Comparing Multiple SAS® Functions for Text Field Matching in Data Linkage:
SOUNDEX, NYSIIS, COMPGED
Yu Sun and Cordell Golden
National Center for Health Statistics, Centers for Disease Control and Prevention

Background:

NCHS Data Linkage Program
- Links survey data with vital and administrative records
- Designed to maximize the scientific value of the NCHS population-based surveys

Motivation:

- Previous linkage algorithms relied heavily on 9 digits of Social Security Number (SSN9)
- Current algorithms are more dependent on name variables due to changes in the way personally identifiable information is collected, only last 4 digits of SSN (SSN4)
- This analysis will assess the value added by incorporating phonetic algorithms and string comparator functions for text field matching in SAS® rather than exact matches

Data Sources:

National Health Interview Survey (NHIS)
Nationally representative, cross-sectional household interview survey conducted by NCHS that serves as an important source of information on the health of the civilian, noninstitutionalized population in the U.S.

Housing and Urban Development (HUD)
Agency responsible for overseeing and managing domestic housing programs and policies, including specialized programs for high-needs populations (e.g., the elderly, homeless, and disabled) in the U.S.
Methods

NCHS Survey Data

Eligible for Linkage

Linked 1999-2006 NHIS-HUD Data
First and Last name present and matched exactly on:
SSN9, Date of birth (M/Y), Sex

Ineligible for Linkage

HUD Administrative Data

Matched exactly on:
• First name
• Last name

Soundex / NYSIIS / COMPGED agreement on:
• First name
• Last name

Goal: Compare the matches when requiring an exact match vs. using phonetic and string comparators

Code used for phonetic and string comparison of First (FN) and Last names (LN):

**SOUNDEX**
SDX=1: SOUNDEX(FN1)=SOUNDEX(FN2) and SOUNDEX(LN1)=SOUNDEX(LN2);

**NYSIIS**
NYS=1: %NYSIIS(name, name_NYS); FN1_NYS=FN2_NYS and LN1_NYS=LN2_NYS;

**COMPGED**
Ged=1: (COMPGED(FN1, FN2)<=100 or COMPGED(FN2,FN1)<=100) and (COMPGED(LN1, LN2)<=100 or COMPGED(LN2,LN1)<=100) ;

x_ged=COMPGED(x, y); y_ged=COMPGED(y, x);
Results

<table>
<thead>
<tr>
<th>Exact match</th>
<th>GED</th>
<th>SDX</th>
<th>NYS</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13,188</td>
<td>86.62</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>910</td>
<td>5.98</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>396</td>
<td>2.60</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>47</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>455</td>
<td>2.99</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>129</td>
<td>0.85</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>81</td>
<td>0.53</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>19</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Value added: 13.4% (n=2,037) additional matches were captured with the phonetic and string comparators

Of the 2,037 additional matches:
44.7% were captured using by all 3 functions
45.7% were captured by SOUNDEX and/or COMPGED
9.6% were captured by NYSIIS only or a combination of NYSIIS and either SOUNDEX or COMPGED

An example of name matched by **NYSIIS only**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x_SDX</th>
<th>y_SDX</th>
<th>x_GED</th>
<th>y_GED</th>
<th>x_NYS</th>
<th>y_NYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catherine</td>
<td>Katherine</td>
<td>C365</td>
<td>K365</td>
<td>200</td>
<td>200</td>
<td>CATARAN</td>
<td>CATARAN</td>
</tr>
</tbody>
</table>

An example of name matched by **SOUNDEX only**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x_SDX</th>
<th>y_SDX</th>
<th>x_GED</th>
<th>y_GED</th>
<th>x_NYS</th>
<th>y_NYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maricela</td>
<td>Marisella</td>
<td>M624</td>
<td>M624</td>
<td>120</td>
<td>120</td>
<td>MARACAL</td>
<td>MARASAL</td>
</tr>
</tbody>
</table>

Example of names matched by **COMPGED only**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x_SDX</th>
<th>y_SDX</th>
<th>x_GED</th>
<th>y_GED</th>
<th>x_NYS</th>
<th>y_NYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hermon-Sisco</td>
<td>Hermon</td>
<td>H65522</td>
<td>H655</td>
<td>280</td>
<td>60</td>
<td>HARNAN-SASC</td>
<td>HARNAN</td>
</tr>
<tr>
<td>Pat</td>
<td>Patricia</td>
<td>P3</td>
<td>P362</td>
<td>50</td>
<td>250</td>
<td>PAT</td>
<td>PATRAC</td>
</tr>
<tr>
<td>Zheng</td>
<td>Zhen</td>
<td>Z52</td>
<td>Z5</td>
<td>50</td>
<td>10</td>
<td>ZANG</td>
<td>ZAN</td>
</tr>
</tbody>
</table>

An example of name matched by **ALL** three functions

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x_SDX</th>
<th>y_SDX</th>
<th>x_GED</th>
<th>y_GED</th>
<th>x_NYS</th>
<th>y_NYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brian</td>
<td>Bryan</td>
<td>B65</td>
<td>B65</td>
<td>100</td>
<td>100</td>
<td>BRAN</td>
<td>BRAN</td>
</tr>
</tbody>
</table>
Conclusions

• All three functions have their unique strengths and were able to identify matches not picked up when an exact match on name was required
  – SOUNDEX and NYSIIS are good in matching names that sound alike and are spelled similarly
  – NYSIIS accounts for differences in the first letter, but SOUNDEX does not
  – COMPGED was better at handling multi-part last names; abbreviations and nicknames; and ethnic and non-traditional spelling variations

• A combination of all three functions appears to work best

• NCHS will continue to research other name comparison functions and algorithms (e.g. Jaro-Winkler, SPEDIS, Perl) for data linkages using text fields
REFERENCES

COMPGED Function

SOUNDEX Function

NYSIIS
http://www.dropby.com/NYSIIS.html


ACKNOWLEDGEMENTS

The authors gratefully acknowledge Lisa Mirel, Eileen Call, Jim Brittain and the Special Projects Branch for their contributions to this work.

CONTACT

NCHS Data Linkage Program: datalinkage@cdc.gov