Tree. Thus the Census Tree becomes the pin board that we can attach all records to and use it to connect those records together.

Engagement

An important purpose of archives is to help a community or country remember and connect with its past so that we can collectively learn from those lessons. The Census Tree makes it possible to allow people to personally connect with different record collections. A book of pictures of soldiers who died in World War I is interesting but it creates a powerful connection when you can point me directly to one of my relatives. A museum can be a great place to learn about the past but that past can become more real when I discover my personal connection to the people who were there.

One example of how this could work is that when a person visits an archive, special collection, or museum, they would have the option of typing in some information about a family member who was alive in 1940. Based on that information, we would help them find their family member in the 1940 census at which point they would be connected to anything that we have connected to the Census Tree. Thus a person visiting an exhibit on World War I could be pointed to sections of the exhibit that include photos or stories of their relatives. A classroom lesson using materials from an archive about the Civil Rights Movement could start by members of the class learning about their personal connection to people involved in those events. We think this level of engagement could dramatically increase the interest by the general public in the data collections kept by archives and museums and create a renewed reason to preserve these records.