

Record-Matching of Names and Addresses Using SAS®

Thomas J. Winn, Jr.

Texas State Auditor's Office, Austin, Texas

[SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.]

Abstract

The paper describes the SAS code used that was written by the author for use at the Texas State Auditor's Office, for matching records from two files on the basis of names and addresses only.

Introduction

Before using more advanced data analysis techniques, frequently we need to combine information from several files, for use by auditors, investigators, classification personnel, or management. This usually is easy, if each of the files contains some "key data" -- columns such as vendor numbers, social security numbers, employer identification numbers, taxpayer numbers, etc. -- which are contained in both files, and can be used as a basis for merging the data.

However, let us suppose that one or more of the files we need to combine with the others does not contain the key data in a crisply comparable format. Specifically, what if the only common link is name and address information? Now, the problem with names and addresses is that they can be represented in several alternative ways, some of which may be readily recognizable as equivalent, while some of the other representations may not appear to be equivalent.

Here is an example:

Mr. Thomas J. Winn II
10906 Rustic Manor Lane
Austin, TX 78750-1134

Tom Winn, M.S.
Rustic Manor Ln, #10906.
Austin, Texas 78750

Winn Jr., Thom
10906 Rustic Manor Rd.
Austin, TX 78750

Winn, Thomas John, Jr.
Rustic Manor Lane, Box 10906
Austin, Texas 78750

Thos. Winn II
10906 Rustic Manor Ln.
Austin, TX 78750-1134

T. J. Winn
10906 Rustic Manor
Austin, Texas 78701

All of the names and addresses shown in this example represent the same individual. Humans are able to recognize the equivalence of those names and addresses fairly readily. Can we instruct a computer to identify their equivalence? If so, how?

On Recognizing "Sameness"

In general, computer software is able to determine when two data values are identical, but it isn't able to discern when two data values are only more-or-less the same. Can anything be done to facilitate the handling of those "slippery" names and addresses? How are we going to deal with this kind of conceptual vagueness and imprecision?

In traditional logic, the Law of the Excluded Middle says that every proposition must either be True or False (0 or 1). Instead of giving up because of the limitations of systems that are based on two-valued logic, we turn to concepts that are grounded in what is known as "fuzzy logic." In fuzzy logic, we extend our logical systems to embrace the concept of "possible" -- and we use measures that operate within the range of real numbers from 0 to 1. Now, what does this have to do with handling names and addresses?

A Solution for Matching Names & Addresses

Toward the end of 2001, and in the summer of 2002, (as a spare-time activity) The author completed the initial development of three SAS programs for matching records in two files on the basis of names and addresses only. The approach is an implementation of fuzzy logic. It follows some of the elements of an outline that was described in a paper by Charles Patridge (*Observations*, #15, Nov. 1998).

The first program breaks the names and addresses into pieces, and then standardizes the information contained in those pieces. The second program uses scoring methods to identify some possible matches. The third program tries a couple of "methods of last resort" on name components only, for the non-matching pairs from the second program.

The standardization program uses SAS functions to perform the following data cleaning steps:

- it eliminates these special characters
'!@#%&*'()_+=[]{}~\|<>.?/~
- it converts everything to uppercase characters,
- it identifies certain titles and suffixes, and it removes them from the other components (for example, Mr., Ms., Dr., Maj., Adm., M.D., Ph.D., Jr., etc.),
- it separates the remaining name- and address-elements into words,
- it replaces name-components which look like first-names with their diminutives (for example, Thomas to Tom, Elizabeth to Liz, Alfred to Al and Fred, etc.),
- it replaces certain words which are used in addresses and business organizational names with abbreviations (for example, Company to Co., Incorporated to Inc., etc.).

The fuzzy duplicates program compares observations in two files, whose name- and address-elements have been standardized, and it attempts to identify those records from each file which appear to represent the

same individuals or entities. It performs the following steps:

- it constructs the Cartesian Product of the two standardized files (all possible pairs of records from the two files), excluding those pairs that are from different areas,
- for every pair, evaluate each name- and address-item from the first standardized file, looking for matching items among the components of the second standardized file in the pair, and count the number and the proportion of matching items,
- identify those pairs which have mostly-matching combined name and address items, mostly-matching name items but few matching address items, and mostly-matching address items but few matching name items.

The "methods of last resort" program performs the following steps on the name items (excluding any titles or suffixes) of the non-matching pairs from the fuzzy duplicates program:

- it looks for matching name components, where the vowels have been removed,
- it looks for name components that sound somewhat alike, even though they may be spelled differently,
- it looks for name components that are spelled similarly.

An Example

The Appendix contains an illustrative example. Two SAS files were used in the example:

- The first file contains some genuine, and some bogus, information pertaining to several individuals, many of which are employees of the State Auditor's Office.
- The second file contains names and addresses extracted from the personnel database for all employees of the State Auditor's Office, as of 12/5/2001.

The first screen shows the SAS data set "Fuzzy_test_data", before standardization. The second screen shows one of the results of standardization of that file: one record contained a zip code that was not consistent with the state postal abbreviation in the input data file. The third screen shows a few of the records from the SAS data file, "Agency308".

After standardization of the names and address components in both files, the fuzzy duplicates program is run. The fourth and fifth screens show the main result from the fuzzy duplicates program, the records from each file which have "mostly-matching" combined name and address items.

Some of the records from the two files that DO NOT have "mostly-matching", COMBINED name and address items, are then evaluated to see if there are some possible matches of the names or addresses, when these components are evaluated separately. The results are in the sixth and seventh screens.

Finally, the non-matching pairs from the fuzzy duplicates program are evaluated for possible matches using the "methods of last resort" program. The results from those comparisons are in the eighth screen. At first glance, this may not appear to be a particularly fruitful approach. The possible matches from this program would need to be evaluated carefully – but a careful inspection will find that this screen does contain one actual match!

Preliminary Results and Intentions

The methods are not "bullet-proof", but they do seem to work pretty well with test data. It is anticipated that enhancements to the programs will be added, as more knowledge and experience concerning this approach is acquired.

Conclusion

This paper describes three SAS programs that are used to perform record-matching operations between two files, on the basis of names and addresses only.

My SAS code is available upon request.

Suggestions for Further Reading

- Bahler, Caroline (2001). "Data Cleaning and Base SAS Functions," SUGI 26 Proceedings, Paper 56.
- Cody, Ronald (2000), "Data Cleaning 101," SCSUG 2000 Proceedings, pp. 44-56.
- Gershteyn, Yefim (2000), "Use of SPEDIS Function in Finding Specific Values", SUGI 25 Proceedings, Paper 86.
- Hermansen, Sigurd (2002), "Searching Character Fields for Nonexact Matches", posted on SAS-L on 31 July 2002.
- Hermansen, Sigurd, "TIP 00276", "Illustrate fuzzy linkage/matching using a highly simplified scoring method", in "SAS Coding Tips and Techniques" in <http://www.sconsig.com/>.
- Partridge, Charles (1998). "The Fuzzy Feeling SAS Provides: Electronic Matching of Records without Common Keys," Observations – The Technical Journal for SAS Users, Vol. 15, November 1998. <http://www.sas.com/service/library/periodicals/obs/obswww15/index.html>

Author Information

Tom Winn, Senior Systems Analyst
Texas State Auditor's Office
P.O. Box 12067
Austin, Texas 78711-2067

Telephone: 512 / 936-9735
E-Mail: twinn@sao.state.tx.us

Appendix – An Example

The SAS Data Set, “Fuzzy_test_data”

ID	Name	Address	City	State	ZIP Code
1	Mr. Thomas J. Winn, Jr.	10906 Rustic Manor Lane	Austin	TX	78750-1134
2	Winn, Thomas John, II	10906 Rustic Manor Ln	Austin	TX	78750
3	Tom Winn	10906 Rustic Manor Ln	Austin	TX	78750
4	Thos. Winn, II	Rustic Manor Ln	Austin	TX	78750
5	Mrs. DeAnne R. Winn	Rustic Manor Lane	Austin	TX	78750-1134
6	De Anne Hart-Winn	10906 Rustic Manor Ln	Austin	TX	78750
7	Dee Winn	10906 Rustic Manor	Austin	TX	78750
8	Ernest D. Cuellar, C.P.A.	3602 Spotted Horse Trail	Austin	TX	78739
9	Cuellar, Ernesto	3602 Spotted Horse Trl	Austin	TX	78739
10	Ernie Cuellar	3602 Spotted Horse Tr.	Austin	TX	78739
11	Olin J. Davis, M.B.A.	P.O. Box 11643	Austin	TX	78711
12	Major D. J. Davis	Box 11643	Austin	TX	78711
13	Offie Davis	PO Box 11643	Austin	TX	78711
14	Katy L. Cossy	504 Dry Creek Court	Round Rock	TX	78681
15	Dr. Kirby Cossy, Ph.D.	504 Dry Creek Ct	Round Rock	TX	78681
16	Mrs. Joyce D. Inman	10526 Jollyville Rd.	Austin	TX	78758
17	Inman, Joy	10526 Jollyville Rd., #1406	Austin	TX	78758
18	Mr. Dorvin D. Hendrick	4010 Halfway Cove	Round Rock	TX	78681
19	D. Hendrick, Esq.	4010 Halfway Cv	Round Rock	TX	78681
20	David Baggett	SAS Institute Inc.	Cary	NC	27513
21	Dave Baggett	The SAS Institute, User Grp Support	Cary	NC	27513-2414
22	Ms. Thomas Winn	17910 28th Avenue North	Plymouth	MN	55447
23	Mr. Thomas Winn, III	Excellence, Incorporated	Plymouth	MN	55447-1658
24	Miss Colleen O'Reilly	12345 Main Street	Newchplace	OK	00000

The SAS function, ZIPSTATE is used to identify certain inconsistent postal abbreviations.

Records in the Input Data File, Fuzzy.Fuzzy_Test_Data,
Having a State Postal Abbreviation that is Inconsistent with the Given Zipcode

11:11 Friday, December 28, 2001

ID_Num	St	ZIPCODE	State
24	OK	00000	

The SAS Data Set, "Agency308" (a few observations)

VIEWTABLE: Fuzzy_Agency308							
Obs	idnum_x	name_x	addr1_x	addr2_x	city_x	st_x	zipcd_x
1	1	ADAIR, SUZANNE R	11519 ANTIQUA		AUSTIN	TX	78799 5122566192
2	2	ADAMS, GREGORY S	201 N KINGS CANYON		CEDAR PARK	TX	78613 5122561927
3	3	ALMARAZ, RODNEY A	8718 COASTAL DR.		AUSTIN	TX	78749 5126896553
4	4	ALVARADO, HELEN S	1306 LAURETTA WOOD DRIVE		CEDAR PARK	TX	78613 5122482757
5	5	ALWIN, LAWRENCE F	510 BENDING OAK		DRIPPING SPRINGS	TX	78620 0
6	6	AMBRES-WADE, COURTNEY R	2016 CHARLOTTE WAY		ROUND ROCK	TX	78684 5126718424
7	7	ASHTON, LESLIE P	4911 WESTFIELD		AUSTIN	TX	78731 5124511814
8	8	AVILA, MARY A	3808 WINFIELD DRIVE		AUSTIN	TX	78704 5124429395
9	9	BAKER, EILEEN M	2908 ANN ARBOR		AUSTIN	TX	78704 5124412788
10	10	BALLMAN, LYNNE E	8032 ABILENE TRAIL		AUSTIN	TX	78749 5122989889
11	11	BARBOZA, ILEANA	6711 CARISBROOKE LANE		AUSTIN	TX	78754 5122781782
12	12	BECKHAM, MARGENE THORNTON	2105 PEASE ROAD		AUSTIN	TX	78703 5124781910
13	13	BEDNARSKI III, FRED	8909 PORTAGE COVE		AUSTIN	TX	78717 5122184188
14	14	BIEBERDORF, SHANNON LYNN	2701 MARIA ANNA		AUSTIN	TX	78703 5124520788
15	15	BOLLINGER, ROBERT H	PO BOX 860067		AUSTIN	TX	78765 5124522567
16	16	BOSTON, EMERY S	1707 DAYFLOWER TRACE		CEDAR PARK	TX	78613 5123361404
17	17	BOSTON, NATALIA Y	1707 DAYFLOWER TRACE		CEDAR PARK	TX	78613 5123361404
18	18	BRADLEY, KIMBERLY M	1605 LANTERN LIGHT DRIVE		ROUND ROCK	TX	78681 5122446130
19	19	BRADSHAW, BRENDA A	2814 AFTONSHIRE WAY, #19-205		AUSTIN	TX	78748 5122825469
20	20	BRANTLEY, SHARON L	2801 WELLS BRANCH PKWY, #316		AUSTIN	TX	78728 5122361600
21	21	BROWN, DOUGLAS C	511 EAST 38 1/2 STREET		AUSTIN	TX	78751 5124528948
22	22	BROWN, TIMOTHY R	2800 STECK AVE #11		AUSTIN	TX	78757 5123024125
23	23	BUECHLEY, PAIGE AMBER	1122 HOLLOW CREEK DRIVE #4		AUSTIN	TX	78704 5124288495
24	24	BURFORD, ADRIANA	4106 CAT MOUNTAIN DRIVE		AUSTIN	TX	78731 5123432175
25	25	BUSHNELL, DENNIS R	P.O. BOX 96		JOHNSON CITY	TX	78636 8309682330
26	26	CANNON, RICHARD D	1306 PIGEON VIEW STREET		ROUND ROCK	TX	78664 5123384019
27	27	CANTU, LUCINDA	3400 DUNLEIGH		AUSTIN	TX	78745 5122802111
28	28	CARDINAL, MARY A	7613 YAUPON DRIVE		AUSTIN	TX	78759 5122567650
29	29	CARMICHAEL, VIRGINIA	1620 WEST 9 1/2 STREET		AUSTIN	TX	78703 5124578557
30	30	CARMICHAEL, VIRGINIA	2014 CARTER STREET		AUSTIN	TX	78707 5127006106

Some "Mostly Matching" Combined Name & Address Items, between Fuzzy_test_data and Agency308

Selected Results From the Matching of Name & Address Components							
From the SAS File Fuzzy_test_data With the SAS File Agency308							
*** Records with 'Mostly-Matching' Combined Name and Address Items ***							
11:04 Friday, December 28, 2001							
Obs	idnum_x	name_x	addr1_x	addr2_x	city_x	st_x	zipcd_x
1	1	Mr. Thomas J. Winn, Jr.	10906 Rustic Manor Lane		Austin	TX	78750-1134
2	2	Winn, Thomas John, II	10906 Rustic Manor Ln		Austin	TX	78750
3	3	Tom Winn	10906 Rustic Manor Ln		Austin	TX	78750
4	4	Thos. Winn, II	Rustic Manor Ln	Box 10906	Austin	TX	78750
5	5	Mrs. Deanne R. Winn	Rustic Manor Lane	*10906	Austin	TX	78750-1134
6	6	De Anne Hart-Winn	10906 Rustic Manor Ln		Austin	TX	78750
7	7	Dee Winn	10906 Rustic Manor Ln		Austin	TX	78750
8	8	Ernest D. Cuellar, C.P.A.	3602 Spotted Horse Trail		Austin	TX	78739
9	9	Cuellar, Ernesto	3602 Spotted Horse Trl.		Austin	TX	78739
10	10	Ernie Cuellar	3602 Spotted Horse Tr.		Austin	TX	78739
11	14	Kirby L. Cossey	504 Dry Creek Court		Round Rock	TX	78681
12	15	Dr. Kirby Cossey, Ph.D.	504 Dry Creek Ct		Round Rock	TX	78681
13	16	Ms. Joyce D. Inman	10926 Jollyville Rd.	*1406	Austin	TX	78759
14	17	Inman, Joy	10926 Jollyville Rd., *1406		Austin	TX	78759
15	18	Mr. Dorvin D. Hendrick	4010 Halfway Cove		Round Rock	TX	78681
Obs	idnum_x	name_x	addr1_x	addr2_x	city_x	st_x	zipcd_x
1	240	WINN JR, THOMAS JOHN	10906 RUSTIC MANOR LN		AUSTIN	TX	78750
2	240	WINN JR, THOMAS JOHN	10906 RUSTIC MANOR LN		AUSTIN	TX	78750
3	240	WINN JR, THOMAS JOHN	10906 RUSTIC MANOR LN		AUSTIN	TX	78750
4	240	WINN JR, THOMAS JOHN	10906 RUSTIC MANOR LN		AUSTIN	TX	78750
5	240	WINN JR, THOMAS JOHN	10906 RUSTIC MANOR LN		AUSTIN	TX	78750
6	240	WINN JR, THOMAS JOHN	10906 RUSTIC MANOR LN		AUSTIN	TX	78750
7	240	WINN JR, THOMAS JOHN	10906 RUSTIC MANOR LN		AUSTIN	TX	78750
8	42	CUELLAR, ERNEST D	3602 SPOTTED HORSE TRAIL		AUSTIN	TX	78739
9	42	CUELLAR, ERNEST D	3602 SPOTTED HORSE TRAIL		AUSTIN	TX	78739
10	42	CUELLAR, ERNEST D	3602 SPOTTED HORSE TRAIL		AUSTIN	TX	78739
11	38	COSSEY, KIRBY L	504 DRY CREEK CT		ROUND ROCK	TX	78681
12	38	COSSEY, KIRBY L	504 DRY CREEK CT		ROUND ROCK	TX	78681
13	106	INMAN, JOYCE D	10926 JOLLYVILLE ROAD *1406		AUSTIN	TX	78759
14	106	INMAN, JOYCE D	10926 JOLLYVILLE ROAD *1406		AUSTIN	TX	78759
15	87	HENDRICK, DORVIN D	4010 HALFWAY COVE		ROUND ROCK	TX	78681

More "Mostly Matching" Combined Name & Address Items, between Fuzzy_test_data and Agency308

SAS

Output (Unloaded)

Selected Results From the Matching of Name & Address Components
From the SAS File Fuzzy_test_data With the SAS File Agency308
*** Records with 'Mostly-Matching' Combined Name and Address Items ***
11:04 Friday, December 28, 2001

Obs	idnum_x	name_x	addr1_x	addr2_x	city_x	st_x	zipcd_x
16	19	D. Hendrick, Esq.	4010 Halfway Cv		Round Rock	TX	78681

Obs	idnum_z	name_z	addr1_z	addr2_z	city_z	st_z	zipcd_z
16	87	HENDRICK, DARVIN D	4010 HALFWAY COVE		ROUND ROCK	TX	78681

SAS

Records with "Mostly-Matching" Name Items, but few Matching Address Items.

SAS

Output (Unloaded)

Selected Results From the Matching of Name & Address Components
From the SAS File Fuzzy_test_data With the SAS File Agency308
*** Records with 'Mostly-Matching' Name Items, But Few Matching Address Items ***
11:04 Friday, December 28, 2001

Obs	idnum_x	name_x	addr1_x	addr2_x	city_x	st_x	zipcd_x	idnum_z
1	1	Mr. Thomas J. Winn, Jr.	16906 Rustic Manor Lane		Austin	TX	78750-1134	201
2	11	Olin J. Davis, W.D.A.	P.O. Box 11643		Austin	TX	78711	46
3	12	Major O. J. Davis	Box 11643		Austin	TX	78711	46

Obs	name_z	addr1_z	addr2_z	city_z	st_z	zipcd_z
1	SHINDELL, THOMAS J	12345 LANPLIGHT VILLAGE AVE, #328		AUSTIN	TX	78759
2	DAVIS, OLIN J	6706 B HART LANE		AUSTIN	TX	78731
3	DAVIS, OLIN J	6706 B HART LANE		AUSTIN	TX	78731

SAS

Records with "Mostly-Matching" Address Items, but few matching Name Items

SAS

Output (Untitled)

Selected Results From the Matching of Name & Address Components
From the SAS File Fuzzy_test_data With the SAS File Agency308
*** Records with 'Mostly-Matching' Address Items, But Few Matching Name Items ***
11:04 Friday, December 28, 2001

Obs	idnum_x	name_x	addr1_x	addr2_x	city_x	st_x	zipcd_x
1	11	Olin J. Davis, M.B.A.	P.O. Box 11648		Austin	TX	78711
2	12	Major D. J. Davis	Box 11648		Austin	TX	78711
3	13	Ollie Davis	PO Box 11648		Austin	TX	78711
4	16	Ms. Joyce D. Inman	10926 Jollyville Rd.	*1406	Austin	TX	78759
5	16	Ms. Joyce D. Inman	10926 Jollyville Rd.	*1406	Austin	TX	78759
6	17	Inman, Joy	10926 Jollyville Rd., *1406		Austin	TX	78759
7	17	Inman, Joy	10926 Jollyville Rd., *1406		Austin	TX	78759

Obs	idnum_z	name_z	addr1_z	addr2_z	city_z	st_z	zipcd_z
1	203	SIMON, MICHAEL A	P.O. BOX 12682		AUSTIN	TX	78711
2	203	SIMON, MICHAEL A	P.O. BOX 12682		AUSTIN	TX	78711
3	203	SIMON, MICHAEL A	P.O. BOX 12682		AUSTIN	TX	78711
4	115	KELLY, NATASHA A	10926 JOLLYVILLE ROAD, APT *617		AUSTIN	TX	78759
5	189	RUSSELL, LAURIE F	9500 JOLLYVILLE ROAD *261		AUSTIN	TX	78759
6	115	KELLY, NATASHA A	10926 JOLLYVILLE ROAD, APT *617		AUSTIN	TX	78759
7	189	RUSSELL, LAURIE F	9500 JOLLYVILLE ROAD *261		AUSTIN	TX	78759

SAS

Fuzzy Name-Matching Using the "Methods of Last Resort"

SAS

Output (Untitled)

Fuzzy Name-Matching Using the 'Methods of Last Resort'
For Non-Matching Pairs from the Fuzzy Duplicates Program,
Using the SAS File Fuzzy_test_data With the SAS File Agency308
15:17 Monday, August 12, 2002

i	d	t	n	s	t	i	t	n	s	t	a	a	s
u	i	a	e	f	e	u	i	e	x	s	c	c	c
-	-	-	-	-	-	-	-	-	-	-	-	-	-
6			DE ANNE HART WINN	4	1			ADAIR SUZANNE R	3	2	0		0.10000
6			DE ANNE HART WINN	4	14			BIEBERDORF SHANNON LYNN	3	2	0		0.10000
6			DE ANNE HART WINN	4	32			CHANDORNEY LISA A	3	2	0		0.13750
6			DE ANNE HART WINN	4	33			CHAVEZ ANTHONY D	3	1	2		0.26250
6			DE ANNE HART WINN	4	90			HARTMAN DEBORAH J	3	2	2		0.20000
6			DE ANNE HART WINN	4	141			MCBRIDE NANCY LEE	3	2	0		0.10000
6			DE ANNE HART WINN	4	180			QUINN FLOYD F	3	2	0		0.10000
6			DE ANNE HART WINN	4	235			WIEDERHOLD JENNIFER R	3	2	0		0.10000
11			OLIN J DAVIS	MBA	3	45		DAVIS FLEMING JOSEPH J	4	2	2		0.70000
11			OLIN J DAVIS	MBA	3	72		FRALIN JOE K	3	2	1		0.28500
11			OLIN J DAVIS	MBA	3	119		KILLINGSWORTH JAMES S	3	2	1		0.16667
11			OLIN J DAVIS	MBA	3	222		WALANDO JUDITH A	3	2	1		0.11667
12	NAJ		O J DAVIS		3	45		DAVIS FLEMING JOSEPH J	4	2	2		0.70000
12	NAJ		O J DAVIS		3	161		OLIPHANT JEMMY	2	1	2		0.16667
13			OLLIE DAVIS		2	46		DAVIS OLIN J	3	1	2		0.69000
14			KIRBY L COSSEY		3	31		CASAS SYLVIA G	3	2	1		0.10000
16	MS		JOYCE D INMAN		3	45		DAVIS FLEMING JOSEPH J	4	1	2		0.16667

SAS