How to code school names more efficiently: Common sense, scripting, and a novel SAS application

Akbar Akbari Esfahani[1], Matt Jans[1], Ninez Ponce[1]
Carl Ganz[1],
[1]UCLA Center for Health Policy Research, 10960 Wilshire Boulevard Suite 1550, Los Angeles, CA 90024

Abstract
For any survey, upcoding is an important, but often time consuming process. At the California Health Interview Survey (CHIS), child and teenage respondents are asked the name of the school they attend. The respondent-reported school names need to match the official names of the schools on the Department of Education (DOE) records for the data to be useful. Here we present an approach for matching school names using the SOUNDEX algorithm, along with a novel SAS application to minimize the amount of time spent manually coding.

Key Words: SOUNDEX, upcoding, natural language processing, data processing

1. Introduction & Motivation

Matching open-ended text against master lists can be an expensive, time-consuming, and error-prone process if done manually. Even automated matching can fail if the input data contain spelling errors and other irregularities, leaving a large remained to be matched by hand. However, with a little forethought, planning, testing, and one or two experienced programmers, significant productivity gains can be had. This paper reports on an automation innovation in the process of cleaning reported school names from the California Health Interview Survey (CHIS) and matching them to a master lists of schools. The procedures involves functions available in default SAS. We demonstrate the steps involved in matching, how we automated them, and the efficiency gains obtained.

1.1 The Opportunities and Problems of Self-feported School Names

Respondents in the CHIS child interview (conducted by proxy of a parent about a sampled child in the HH), the respondent reports the name of the school in which the sampled child is enrolled. We geocode the locations of these schools so researchers can conduct analyzes involving the distance between home and school (e.g., distance walked to school). This involves matching the self-reported school names to a master list of schools and their locations provided by the California

Department of Education (CA-DOE). This is a very unique and useful data product[1]

The self-reported data are often incomplete or incorrect in some way, making direct matching to the CA-DOE master list difficult or impossible without significant manual intervention. There are at least two ways in which the school information data can be incomplete or incorrect.

1) The parent provides only partial information (e.g., "I think it's Johnson School"), sometimes because they don't know or are unsure about the name of their student's school.

2) The parent provides complete information but it is incorrect in some way.

    1) School name is not worded exactly as it is on the CA-DOE master list
    2) School's common name is used rather than its formal name (e.g,. Johnson High v. Lyndon B. Johnson Technical High School)
    3) Simple misspellings or abbreviations due to either the respondent interviewer

All of these errors provide a roadblock to direct text matching. The simple solution too many researchers is human adjudication of mismatches, since human reasoning and intelligence, and brief familiarity with the lists and schooling system can easily adapt to these discrepancies. But the time and resource cost of this approach can expand quickly when the number of cases to resolve is beyond a few score.

Besides being time consuming, the matching process itself is prone to human error, which likely increases with fatigue inherent in a large job. Human matching is also not reproducible (at least not without significant documentation, and even them the replicability can be questionable), making errors or discrepancies hard to trace. The alternative to manually checking and matching that we present uses a basic natural language processing function in SAS called "SOUNDEX".

Our two overall goals for this proof-of-concept were to see how we could reduce:

1. Overall processing time, and
2. Human intervention/error?

1.1 SOUNDEX

Despite its association with SAS and data processing, The SOUNDEX algorithm was developed before the advent of the computer. It was developed in the 1918 [1] to help the U.S. Census Bureau deal with the range of last name spellings that are found in the population (e.g., Olson v. Olsen or Smith v. Smyth). The algorithm has also been popular for genealogical research for the same reason like with www.familysearch.com.

[1] See LINK for access to CHIS geocode data.

SOUNDEX has been programmed is SAS since SAS 6.07, and is part of the default SAS installation.

The purpose of SOUNDEX is to create a unique way to represent a set of homophones common to last names. SOUNDEX works by converting words into 4-character codes based on how they sound in English, thus correcting for slight spelling differences and mistakes. For example, using the rules in Table 1, the word "Washington" has a SOUNDEX value of "W252". Similarly, "Woshingto", "Wassington", "Wesington", and "Washingtin" also have a SOUNDEX of "W252", so any misspellings like this would be reduced to the same SOUNDEX value as the master list, making match possible. Because SOUNDEX only deals with the first four consonants, it can also correct abbreviations like "Wash'n". The algorithm works like this:

1) Keep the first letter
2) Ignore A, E, I, O, U, H, W, and Y after the first letter
3) Assign a numeric code from the "Code Number" column in Table 1 for the first four consonants
    a. Consonants are grouped by phonetic similarity

This makes dealing with surnames much more manageable considering the multitude of spellings of names that have similar or the same ethnic origin. For CHIS, the SOUNDEX algorithm allows for matching of respondent-reported school names with the CA-DOE list of schools even if the interviewer misspelled the name, or the respondent mispronounced the name.

Table 1.

Code Number	Represents Letters
1	B, F, P V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R
Disregard A, E, I, O, U, H, W, Y and repeating letters	

There are, of course, some misspellings that cannot be captured by SOUNDEX such as when the first letter of the word is wrong.

2. Application and Results (i.e., Our SAS Program)

Our SAS program was developed on California Health Interview Survey (CHIS) 2014 phone interview data from the sample child interview. The self-reported school names were provided to interviewers by a parent who reported about the sample child, and where hand-entered verbatim by the interviewer into an open-

text field. We also included sampled teens' school, which was reported by the teens themselves, but was also entered by the interviewer into an open-text field.

The CA-DOE master school name list was an annually-updated list of operating and closed public and private schools in California and their locations. The matching of self-reported school name to the CA-DOE master list was based on school name, county, and ZIP Code.

2.1 Eight matching stages

Table 2 describes the eight matching stages that defined our program, and the match rates for teen and child cases at each stage. The first three stages were conducted without conversion to SOUNDEX values. In the first stage, cases were considered matched to the CA-DOE list if the self-reported school name, county of residence, and zip code of residence were the same as a school on the DOE list. After each level of matching we remove matched observations and the remaining observations are passed on to the next level. Percentages in Table 2 are based on the entire sample submitted to the program for teen and child interviews separately.

Matching stage 2, based on cases that were not matched in stage 1, required self-reported school name and zip to match the CA-DOE list. Stage 3 required school name and county.

Up to this point we had successfully matched about 70% of teen responses and 68% of parent proxy responses for the child interview, so we converted school name on both lists (self-reported and CA-DOE) to SOUNDEX values and re-ran the match on the SOUNDEX values of school and alphabetic spelling of county (Stage 4), and then by just the SOUNDEX values for school (Stage 5) Stages 4 and 5 matched an additional 10% of teen cases and 9.4% of child cases. Stage 6 involved adding the word "School" to the self-reported school name, which matched a few more cases in each interview type.

Steps 7 and 8 used the cases that still could not be matched, and required manual intervention. Step 7 involved visually checking unmatched cases, and resolving them manually. Between step 7 and 8, 90% of teen remaining unmatched and This matched was about 95% of the teen cases and 97% of the child cases remaining unmatched. At the end of all matching steps, 1.0% of teen cases and 0.7% of child cases remained unmatched.

Table 2.

Stage	*Match based on…*	Teen	Child
1	School name, County, & ZIP	46.8%	48.1%
2	School name & ZIP	2.2%	0.6%
3	School Name & County	21.3%	19.1%
School name (on both lists) converted to SOUNDEX			
4	*SOUNDEX* <School>, County	3.9%	1.0%
5	*SOUNDEX* <School>	6.3%	8.4%
6	Add "School" to name	0.3%	0.7%
7	Visually check unmatched	16.1%	19.1%
8	Hand-coding unmatched cases	2.0%	2.3%
	Uncodeable after Stages 1-8	1.0%	0.7%
	Total (columns may not add to 100% due to rounding)	100.0%	100.0%
	n	1042	1502

2.2 SAS's SOUNDEX function

SAS (as well as Stata, SPSS, and R) offers a SOUNDEX function. It simply inputs a character (or character vector) and converts it to its SOUNDEX equivalent. We apply the SOUNDEX function to the respondent's school name as well as the DOE list.

We then apply the fourth level of matching which is by SOUNDEX school name and county. The next level is just by SOUNDEX school name. This matches about 10% of the respondents leaving us with about 20% left unmatched.

2.4 Quality Control

We create a "match quality" variable reflecting the degree of agreement among the matched cases across the multi-stage linking. In stage 1 matches received a score of 5 (i.e., School Name, County, and Zip code all matched based on raw data). This accounted for less than one percent of the teen and child samples.

Figure two shows the percentage of the remaining cases falling into match quality scores below five.

Table 3.

Match Quality	*Respondent data match DOE based on…*	Teen	Child
4	School + County + ZIP	64%	68%
3	School (or SOUNDEX) + County or ZIP	28%	21%
2	SOUNDEX <School>	6%	8%
1	Hand up-coded	2%	3%
	Total	100%	100%
	n	1032	1492

2.5 Efficiency Gains

Using information from our past, manual matching process and the process reported here, Figure 1 shows the number of cases that can be completed per hour. Note that the "Automated" match rate of 112 cases per minute includes all match steps. The auto-coded portion of that (i.e., the part requiring no human intervention) accounted for most of the matching (n = 1007 cases from stages 1-5) and could be run in about an hour.

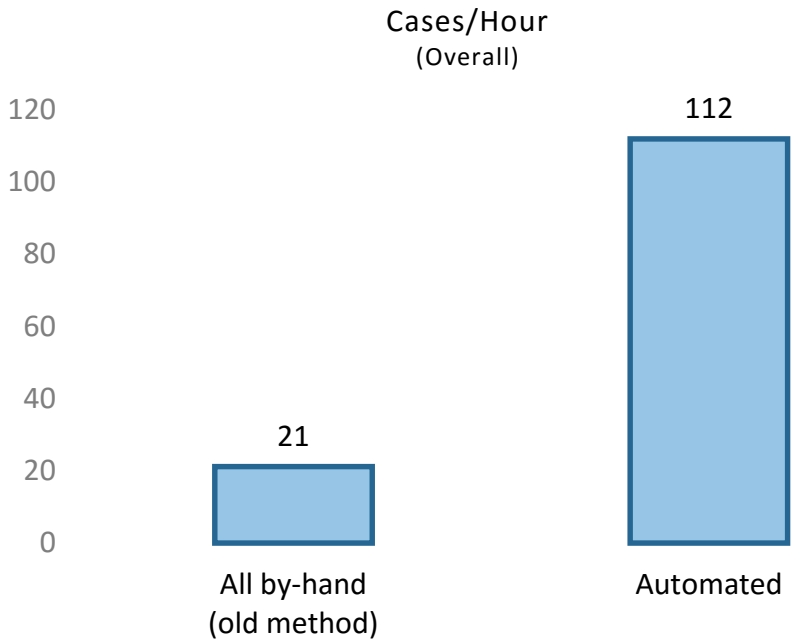

Figure 1. Cases/hour by hand-coding (old method) v. automated method

Figure 2 shows that overall hand-coding rate from past CHIS efforts (same as in Figure 1) compared to the hand-coded residual per-hour rate from the new method. As expected, coding the residual cases takes a little longer per hour because these are the "hardest of the hard." The gains in efficiency of the overall process displayed in Figure 1 remain clear.

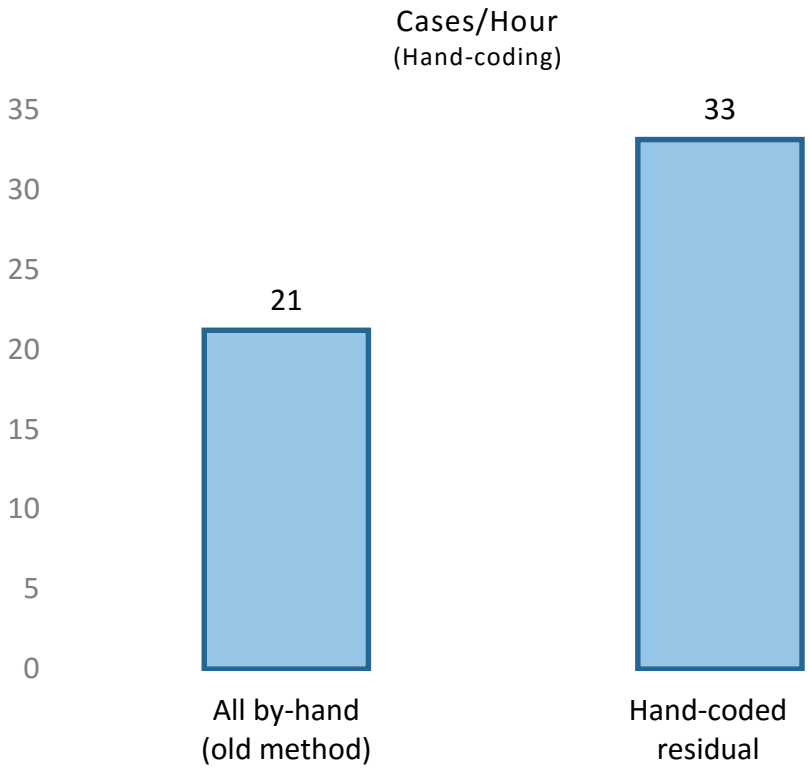

Figure 2. TITLE

3. Discussion and Future Developments

We demonstrated a modified and automated matching process for self-reported school names that turned a 3-4-week process into a process of about one day, with the majority of cases being matched automatically in about one hour. The setup cost of this program was about 3 days' time for an experienced programmer. The fixed startup costs are clearly covered by the overall efficiency of the matching process, and these are not costs that we will incur in future years using the same program. Further, the time saved matching means that we can make these data available to researchers much quicker than we could in the past.

Of course, this is just one possible way to match school names. We explored using a Levenshtein [2] distance to match at stage X, but we found that it did not create enough successful matches to be useful. This was because the unmatched cases where ones where whole words were missing from the school name, rather than misspellings.

One possible program improvement could be to finding more words that are frequently missing and applying them. For example, perhaps if first names are

often left off of self-reported school names (e.g., “Washington High School”), but are on the official CA-DOE list (e.g., “George Washington High School”) we should build that into the automated cleaning.

Another possible advance would be to use fuzzy linkage algorithms [3] instead of, or in addition to SOUNDEX. These may allow for wider variability in misspellings to match, and counter some of the limitations of SOUNDEX discussed above.

Finally, we are exploring the addition of an automated approach to expanding the DOE yearly database by augmenting it with any match that has quality of 5. This would allow the algorithm to “learn” from past matches and allow for the production of higher quality results than before.

Acknowledgements

California Health Interview Survey

References

[1] US patent 1435663, R. C. Russell, "(untitled)", issued 1922-11-14.
[2] Fellegi, I.; Sunter, A. (1969). “A Theory for Record Linkage.” Journal of the American Statistical Association 64 (328): pp. 1183-1210.
[3] Baeza-Yates R, Navarro G (June 1996). "A faster algorithm for approximate string matching". In Dan Hirchsberg, Gene Myers. Combinatorial Pattern Matching (CPM'96), LNCS 1075. Irvine, CA. pp. 1–23.