A Short Survey on Census Matching Methods

Xiaoyun QIU

August 23, 2017

1 Introduction

Record linkage is not a new topic. On the contrary, there is a strand of researches in statistics studying the theory of record likage. In economics, records linkage can be applied to many different cases. For example, historical census records usually lack unique identifier and thus constructing a panel census records requires matching individuals across years. Another example is to link records of different sources of data without unique identifier. Instead, only names and addresses are available.

Historical census records are usually with large size. Manually constructing matched census records are costly and hard to replicate. The main challenge in matching cencus records is the inavailability of unique identifier, for example, social security number. Instead, pre-determined information is used to identify the same person in different census data, including names, age, gender, birth place and race, etc.. However, the combination of this information is not a perfect unique identifier. There are often noises in these data and they are not unchanged between census waves. In addition, the size of data further increases the implementation possibility of systematically linking the entire census records across years. As the advancement of computational powers, econmists can now develop automated methods to link census records across years, in order to utilize richer dimensions of information.

Fellegi and Sunter (1969) developes a theory for record linking and disccues its practice using traditional statistical methods. However, not until recently, machine learning is becoming a popular method, which is famous for its strength in prediction. Applying machine learning methods creates new possibility of record linkage.

In this article, I try to formalize the census matching problem and surveys the matching methods used in economics literature.

2 Formalization of the Matching Problem

2.1 A Classification Problem

The core of census matching is a classification problem. Given any pair of records from different census years, finding a true match is to find the mapping that classifies the pair as matched or unmatched based on the set of pre-determined features, including names, gender, age, race and birth place. However, since this set of features is far from unique, there are cases that the same individual is

showing in different pairs classified as matched. In other words, one individual is matched to several candidates in another census year. Therefore, census matching is not just a classification problem, but we need to consider the fact that each individual in one census year only has at most one true self in another census year.

Supose there are two censuses $A = \{a_1, a_2, ...a_i, ..., a_n\}$ and $B = \{b_1, b_2, ..., b_j, ..., b_m\}$. One can view A as the earlier year census while B is the latter year census. Set

$$X = A \times B = \{(a_i, b_j), a_i \in A, b_j \in B\}$$

is the set of all possible pairs. Matching is to find a classifer

$$h : X \to \{0, 1\},$$

where 0 means not a match while 1 means a match. Therefore, set X can be expressed as

$$X = M \cup U,$$

where

$$M = \{x \in X, h(x) = 1\},$$

and

$$U = \{x \notin X, h(x) = 0\}.$$

2.2 Distinguishing Different Types of Match

Based on the classifer h, set A and set B can be expressed as the union of a sequence of disjoint equivalence class, where in each equivalence class, the elements are classified as true matched to the same record in another census year. Define

$$[a_i] = \{b_j \in B, h(a_i, b_j) = 1\},$$

which includes all records in set B that are matched to a_i in set A. However, in each set, there are some records that cannot be matched. For example, for those showing up in the earlier census, some of them are able to be matched to any records in the latter year because of names changed, mortality, migration and so on, while for those showing up in the latter census, some of them are not able to be matched bacause they are not born. Therefore, there are some a_i such that $[a_i] = \emptyset$. Suppose there are n_1 of them, and reorder the elements in set A such that

$$[a_i] = \emptyset, i = 1, 2, ..., n_1.$$

Denote the subset A_0 to include those not be able to be matched in set A and the subset B_0 to include those not be able to be matched in set B. Set B can be partitioned into the union of the disjoint sets

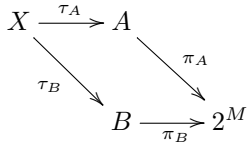
$$A = (\bigcup_{i=1}^{n_1} [a_i]) \cup (\bigcup_{i=n_1+1}^{n} [a_i]) \cup B_0.$$

Similarly, define

$$[b_j] = \{a_i \in A, h(a_i, b_j) = 1\},$$

and reorder the elements in set B such that

$$[b_j] = \emptyset, j = n_2 + 1, n_2 + 2, ..., m.$$



Therefore, set A can be expressed as the union of disjoint sets

$$B = A_0 \cup (\bigcup_{j=1}^{n_2} [b_j]) \cup (\bigcup_{j=n_2+1}^{m} [b_j]).$$

Before we distinguish the 4 types of matched pairs, we define some mapping rules among sets.

$$\tau_A : X \to A$$
$$\tau_A((a_i, b_j)) = a_i$$

Similarly,

$$\tau_B : X \to B$$
$$\tau_B((a_i, b_j)) = b_j$$

$$\pi_A : A \to 2^M$$
$$\pi_A(a_i) = \{(a_i, b_j) \in X, (a_i, b_j) \in M\}$$

Similarly,

$$\pi_B : B \to 2^M$$
$$\pi_B(b_j) = \{(a_i, b_j) \in X, (a_i, b_j) \in M\}$$

The first type is called 1-1 match.

$$M_1 = \{(a_i, [a_i]) \in M, |\,[a_i]\,| = 1 \quad and \quad |\pi_B([a_i])| = 1\}$$

The second type is called 1-x match.

$$M_2 = \{(a_i, [a_i]) \in M, |\,[a_i]\,| > 1 \quad and \quad \forall b_j \in [a_i]\,, |\pi_B(b_j)| = 1\}$$

The third type is called x-1 match.

$$M_3 = \{(a_i, [a_i]) \in M, |\,[a_i]\,| = 1 \quad and \quad |\pi_B([a_i])| > 1\}$$

The forth type is called x-x match.

$$M_4 = \{(a_i, [a_i]) \in M, |\,[a_i]\,| > 1 \quad and \quad \exists b_j \in [a_i]\,, |\pi_B(b_j)| > 1\}$$

The 4 types of match can be defined using the notation of $[b_j]$ and everything remains the same.

In current matching literature, most of them focus on the first two types of match, while the third of forth types of match are rarely mentioned. Ignoring the third and forth types of match pairs will increase matching rate. However, in matching census data, linking different individuals from the previous year to the same person in the latter year will introduce false links with no doubt and thus should be avoided. (Is this one of the innovations of our matching method?)

2.3 A Pruning Process

The first stage of census matching is a standard classification problem. However, given the specific setting of the question, we need to implement the second stage to filter true match from non-unique match as much as possible.

The filer procedure or pruning procedure is a mapping

$$g : M \to \{0, 1\},$$

where 0 means not passing the filter procedure, while 1 means passing the filter procedure.

For the first type of match,

$$\forall y \in M_1, g(y) = 1.$$

For the second type of match, given the existence of the best match,

$$\exists! y_i \in (a_i, [a_i]), s.t. g(y_i) = 1, i = n_1 + 1, ..., n$$

$$\forall x \in (a_i, [a_i]), x \neq y_i, g(x) = 0, i = n_1 + 1, ..., n$$

For the third type of match, given the existence of the best match, it is uniquely selected in the same way.

For the forth type, first select the best match for each $[a_i]$, which is similar to selecting the best match from 1-x matched pairs. Then for each x-1 matched pairs, select the best match.

The above reasoning can be established using the eqivalence class of elements in set B. Therefore, there should not be any difference in forward matching and backward matching in theory.

The census matching method should be a combination of the classification method and the filter method.

3 Survey on Matching Literature

Using the terminology of machine learning, the current census matching methods used in economic history can be classified into 3 lines. In a typical classification task, the method models the probability distribution of each outcome category explicitly is called generative learning algorithm, while the method finiding the clssifier that directly maps features to different outcome categories is called discriminative algorithm.

- Decision tree method: Ferrie (1996), Abramitzky, Boustan, and Eriksson(2011, 2012, 2014); Long (2005, 2006, 2008); Long (2007, 2013); Nix and Qian (2015)
- (Unsupervised) generative learning method: Fellegi and Sunter (1969); Mill and Stein (2016), Perez(2017)
- Supervised discriminative learning method: Vick and Huynh (2011); Goeken, Huynh, Lenius, and Vick (2011); Feigenbaum (2016)

Paper	Census	Country	Gender, Age and Race	Methods	Matching Goal	Matching Rate
Long and Ferrie (2005)	1851-1881	Britain	Male, age 25 and under	Decision Tree	accuracy	20
Long and Ferrie (2005)	1850-1880	US	Male	Decision Tree	accuracy	22
Abramitzky, Boustan, Eriksson (2012)	1865-1900	Norway	Male	Desicion Tree	accuracy	8.34
Nix and Qian (2015)	1900-1910	US	Male, age 3-15 in 1865	Desicion Tree	efficiency	72
Mill and Stein (2016)	1910-1940	US	Male, Black and Mulatto	Generative	efficiency, accuracy and representiveness	11 or 34
Perez (2017)	1869-1895	Argentina	Male	Generative	efficiency and accuracy	around 10
Vick and Huynh (2011)	1870-1880	Maryland, US	Male	Discriminative, SVM	accuracy	1.11
Vick and Huynh (2011)	1870-1880	Maryland, US	Male, White	Discriminative, SVM	accuracy	1.17
Vick and Huynh (2011)	1870-1880	Maryland, US	Male, African Americans	Discriminative, SVM	accuracy	0.98
Vick and Huynh (2011)	1870-1880	Torm, Norway	Male	Discriminative, SVM	accuracy	1.97
Feigenbaum (2016)	1915-1940	Iowa State, US	Male	Disciminative, probit	efficiency and accuracy	57.37

In the decision tree method, researchers specify a set of ad-hoc rules defining what is a true match in practice. For example, in Abramitzky, Boustan, and Eriksson(2012), records with the same standardized names, with the same birth place and their age gap is within 3 years are considered as a true match. One of the main advantages of decision tree method is simplicity and thus it is easy to implement. However, the simplicity usually comes at a cost that the matching results may not be solid.

While the last method is a supervised machine learning method using a training sample to tune some important decision parameters, the second method can be supervised or unsupervised, depending on whether a training sample is used to select the parameters. Apart from the usage of training sample, another main difference is that generative learning methods explicitly models the underlying probability distribution of each category while discriminative learning mathod only cares about finding the classifier. For example, in Perez (2017), the author assumes that conditional on true match, the distance of JW distance of names and age difference follows a multinomial dirstibution. Applying Bayes rule, the method tunes threshold parameters for the distances that maximize the probability of the observed distance vector. However, without using training sample, the parameters are chosen by using an iterative algorithm called EM algorithm. Compared to the decision tree method, where researchers determine ad-hoc thresholds for the distances, generative learning method uses parametric model to select thresholds.

Different from generative learning method, discriminative learning method does not explicitly model the probability distribution of each outcome category. On the contrary, it aims to find the optimal "boundary" (or hyperplane in machine learning methods) for different outcome categories without making assumption on the underlying probability distribution. In modeling methods, researchers can choose parametric model (Feigenbaum, 2016), or non-parametric model (Vick and Huynh (2011); Goeken, Huynh, Lenius, and Vick (2011)), for example, machine learning algorithm, SVM, RFV, etc..

4 Evaluation of Matching Results

Few of the current matching literature evaluates the matching results, for example, the false positive rate and true positive rate, except for Feigenbaum (2016). The main reason is that it is costly to manually match the records and then evalute the matching results. However, evaluation is feasible for supervised matching methods. Using the training sample, one can employ the cross-validation method to select the appropriate models and evalute the matching results. For example, in hold-out cross vaidation, the training sample is ramdomly separated into the training set and cross validation set, where training set contains 70% of data points. One can train the model on the training set only and use the

cross validation set to evaluate the methods.

References

[1] Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review*, 102(5):1832–1856, August 2012.

[2] James J. Feigenbaum. Intergenerational Mobility during the Great Depression. *Unpublished working paper*, 2015.

[3] James J. Feigenbaum. A Machine Learning Approach to Census Record Linking. 2016.

[4] Ron Goeken, Lap Huynh, T. A. Lynch, and Rebecca Vick. New Methods of Census Record Linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44(1):7–14, January 2011.

[5] Michael D. Larsen. Hierarchical bayesian record linkage theory. *Iowa State University, Statistics*, 2005.

[6] Jason Long and Joseph Ferrie. Intergenerational occupational mobility in Great Britain and the United States since 1850. *The American Economic Review*, 103(4):1109–1137, 2013.

[7] Roy Mill and Luke CD Stein. Race, skin color, and economic outcomes in early twentieth-century America. 2016.

[8] Emily Nix and Nancy Qian. The Fluidity of Race:Passing in the United States, 1880-1940. Technical report, National Bureau of Economic Research, 2015.

[9] Rebecca Vick and Lap Huynh. The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44(1):15–24, January 2011.

[10] William E. Winkler. Matching and record linkage. *Business survey methods*, 1:355–384, 1995.

[11] Maria J. Wisselgren, S?ren Edvinsson, Mats Berggren, and Maria Larsson. Testing Methods of Record Linkage on Swedish Censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 47(3):138–151, July 2014.