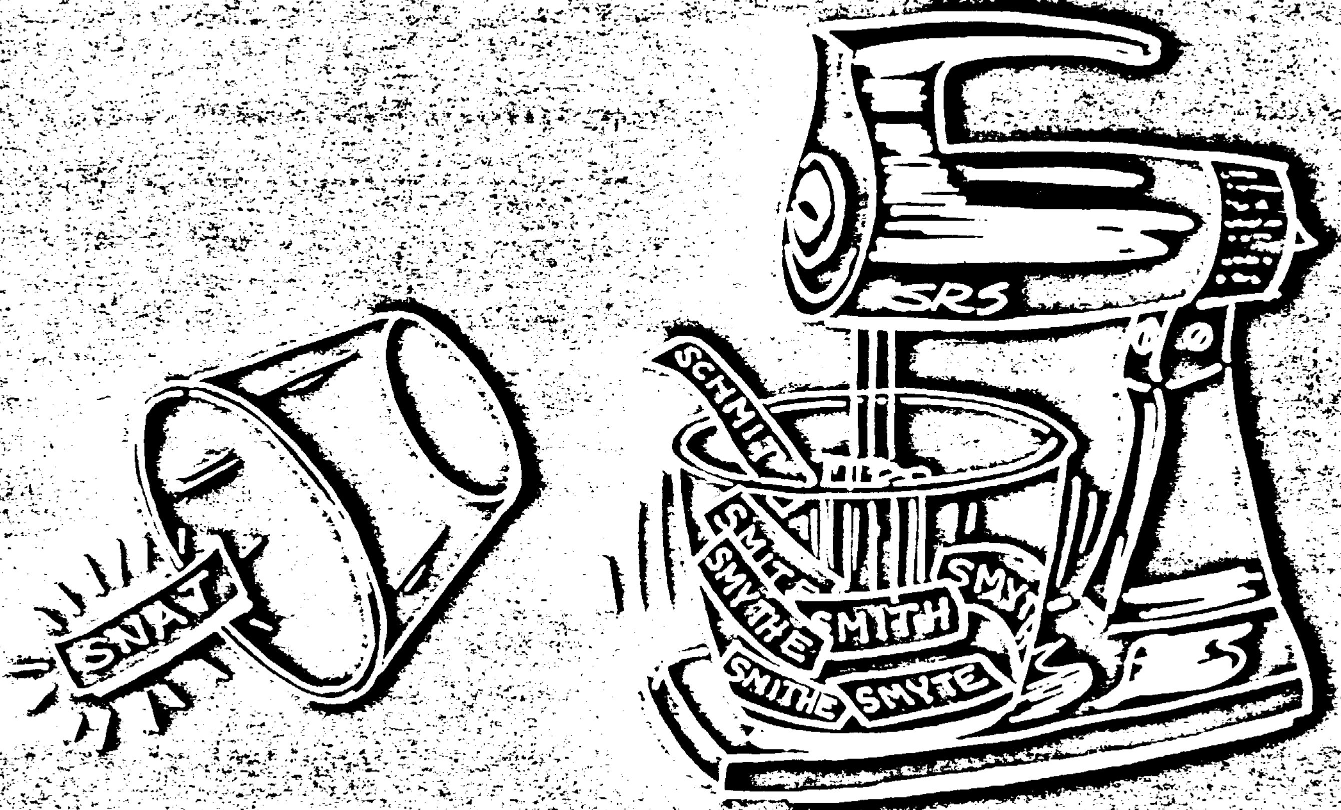
Power

SELECTION OF A SURNAME CODING PROCEDURE FOR THE SRS RECORD LINKAGE SYSTEM



Sample Survey Research Branch

Research Division

Statistical Reporting Service

U. S. Department of Agriculture

Washington, D. C.

February 1977

SELECTION OF A SURNAME CODING PROCEDURE FOR THE SRS RECORD LINKAGE SYSTEM

by

BILLY T. LYNCH

WILLIAM L. ARENDS

TABLE OF CONTENTS

SELECTION OF A SURNAME CODING PROCEDURE FOR THE SRS RECORD LINKAGE SYSTEM

INTRODUCTION

The Statistical Reporting Service (SRS) is developing a record linkage system to create a master list sampling frame of farm operators in each State. All samples for probability and non-probability surveys conducted by each State Statistical Office (SSO) will be selected from this list. This system uses a probability model which incorporates some of the theoretical concepts developed by Ivan P. Fellegi and Alan B. Sunter.[1/] Implicit in the development of their theory is the assumption that if two files are linked then all possible comparisons of all the records of both files will be attempted. However, SRS is really dealing with the "one super file" unduplication problem. That is, different files have been combined into one composite file. The ideal situation in this case is still to make all possible pairwise comparisons. It is clear that even for medium-sized files the number of comparisons under this assumption would be very large, (e.g. 10^5 records in each file would imply $\begin{bmatrix}10^5\\2\end{bmatrix}$ comparisons). Some technique has to be used to reduce these comparisons to a more manageable number. In order to reduce the number of comparisons, it is necessary to carry out the comparisons only within specified sub-groups of the file. These sub-groups should be organized so that the proportion of duplicate records within the sub-groups is maximized with respect to the proportion of undetected duplication. In other words, these sub-groups should be formed to maximize the possibility of determining duplication. Therefore, the files have to be "blocked" in some fashion and comparisons made only within corresponding blocks. A block, therefore, can be defined as a group of records which has a high likelihood of containing duplicates.

The variables that are used to create these blocks must meet the following criteria:

1. They must be present in all records in the file (Ubiquity).
2. They should be a permanent form of identification (Permanence).
3. They must be recorded with a high degree of accuracy (Reliability).

Among the variables which are present in agricultural files, the surname comes closest to satisfying the preceding requirements:

1. The surname is present in all records.
2. Among members of the farming population, it is subject to very little change as a form of permanent identification.
3. The surname is subject to minor recording errors, but it is still more reliable than other variables present.

[1/] Fellegi, Ivan P. and Sunter, Alan B., "A Theory for Record Linkage", Journal of American Statistical Association, pp. 1183-1210, December 1969.

OBJECTIVES

The primary objective of this research project was to select the best surname coding technique that could be used to create linkage blocks for the SRS System. The "best" surname coding technique can be defined as one which:

1. Places all variations of a given surname in the same code.
2. Limits the size of the codes; that is limiting the number of records assigned a given code.
3. Creates codes that contain few dissimilar surnames.
4. Requires minimal processing time and costs.

PROCEDURES

The research project was conducted in two phases. In the first phase, a sample of individual names was selected from four files containing names from eight different states. In the second phase, a complete list of individual names from a file supplied by a single state was analyzed. By conducting the analysis in two phases, we were able to evaluate the coding techniques on surnames from a cross section of states, and then evaluate the techniques in one state. Table 1 displays the size of the files and the states contained in each file.

The sample for the first phase of the analysis consisted of the 18,830 names with the employer identification number (EIN) plus a systematic sample drawn with sampling rates of 1/2 from file 2, 1/3 from file 3, and 1/2 from file 4. This produced a total sample of 250,431 names.

TABLE 1 FILE SIZE OF TEST STATES

File	States	No. With Employer Identification Number	Other	Sample Size
1	Kentucky	3,475	201,021*	3,475
2	Nebraska, Pennsylvania	3,046	99,647	52,869
3	Michigan	2,567	164,438	57,179
4	South Carolina, Louisiana New Mexico, Washington	9,742	253,928	136,706
	Totals	18,830	719,034	250,229

*The Kentucky File was used in the second test of the name coding techniques.

For the second phase of the analysis, file 1 containing 201,021 names was used.

Comparison of these two sample sizes with the sample size totals in the tables in Appendix A indicates that not all the names selected were coded by the name coding procedures. These discrepancies exist because some of the records were dropped from the test file by the preparational programs which execute prior to the name coding procedures. These programs detected certain error conditions that eliminated these records from further processing.

Five name coding techniques were examined. These techniques were Lein, Roger Root, Census Canada, New York State Identification and Intelligence System (NYSIIS), and Central Intelligence Agency (CIA) Dictionary. Detailed descriptions of each technique are in Appendix B. Each of these techniques was used to code the surnames in each phase of the analysis.

Tables summarizing means and frequency distributions for each coding technique are included in Appendix A. Other outputs that were used consisted of surname codes containing large numbers of records, surname codes containing large numbers of unique surnames, complete listings of the records in these codes, and a listing of each surname and the code it received from each technique. Cost comparisons were also made of the five techniques. The descriptive statistics were used to analyze the coding techniques by computing the number of records per code and the number of unique surnames per code. The other output, listings of the composition of each code, were used to compare how spelling variations, recording errors (surnames recorded incorrectly on list), etc. are handled by each code; and which codes place a large number of dissimilar surnames into the same code.

ANALYSIS

Cost comparisons for the five procedures indicate that the computer processing cost of each of these coding techniques is not an important factor in selecting the best technique to use. Table 2 displays the processing time of each method that was incurred in the first phase of the analysis as well as the costs. The cost of each name coding technique was so small in relation to the estimated cost of the overall linkage system, it was not considered very strongly in the final selection. However, one entry in Table 2 that requires an additional comment is the cost of the CIA Dictionary method. Although it appears to be the cheapest, it required an additional expenditure of $800 (.35 cents/record) to prepare the surname file for the dictionary look-up procedure. This cost was for a series of sorts required to prepare the CIA Dictionary for execution.

TABLE 2 COST COMPARISONS OF NAME CODING TECHNIQUES FOR 226,600 RECORDS

Coding Technique	Processing Time	Processing Cost
Roger Root	22.50 Sec.	$38.93
Lein	19.58 Sec.	34.02
Census Canada	16.43 Sec.	28.76
NYSIIS	11.65 Sec.	21.98
CIA Dictionary	5.31 Sec.	10.70

Two other important aspects of a good name coding technique that were analyzed for each coding method were the ability of a given technique to place all variations of a given surname in the same code and yet limit the size of the codes. This size limitation, as previously explained on Page 1, would help reduce computer costs that would be incurred in the acutual linkage process.

To analyze these two characterisitics, the information in the Tables of Appendix A was used. These distributions provided data concerning the average number of unique surnames contained in each code and the average number of records contained in each code. Also used in this portion of the analysis, were printouts listing the codes created by each technique and the surnames that each code contained. These two sources were used in conjunction to evaluate these two aspects of a good name coding technique. The distributions provided an objective analysis while the observation of the surnames in each code provided a more subjective analysis.

Tables in Appendix A show that the Lein and the Roger Root techniques placed more unique surnames per code than any of the other techniques. However, observation of the surnames in the Lein and Roger Root codes also shows that these two codes contain many dissimilar surnames. The NYSIIS and Census Canada techniques placed fewer unique surnames per code than the Lein or Roger Root technique. Also, examination of the actual surnames in these codes indicated that these two methods created codes that had fewer dissimilar surnames per code than Lein or Roger Root. This is another characteristic of a good coding technique. These two techniques also placed a smaller percentage of records in codes containing 1,000 records or more.

A note should be included at this point concerning the CIA Dictionary coding method. Although the average number of surnames per code and average number of records per code compare favorably with those of the NYSIIS and Census Canada methods, the CIA Dictionary did not code all the surnames in the files. Those surnames that were not in the CIA Dictionary were assigned a miscellaneous code and grouped together; therefore, only 58.7 percent of the records in the eight state file and 68.3 percent of the records in the Kentucky file were coded using this technique. The CIA coding incompleteness results from the fact that not all variations of each surname can be included in a dictionary.

Tables 3 and 4 below summarize the information contained in the tables in Appendix A.

TABLE 3 Summary Table for Eight State Volume Tests

Surname Coding Technique	Unique Surnames/Code	Surnames/Code	Total Surname Codes	Total Surnames
Eight Character NYSIIS	2.2	11.1	20,505	226,600
Six Character NYSIIS	2.7	13.7	16,592	226,600
Census Canada	3.2	16.3	13,917	226,600
Roger Root	6.6	33.9	6,694	226,600
Lein	15.8	80.3	2,822	226,600
CIA	1.5	23.1	5,765	226,600

TABLE 4 SUMMARY TABLE FOR SINGLE STATE TEST

Surname Coding Technique	Unique Surnames/Code	Surnames/Code	Total Surname Codes	Total Surnames
Modified NYSIIS	1.9	28.5	6,881	196,407
Eight Character NYSIIS	1.9	27.2	7,223	196,407
Six Character NYSIIS	2.0	29.8	6,590	196,407
Census Canada	2.3	33.9	5,793	196,407
Roger Root	4.0	58.0	3,385	196,407
Lein	6.8	100.4	1,957	196,407
CIA	1.4	47.3	2,834	196,407

Based on the above evidence, the remainder of the analysis was concentrated on the NYSIIS and Census Canada coding techniques. The comparison of these two coding methods was made by examining the surnames placed in each code. These observations resulted in the following findings:

1. The NYSIIS technique always placed surnames ending with an "s" (John, Johns) in the same code with those that didn't but the Census Canada method did not always do this.

2. The NYSIIS technique placed surnames ending with an "s" and a "z" in the same code but the Census Canada method didn't.

3. The NYSIIS technique placed surnames that begin with a "k" and a "c" in the same code but the Census Canada method didn't.

4. The NYSIIS technique put similar surnames like Louis and Lewis in the same code but the Census Canada method didn't.

5. The NYSIIS technique could create codes of any length which would place the longer syllable words in separate codes rather than grouping them with shorter syllable surnames. The Census Canada method created only a four character code.

Based on the preceding findings, the eight character NYSIIS coding technique was selected as the surname coding method to be used in the SRS record linkage system. However, some modifications were made. These modifications are in Appendix B in the modified NYSIIS technique.

Each of the modifications was made to improve the ability of the NYSIIS technique to place all variations of a given surname in the same code. The distribution of the modified NYSIIS technique in Table 15 (Kentucky) indicates that these modifications only increased the average number of records per code to 28.5 as compared to an average of 27.2 records per code for the original eight character NYSIIS technique. However, these modifications also enabled the modified NYSIIS technique to put more spelling variations of a given surname in the same code.

SUMMARY

The eight character modified NYSIIS coding technique was selected as the surname coding method to be used in the SRS Record Linkage System. This technique satisfied the criteria desired in a coding technique. It 1) placed variations of a given surname in the same code, 2) limited the size of each code, and 3) created codes that contain few dissimilar surnames.

APPENDIX A

TABLES 1 - 13

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	13,696	66.8
2	3,136	82.1
3	1,309	88.5
4	657	91.7
5	416	93.7
6	298	95.2
7	198	96.1
8	128	96.7
9	94	97.2
10	77	97.6
11	72	97.9
12	49	98.2
13	37	98.4
14	41	98.6
15	39	98.7
16	27	98.9
17	23	99.0
18	20	99.1
19	17	99.2
20	12	99.2
21-30	85	99.6
31-40	42	99.8
41-50	15	99.9
51-60	7	99.9
61-70	6	99.9
71-80	3	99.9
81-90	1	100.0
TOTAL	20,505	
Avg. No. of Unique Surnames Per Code		2.2
Avg. No. of Surnames Per Code		11.1

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	8,840	8,840	43.1	3.9
2	3,185	6,370	58.6	6.7
3	1,619	4,857	66.5	8.9
4	1,086	4,344	71.8	10.8
5	731	3,655	75.4	12.4
6	509	3,054	77.9	13.7
7	438	3,066	80.0	15.1
8	358	2,864	81.8	16.4
9	293	2,637	83.2	17.5
10	243	2,430	84.4	18.6
11	240	2,640	85.5	19.8
12	158	1,896	86.3	20.6
13	163	2,119	87.1	21.5
14	147	2,058	87.8	22.4
15	123	1,845	88.4	23.2
16	116	1,856	89.0	24.1
17	101	1,717	89.5	24.8
18	80	1,440	89.9	25.5
19	99	1,881	90.4	26.3
20	80	1,600	90.8	27.0
21-30	552	13,769	93.4	33.1
31-40	335	11,761	95.1	38.3
41-50	201	9,035	96.1	42.2
51-60	128	7,003	96.7	45.3
61-70	105	6,859	97.2	48.4
71-80	73	5,482	97.6	50.8
81-90	56	4,785	97.8	52.9
91-100	55	5,272	98.1	55.2
101-140	129	15,353	98.7	62.0
141-180	75	12,027	99.1	67.3
181-220	46	9,029	99.3	71.9
221-260	25	5,962	99.4	73.9
261-300	26	7,253	99.6	77.1
301-400	34	11,538	99.7	82.2
401-500	21	9,299	99.8	86.3
501-1000	25	15,648	99.9	93.2
Over 1000	10	15,356	100.0	100.0
TOTAL	20,505	226,600		

TABLE 1--EIGHT CHARACTER NYSIIS NAME CODE DISTRIBUTION ANALYSIS FOR EIGHT STATE VOLUME

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	9,394	56.6
2	2,839	73.7
3	1,384	82.1
4	811	87.0
5	529	90.1
6	367	92.4
7	254	93.9
8	173	94.8
9	129	95.7
10	105	96.3
11	90	96.9
12	65	97.3
13	58	97.6
14	59	98.0
15	46	98.3
16	33	98.5
17	27	98.6
18	26	98.8
19	18	98.9
20	14	99.0
21-30	93	99.5
31-40	45	99.8
41-50	15	99.9
51-60	7	99.9
61-70	6	99.9
71-80	5	100.0
TOTAL	16,592	

Avg. No. of Unique Surnames Per Code	2.7
Avg. No. of Surnames Per Code	13.7

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	6,042	6,042	36.4	2.7
2	2,486	4,972	51.4	4.9
3	1,347	4,041	59.5	6.6
4	952	3,808	65.3	8.3
5	688	3,440	69.4	9.8
6	481	2,886	72.3	11.1
7	414	2,898	74.8	12.4
8	345	2,760	76.9	13.6
9	296	2,664	78.7	14.8
10	252	2,520	80.2	15.9
11	224	2,464	81.5	17.0
12	160	1,920	82.5	17.8
13	159	2,067	83.4	18.7
14	161	2,254	84.4	19.7
15	115	1,725	85.1	20.5
16	122	1,952	85.8	21.4
17	90	1,530	86.4	22.0
18	86	1,548	86.9	22.7
19	90	1,710	87.5	23.5
20	85	1,700	88.0	24.2
21-30	599	14,924	91.6	30.8
31-40	340	11,938	93.6	36.1
41-50	214	9,638	94.9	40.3
51-60	141	7,728	95.8	43.7
61-70	107	6,987	96.4	46.8
71-80	74	5,548	96.9	49.3
81-90	58	4,939	97.2	51.5
91-100	60	5,754	97.6	54.0
101-140	133	15,739	98.4	60.9
141-180	81	12,969	98.9	66.7
181-220	46	9,003	99.1	70.6
221-260	25	5,947	99.3	73.3
261-300	23	6,342	99.4	76.1
301-400	38	12,906	99.7	81.8
401-500	22	9,759	99.8	86.1
501-1000	26	16,171	99.9	93.2
Over 1000	10	15,407	100.0	100.0
TOTAL	16,592	226,600		

TABLE 2-- SIX CHARACTER NYSIIS NAME CODE DISTRIBUTION ANALYSIS FOR EIGHT STATE VOLUME TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	6,772	48.7
2	2,472	66.4
3	1,365	76.2
4	807	82.0
5	546	86.0
6	410	88.9
7	311	91.1
8	220	92.7
9	153	93.8
10	140	94.8
11	118	95.7
12	80	96.2
13	83	96.8
14	56	97.2
15	52	97.6
16	43	97.9
17	38	98.2
18	34	98.4
19	17	98.6
20	20	98.7
21-30	106	99.5
31-40	40	99.8
41-50	19	99.9
51-60	5	99.9
61-70	4	99.9
71-80	2	99.9
81-90	2	99.9
91-100	0	99.9
Over 100	2	100.0
TOTAL	13,917	

Avg. No. of Unique Surnames Per Code 3.2
Avg. No. of Surnames Per Code 16.3

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	4,425	4,425	31.8	2.0
2	1,999	3,998	46.2	3.7
3	1,121	3,363	54.2	5.2
4	776	3,104	59.8	6.6
5	660	3,300	64.5	8.0
6	425	2,550	67.6	9.2
7	361	2,527	70.2	10.3
8	313	2,504	72.4	11.4
9	287	2,583	74.5	12.5
10	232	2,320	76.2	13.5
11	215	2,365	77.7	14.6
12	177	2,124	79.0	15.5
13	146	1,898	80.0	16.4
14	142	1,988	81.0	17.2
15	126	1,890	82.0	18.1
16	101	1,616	82.7	18.8
17	97	1,649	83.4	19.5
18	114	2,052	84.2	20.4
19	87	1,653	84.8	21.1
20	86	1,720	85.4	21.9
21-30	546	13,597	89.4	27.9
31-40	337	11,885	91.8	33.1
41-50	206	9,345	93.3	37.3
51-60	147	8,150	94.3	40.9
61-70	119	7,761	95.2	44.3
71-80	108	8,135	95.9	47.9
81-90	79	6,725	96.5	50.9
91-100	56	5,316	96.9	53.2
101-140	157	18,445	98.0	61.3
141-180	73	11,506	98.6	66.4
181-220	54	10,640	99.0	71.1
221-260	32	7,673	99.2	74.5
261-300	27	7,614	99.4	77.9
301-400	39	13,805	99.7	83.9
401-500	14	6,221	99.8	86.7
501-1000	23	15,085	99.9	93.4
Over 1000	10	15,068	100.0	100.0
TOTAL	13,917	226,600		

TABLE 3--CENSUS CANADA NAME CODE DISTRIBUTION ANALYSIS FOR EIGHT STATE VOLUME TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	2,494	37.3
2	1,065	53.2
3	637	62.7
4	395	68.6
5	310	73.2
6	244	76.9
7	195	79.8
8	163	82.2
9	118	84.0
10	87	85.3
11	66	86.3
12	82	87.5
13	65	88.5
14	60	89.3
15	36	89.9
16	35	90.4
17	41	91.0
18	40	91.6
19	29	92.1
20	32	92.5
21-30	217	95.8
31-40	106	97.4
41-50	59	98.2
Over 50	118	100.0
TOTAL	6,694	
Avg. No. Unique Surnames Per Code		6.6
Avg. No. Surnames Per Code		33.9

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	1,685	1,685	25.2	.7
2	803	1,606	37.2	1.5
3	524	1,572	45.0	2.1
4	358	1,432	50.3	2.8
5	265	1,325	54.3	3.4
6	206	1,236	57.4	3.9
7	175	1,225	60.0	4.4
8	148	1,184	62.2	5.0
9	145	1,305	64.4	5.5
10	124	1,240	66.2	6.1
11	95	1,045	67.6	6.6
12	80	960	68,8	7.0
13	89	1,157	70.2	7.5
14	65	910	71.1	7.9
15	70	1,050	72.2	8.4
16	72	1,152	73.3	8.9
17	49	833	74.0	9.2
18	51	918	74.8	9.6
19	59	1,121	75.6	10.1
20	53	1,060	76.4	10.6
21-30	356	9,013	81.7	14.6
31-40	206	7,255	84.8	17.8
41-50	154	6,974	87.1	20.9
51-60	96	5,280	88.6	23.2
61-70	95	6,157	90.0	25.9
71-80	59	4,495	90.9	27.9
81-90	54	4,590	91.7	29.9
91-100	56	5,304	92.5	32.3
101-140	153	18,162	94.8	40.3
141-180	80	12,699	96.0	45.9
181-220	53	10,498	96.8	50.5
221-260	43	10,379	97.4	55.1
261-300	34	9,442	97.9	59.3
301-400	39	13,207	98.5	65.1
401-500	26	11,627	98.9	70.2
501-1000	54	36,496	99.7	86.3
Over 1000	20	31,006	100.0	100.0
TOTAL	6,694	226,600		

TABLE 4--ROGER ROOT NAME CODE DISTRIBUTION ANALYSIS FOR EIGHT STATE VOLUME TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	434	15.4
2	310	26.4
3	206	33.7
4	155	39.2
5	158	44.8
6	97	48.2
7	102	51.8
8	87	54.9
9	81	57.8
10	62	60.0
11	68	62.4
12	66	64.7
13	55	66.7
14	47	68.3
15	51	70.1
16	37	71.4
17	45	73.0
18	43	74.6
19	37	75.9
20	28	76.9
21-30	214	84.4
31-40	145	89.6
41-50	95	92.9
51-60	56	94.9
61-70	43	96.5
71-80	20	97.2
81-90	26	98.1
91-100	13	98.5
Over 100	41	100.0
TOTAL	2,822	
Avg. No. Unique Surnames Per Code		15.8
Avg. No. of Surnames Per Code		80.3

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	313	313	11.1	.1
2	184	368	17.6	.3
3	138	414	22.5	.5
4	114	456	26.5	.7
5	93	465	29.8	.9
6	76	456	32.5	1.1
7	59	413	34.6	1.3
8	73	584	37.2	1.5
9	64	576	39.5	1.8
10	31	310	40.6	1.9
11	40	440	42.0	2.1
12	44	528	43.6	2.3
13	38	494	44.9	2.6
14	45	630	46.5	2.8
15	31	465	47.6	3.1
16	23	268	48.4	3.2
17	29	493	49.4	3.4
18	29	522	50.5	3.7
19	31	589	51.6	3.9
20	25	500	52.4	4.1
21-30	204	5,101	59.7	6.4
31-40	138	4,886	64.6	8.5
41-50	121	5,450	68.9	11.0
51-60	84	4,685	71.8	13.0
61-70	78	5,056	74.6	15.3
71-80	55	4,167	76.5	17.1
81-90	59	5,041	78.6	19.3
91-100	40	3,823	80.0	21.0
101-140	128	15,344	84.6	27.8
141-180	85	13,443	87.6	33.7
181-220	77	15,219	90.3	40.4
221-260	48	11,248	92.0	45.4
261-300	33	9,121	93.2	49.4
301-400	70	23,974	95.7	60.0
401-500	36	16,110	97.0	67.1
501-1000	66	44,453	99.3	86.7
Over 1000	20	30,095	100.00	100.0
TOTAL	2,822	226,600		

TABLE 5-- LEIN NAME CODE DISTRIBUTION ANALYSIS FOR EIGHT STATE VOLUME TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	4,180	72.5
2	897	88.1
3	347	94.1
4	157	96.8
5	82	98.2
6	43	99.0
7	32	99.5
8	10	99.7
9	6	99.8
10	4	99.9
11	3	99.9
12	1	99.9
13	1	99.9
55	1	99.9
35,667	1	100.0
TOTAL	5,765	
Avg. No. Unique Surnames Per Code		1.5
Avg. No. Surnames Per Code		23.1

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	1,391	1,391	24.1	.6
2	710	1,420	36.4	1.2
3	481	1,443	44.8	1.9
4	342	1,368	50.7	2.5
5	240	1,200	54.9	3.0
6	198	1,188	58.3	3.5
7	153	1,071	61.0	4.0
8	162	1,296	63.8	4.6
9	119	1,071	65.8	5.1
10	98	980	67.5	5.5
11	93	1,023	69.2	5.9
12	98	1,176	70.9	6.5
13	84	1,092	72.3	6.9
14	74	1,036	73.6	7.4
15	68	1,020	74.8	7.8
16	58	928	75.8	8.3
17	41	697	76.5	8.6
18	47	846	77.3	8.9
19	46	874	78.1	9.3
20	47	940	78.9	9.7
21-30	345	8,629	84.9	13.5
31-40	175	6,197	87.9	16.3
41-50	130	5,889	90.2	18.9
51-60	93	5,131	91.8	21.1
61-70	65	4,225	92.9	23.0
71-80	57	4,283	93.9	24.9
81-90	44	3,729	94.7	26.5
91-100	33	3,146	95.3	27.9
101-140	90	10,729	96.8	32.7
141-180	61	9,862	97.9	37.0
181-220	22	4,294	98.3	38.9
221-260	20	4,803	98.6	41.0
261-300	15	4,203	98.9	42.9
301-400	28	9,524	99.4	47.1
401-500	17	7,441	99.7	50.4
501-1000	12	8,048	99.9	53.9
Over 1000	8	104,407	100.0	100.0
TOTAL	5,765	226,600		

TABLE 6--CIA NAME CODE DISTRIBUTION ANALYSIS FOR EIGHT STATE VOLUME TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	4,465	64.9
2	1,196	82.3
3	508	89.7
4	257	93.4
5	117	95.1
6	97	96.5
7	52	97.3
8	39	97.8
9	21	98.1
10	30	98.6
11	20	98.9
12	15	99.1
13	6	99.2
14	9	99.3
15	5	99.4
16	2	99.4
17	6	99.5
18	7	99.6
19	6	99.7
20	2	99.7
21-30	17	99.9
31-40	3	99.9
Over 40	1	100.0
TOTAL	6,881	
Avg. No. of Unique Surnames Per Code		1.9
Avg. No. of Surnames Per Code		28.5

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	2,093	2,093	30.4	1.1
2	711	1,422	40.7	1.8
3	443	1,329	47.2	2.5
4	291	1,164	51.4	3.1
5	218	1,090	54.6	3.6
6	192	1,152	57.4	4.2
7	171	1,197	59.9	4.8
8	135	1,080	61.8	5.4
9	133	1,197	63.8	6.0
10	118	1,180	65.5	6.6
11	110	1,210	67.1	7.2
12	109	1,308	68.7	7.9
13	85	1,105	69.9	8.4
14	79	1,106	71.0	9.0
15	83	1,245	72.2	9.6
16	62	992	73.1	10.1
17	55	935	73.9	10.6
18	55	990	74.7	11.1
19	56	1,064	75.6	11.6
20	50	1,000	76.3	12.1
21-30	404	10,125	82.2	17.3
31-40	239	8,374	85.6	21.6
41-50	158	7,173	87.9	25.2
51-60	114	6,292	89.6	28.4
61-70	99	6,470	91.0	31.7
71-80	64	4,838	91.9	32.2
81-90	71	6,040	93.0	37.3
91-100	41	3,920	93.6	39.3
101-140	134	15,921	95.5	47.4
141-180	83	13,208	96.7	54.1
181-220	56	11,052	97.5	59.7
221-260	37	8,773	98.1	64.2
261-300	26	7,241	98.5	67.9
301-400	38	12,843	99.0	74.4
401-500	22	9,690	99.3	79.3
501-1000	35	23,237	99.8	91.2
Over 1000	11	17,351	100.0	100.0
TOTAL	6,881	196,407		

TABLE 7--MODIFIED NYSIIS NAME CODE DISTRIBUTION ANALYSIS FOR KENTUCY TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	4,797	66.6
2	1,260	83.9
3	495	90.7
4	242	94.1
5	121	95.7
6	89	97.0
7	48	97.6
8	32	98.1
9	28	98.5
10	23	98.8
11	15	99.0
12	15	99.2
13	14	99.4
14	3	99.4
15	4	99.5
16	5	99.6
17	3	99.6
18	8	99.7
19	5	99.8
20	3	99.8
21-30	11	99.9
Over 30	2	100.0
TOTAL	7,223	
Avg. No. of Unique Surnames Per Code		1.9
Avg. No. of Surnames Per Code		27.2

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	2,214	2,214	30.7	1.1
2	762	1,524	41.2	1.9
3	464	1,392	47.6	2.6
4	301	1,204	51.8	3.2
5	254	1,270	55.3	3.9
6	204	1,224	58.1	4.5
7	182	1,274	60.7	5.1
8	138	1,104	62.6	5.7
9	136	1,224	64.4	6.3
10	116	1,160	66.1	6.9
11	117	1,287	67.7	7.6
12	110	1,320	69.2	8.2
13	87	1,131	70.4	8.8
14	86	1,204	71.6	9.4
15	86	1,290	72.8	10.1
16	64	1,024	73.7	10.6
17	57	969	74.5	11.1
18	58	1,044	75.3	11.6
19	61	1,159	76.1	12.2
20	57	1,140	76.9	12.8
21-30	421	10,563	82.7	18.2
31-40	244	8,541	86.1	22.5
41-50	172	7,808	88.5	26.5
51-60	111	6,115	90.0	29.6
61-70	106	6,933	91.5	33.2
71-80	66	4,978	92.4	35.7
81-90	65	5,567	93.3	38.5
91-100	37	3,536	93.8	40.3
101-140	140	16,518	95.7	48.7
141-180	92	14,605	97.0	56.2
181-220	51	10,094	97.7	61.3
221-260	38	9,043	98.3	65.9
261-300	25	6,909	98.6	69.4
301-400	40	13,636	99.2	76.4
401-500	18	7,951	99.4	80.4
501-1000	32	21,198	99.8	91.2
Over 1000	11	17,254	100.0	100.0
TOTAL	7,223	196,407		

TABLE 8--EIGHT CHARACTER NYSIIS NAME CODE DISTRIBUTION ANALYSIS FOR KENTUCKY TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	4,001	60.7
2	1,265	79.9
3	537	88.1
4	280	92.3
5	155	94.7
6	102	96.2
7	63	97.2
8	38	97.7
9	31	98.2
10	26	98.6
11	14	98.8
12	19	99.1
13	15	99.3
14	3	99.4
15	4	99.4
16	5	99.5
17	3	99.5
18	8	99.7
19	5	99.8
20	3	99.8
21-30	11	99.9
Over 30	2	100.0
TOTAL	6,590	
Avg. No. of Unique Surnames Per Code		2.0
Avg. No. of Surnames Per Code		29.8

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	1,846	1,846	28.0	.9
2	688	1,376	38.5	1.6
3	401	1,203	44.5	2.3
4	281	1,124	48.8	2.8
5	225	1,125	52.2	3.4
6	191	1,146	55.1	4.0
7	173	1,211	57.5	4.6
8	133	1,064	59.8	5.1
9	127	1,143	61.7	5.7
10	114	1,140	63.4	6.3
11	113	1,243	65.1	6.9
12	101	1,212	66.7	7.6
13	81	1,053	67.9	8.1
14	83	1,162	69.2	8.7
15	90	1,350	70.5	9.4
16	58	928	71.4	9.8
17	52	884	72.2	10.3
18	50	900	72.9	10.7
19	64	1,216	73.9	11.4
20	55	1,100	74.7	11.9
21-30	414	10,396	81.0	17.2
31-40	241	8,456	84.7	21.5
41-50	168	7,611	87.2	25.4
51-60	115	6,357	89.0	28.6
61-70	102	6,680	90.5	32.0
71-80	64	4,804	91.5	34.5
81-90	67	5,726	92.5	37.4
91-100	46	4,387	93.2	39.6
101-140	136	16,134	95.3	47.8
141-180	92	14,653	96.7	55.3
181-220	55	10,967	97.5	60.9
221-260	38	9,050	98.1	65.5
261-300	23	6,329	98.4	68.7
301-400	41	14,094	99.1	75.9
401-500	18	7,995	99.3	80.0
501-1000	33	22,084	99.8	91.2
Over 1000	11	17,258	100.0	100.0
TOTAL	6,590	196,407		

TABLE 9--SIX CHARACTER NYSIIS NAME CODE DISTRIBUTION ANALYSIS FOR KENTUCKY TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	3,113	53.7
2	1,176	74.0
3	563	83.8
4	319	89.3
5	182	92.4
6	132	94.7
7	76	96.0
8	67	97.2
9	33	97.7
10	31	98.3
11	17	98.5
12	11	98.7
13	13	99.0
14	10	99.1
15	9	99.3
16	8	99.4
17	5	99.5
18	10	99.7
19	1	99.7
20	5	99.8
21-30	10	99.9
31-40	2	100.0
TOTAL	5,793	

Avg. No. of Unique Surnames Per Code	2.3
Avg. No. of Surnames Per Code	33.9

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	1,471	1,471	25.4	0.7
2	568	1,136	35.2	1.3
3	343	1,029	41.1	1.9
4	272	1,088	45.8	2.4
5	189	945	49.1	2.9
6	175	1,050	52.1	3.4
7	146	1,022	54.6	3.9
8	132	1,056	56.9	4.5
9	128	1,152	59.1	5.1
10	82	820	60.5	5.5
11	100	1,100	62.2	6.0
12	93	1,116	63.9	6.6
13	71	923	65.1	7.1
14	70	980	66.3	7.6
15	63	945	67.4	8.1
16	58	928	68.4	8.5
17	47	799	69.2	8.9
18	48	864	70.0	9.4
19	56	1,064	71.0	9.9
20	48	960	71.8	10.4
21-30	368	9,299	78.2	15.1
31-40	299	8,096	82.1	19.2
41-50	155	7,035	84.8	22.8
51-60	114	6,288	86.8	26.0
61-70	87	5,683	88.3	28.9
71-80	73	5,477	89.5	31.7
81-90	73	6,211	90.8	34.9
91-100	57	5,471	91.8	37.6
101-140	150	17,633	94.4	46.6
141-180	101	16,052	96.1	54.8
181-220	60	11,853	97.1	60.8
221-260	39	9,237	97.8	65.5
261-300	23	6,358	98.2	68.8
301-400	34	11,660	98.8	74.7
401-500	24	10,755	99.2	80.2
501-1000	36	23,815	99.8	92.3
Over 1000	10	15,106	100.0	100.0
TOTAL	5,793	196,407		

TABLE 10--CENSUS CANADA NAME CODE DISTRIBUTION ANALYSIS FOR KENTUCKY TEST

NUMBERS OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	486	24.8
2	283	39.3
3	176	48.3
4	159	56.7
5	125	62.8
6	87	67.2
7	89	71.8
8	81	75.9
9	53	78.6
10	41	80.7
11	41	82.8
12	28	84.3
13	33	85.9
14	23	87.1
15	22	88.2
16	20	89.3
17	26	90.6
18	23	91.8
19	13	92.4
20	8	92.8
21-30	82	97.0
31-40	39	99.0
41-50	11	99.6
51-60	4	99.8
61-70	2	99.9
71-80	2	100.0
TOTAL	1,957	

Avg. No. of Unique Surnames Per Code 6.8

Avg. No. of Surnames Per Code 100.4

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	255	255	13.0	0.1
2	124	248	19.4	0.3
3	87	261	23.8	0.4
4	50	200	26.4	0.5
5	46	230	28.7	0.6
6	41	246	30.8	0.7
7	33	231	32.5	0.9
8	37	296	34.4	1.0
9	29	261	35.9	1.1
10	32	320	37.5	1.3
11	19	209	38.5	1.4
12	26	312	39.8	1.6
13	18	234	40.7	1.7
14	22	308	41.8	1.8
15	8	120	42.3	1.9
16	22	352	43.4	2.1
17	26	442	44.7	2.3
18	16	288	45.5	2.5
19	16	304	46.3	2.6
20	16	320	47.2	2.8
21-30	168	4,170	55.7	4.9
31-40	89	3,180	60.3	6.5
41-50	63	2,866	63.5	8.0
51-60	62	3,434	66.7	9.7
61-70	58	3,804	69.6	11.7
71-80	42	3,157	71.8	13.3
81-90	27	2,299	73.2	14.4
91-100	26	2,501	74.5	15.7
101-140	119	14,077	80.6	22.9
141-180	87	13,842	85.0	29.9
181-220	46	9,285	87.4	34.6
221-260	41	9,815	89.5	39.6
261-300	28	7,764	90.9	43.6
301-400	56	19,058	93.8	53.3
401-500	36	15,839	95.6	61.4
501-1000	63	42,390	98.8	82.9
Over 1000	23	33,489	100.0	100.0
TOTAL	1,957	196,407		

TABLE 11--LEIN CODE DISTRIBUTION ANALYSIS FOR KENTUCKY TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	1,453	42.9
2	598	60.6
3	399	70.6
4	239	77.7
5	140	81.8
6	116	85.2
7	72	87.4
8	59	89.1
9	64	91.0
10	37	92.1
11	35	93.1
12	34	94.1
13	17	94.6
14	23	95.3
15	22	96.0
16	13	96.3
17	12	96.7
18	14	97.1
19	8	97.3
20	5	97.5
21-30	53	99.1
31-40	13	99.4
41-50	10	99.7
51-60	3	99.8
61-70	1	99.9
71-80	2	99.9
81-90	3	100.0
TOTAL	3,385	

Avg. No. of Unique Surnames Per Code 4.0

Avg. No. of Surnames Per Code 58.0

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	728	728	21.5	.4
2	322	644	31.0	.7
3	179	537	36.3	1.0
4	127	508	40.1	1.2
5	94	470	42.8	1.5
6	81	486	45.2	1.7
7	61	427	47.0	1.9
8	70	560	49.1	2.2
9	65	585	51.0	2.5
10	53	530	52.6	2.8
11	61	671	54.4	3.1
12	57	684	56.1	3.5
13	44	572	57.4	3.8
14	38	532	58.5	4.0
15	25	375	59.2	4.2
16	42	672	60.5	4.8
17	27	459	61.3	4.8
18	37	666	62.4	5.1
19	33	627	63.3	5.5
20	28	560	64.2	5.7
21-30	223	5,660	70.8	8.6
31-40	144	5,031	75.0	11.2
41-50	104	4,680	78.1	13.6
51-60	82	4,539	80.5	15.9
61-70	77	4,987	82.8	18.4
71-80	46	3,482	84.1	20.2
81-90	39	3,312	85.3	21.9
91-100	38	3,605	86.4	23.7
101-140	124	14,674	90.1	31.2
141-180	74	11,965	92.3	37.3
181-220	56	11,176	93.9	43.0
221-260	41	9,900	95.1	48.0
261-300	22	5,984	95.8	51.1
301-400	45	15,552	97.1	59.0
401-500	21	9,387	97.7	63.8
501-1000	53	35,104	99.3	81.6
Over 1000	24	36,076	100.0	100.0
TOTAL	3,385	196,407		

TABLE 12--ROGER ROOT NAME CODE DISTRIBUTION ANALYSIS FOR KENTUCKY TEST

NUMBER OF UNIQUE SURNAMES PER CODE	TOTAL NUMBER OF CODES	CUMULATIVE PERCENT OF CODES
1	2,150	75.9
2	451	91.8
3	136	96.6
4	57	98.6
5	28	99.6
6	6	99.8
7	1	99.8
8	3	99.9
11	1	99.9
9,471	1	100.0
TOTAL	2,834	
Avg. No. of Unique Surnames Per Code		1.4
Avg. No. of Surnames Per Code		47.3

NUMBER OF SURNAMES PER CODE	TOTAL NUMBER OF CODES	TOTAL NUMBER OF SURNAMES	CUMULATIVE PERCENT OF CODES	CUMULATIVE PERCENT OF SURNAMES
1	570	570	20.1	0.3
2	241	482	28.6	0.5
3	173	519	34.7	0.8
4	112	448	38.7	1.0
5	86	430	41.7	1.2
6	67	402	44.1	1.5
7	67	469	46.4	1.7
8	52	416	48.3	1.9
9	45	405	49.9	2.1
10	43	430	51.4	2.3
11	48	528	53.1	2.6
12	38	456	54.4	2.8
13	43	559	55.9	3.1
14	26	364	56.8	3.3
15	35	525	58.1	3.6
16	43	688	59.6	3.9
17	29	493	60.6	4.2
18	26	468	61.5	4.4
19	24	456	62.4	4.6
20	24	480	63.2	4.9
21-30	191	4,816	70.0	7.3
31-40	141	4,971	74.9	9.9
41-50	89	4,047	78.1	11.9
51-60	69	3,812	80.5	13.9
61-70	63	4,142	82.7	16.0
71-80	58	4,380	84.8	18.2
81-90	41	3,508	86.2	20.0
91-100	37	3,540	87.5	21.8
101-140	117	13,949	91.7	28.9
141-180	65	10,331	94.0	34.2
181-220	40	7,950	95.4	38.2
221-260	28	6,592	96.4	41.6
261-300	18	5,070	97.0	44.1
301-400	34	11,702	98.2	50.1
401-500	16	6,998	98.8	53.7
501-1000	28	18,798	99.8	63.2
Over 1000	7	72,213	100.0	100.0
TOTAL	2,834	196,407		

TABLE 13--CIA NAME CODE DISTRIBUTION ANALYSIS FOR KENTUCKY TEST

APPENDIX B

THE NYSIIS NAME CODING PROCEDURE

1. If the first letters of the name are:
'MAC' then change these letters to 'MCC'
'KN' then change these letters to 'NN'
'K' then change this letter to 'C'
'PH' then change these letters to 'FF'
'PF' then change these letters to 'FF'
'SCH' then change these letters to 'SSS'

2. If the last letters of the name are:
'EE' then change these letters to 'Yƀ'
'IE' then change these letters to 'Yƀ'
'DT' or 'RT' or 'RD' or 'NT' or 'ND' then change these letters to 'Dƀ'

3. The first character of the NYSIIS code is the first character of the name.

4. In the following rules, a scan is performed on the characters of the name. This is described in terms of a program loop. A pointer is used to point to the current position under consideration in the name. This step begins the loop and sets this pointer to point to the second character of the name.

5. Considering the position of the pointer, only one of the following statements can be executed.

   If blank, then go to rule 7.
   If the current position is a vowel (AEIOU) then if equal to 'EV' then change to 'AF', otherwise, change current position to 'A'.

   If the current position is the letter:
   'Q' then change the letter to 'G'
   'Z' then change the letter to 'S'
   'M' then change the letter to 'N'

   If the current position is the letter 'K', then if the next letter is 'N' then replace the current position by 'N' otherwise, replace the current position by 'C'.

   If the current position points to the letter string
   'SCH' then replace the string with 'SSS'
   'PH' then replace the string with 'FF'

   If the current position is the letter 'H' and either preceding or following letter is not a vowel (AEIOU) then replace the current position with the preceding letter.

   If the current position is the letter 'W' and the preceding letter is a vowel, then replace the current position with the preceding position.

If none of these rules applies, then retain the current position letter value.

6. If the current position letter is equal to the last letter placed in the code, then set the pointer to point to the next letter and go to step 5.

   The next character of the NYSIIS code is the current position letter.

   Increment the pointer to point at the next letter.

   Go to step 5.

7. If the last character of the NYSIIS code is the letter 'S', then remove it.

8. If the last two characters of the NYSIIS code are the letters 'AY', then replace them with the single character 'Y'.

9. If the last character of the NYSIIS code is the letter 'A', then remove this letter.

THE MODIFIED NYSIIS NAME CODING PROCEDURE

1. If the first letters of the name are:
'MAC' then change these letters to 'MCC'
'KN' then change these letters to 'NN'
'K' then change this letter to 'C'
'PH' then change these letters to 'FF'
'PF' then change these letters to 'FF'
'SCH' then change these letters to 'SSS'
*'WR' then change these letters to 'RR'
*'RH' then change these letters to 'RR'
*'DG' then change these letters to 'GG'
*'A,E,I,O,U then change these letters to 'Ab̸'

*2. Drop terminal S or Z from all names before coding begins.

3. If the last letters of the names are:
'EE' then change these letters to 'Yb̸'
'IE' then change these letters to 'Yb̸'
*'YE' then change these letters to 'Yb̸'
'DT' or 'RT' or 'RD' then change these letters to 'Db̸'
*'NT' or 'ND' then change these letters to 'Nb̸'
*'IX' then change theseletters to 'ICK'
*'EX' then change these letters to 'ECK'
*'JR' or 'SR' then call this name an error and include it in table 2 of error output.

4. The first character of the NYSIIS code is the first character of the name.

5. In the following rules, a scan is performed on the character of the name. This is described in terms of a program loop. A pointer is used to point out the current position under consideration in the name. This step begins the loop and sets this pointer to point to the second character of the name.

6. Considering the position of the pointer, only one of the following statements can be executed.
(a) If blank, go to rule 7.
(b) If the current position is a vowel (AEIOU) then if equal to 'EV' then change to 'AF', otherwise, change current position to 'A'.
*(c) If the current position is a Y and it is not the last letter of the name, then change the current position to an 'A'.
(d) If the current position of the letter is:
'Q' then change the letter to 'G'
'Z' then change the letter to 'S'
'M' then change the letter to 'N'
(e) If the current position is the letter 'K', then if the next letter is 'N' then replace the current position by 'N' otherwise, replace the current position by 'C'.

*(f) If the current position is the letter 'S' and the next letter 'CH' then change to 'SSA' if end of the word or change to 'SSS' if not end of word.

*(g) If the current position is the letter 'S' and the next letter 'H' then change to 'SA' if end of the word or change to 'SS' if not end of word.

(h) If the current position is the letter 'P' and the next letter 'H' then change 'PH' to 'FF'.

*(i) If the current position is the letter 'G' and the next two letters are 'HT', then change 'GHT' to 'TTT'.

*(j) If the current position is the letter 'D' and the next letter is 'G', then change 'DG' to 'GG'.

*(k) If the current position is the letter 'W' and the next letter is 'R', then change 'WR' to 'RR'.

(l) If the current position is the letter 'H' and either preceding or following letter is not a vowel then replace the current position with the preceding letter.

(m) If the current position is the letter 'W' and the preceding letter is a vowel then replace the current position ('W') with the preceding position.

(n) If none of these rules apply, then retain the current position letter value.

7. If the current position letter is equal to the last letter placed in the code, then set the pointer to point to the next letter and go to step 6. The next character of the NYSIIS code is the current position letter. Increment the pointer to point at the next letter. Go to step 6.

8. If the last character of the NYSIIS code is the letter 'S', then remove it.

9. If the last two characters of the NYSIIS code are the letters 'AY', then replace the letters 'AY' with the single character 'Y'.

10. If the last character of the NYSIIS code is the letter 'A', then remove this letter.

*11. If the first character of the NYSIIS code is either 'A' or space, then replace it with the first letter of the original name.

*Modifications made to the original NYSIIS coding technique.

THE CENSUS MODIFIED STATISTICS CANADA NAME CODING PROCEDURE

1. Insert first character of name in first code position.

2. Examine remaining characters of name deleting all vowels and the letter 'Y'.

3. Make all multiple adjacent letters occurrence single.

4. Compress the name removing all embedded blanks.

5. Truncate to four character. If the procedures yield a code of less than four characters, blanks to the right are valid and do not need change.

THE LEIN NAME CODING PROCEDURE

1. Insert first character of name word in first code position.

2. Examine the remaining letters of the name words removing all vowels and the letters 'Y', 'W', and 'H'.

3. Make all multiple adjacent letters single and <u>truncate</u> to four characters.

4. Code the 2nd thru 4th characters with the table below padding with 0's to the right if needed to make four characters.

NOTE: In step 3 and 2, you would compress the name removing all embedded blanks before continuing.

Table for Lein Name Coding Method

Letters	Code Number
D, T	1
M, N	2
L, R	3
B, F, P, V	4
C, J, K, G, Q, S, X, Z	5

THE ROGER ROOT NAME CODING PROCEDURE

**The phonic code consists of five numeric digits.
Example: BROWNER (09424)
STANLEY (00125)

**The first letter or combination of letters are coded from the '1st Letter' table. The remainder of the letters are coded from the 'Basic' table. When vowels and the letters H, Y, and W appear other than as first letters, they are not coded.
Example: CHALMAN (06532) would be coded as follows-
CH - 06 (as shown in '1st Letter' table)
A - not coded
L - 5
M - 3
A - not coded
N - 2

**If a fully coded name results in less than five digits, pad with zeros.
Example: CHING (06270)

**If a name is too long for the five-digit code, code as many letters as possible and ignore remainder.
Example: ANDERSON (12140)
OVERSTREET (18401)

**When two letters with the same numerical value are together, they are considered as one letter.
Example: HECKEL (27500)
WYSZYNSKI (40207)

**Consonants separated by a vowel or by the letters H, Y, or W are coded separately and carry their individual values.
Example: WHITTED (41100)
ONGOOO (12770)

**The ten most common names on file would be coded as follows:

JOHNSON	(32020)
WILLIAMS	(45300)
SMITH	(00310)
JONES	(32000)
BROWN	(09420)
DAVIS	(01800)
JACKSON	(37020)
WILSON	(45020)
LEE	(05000)
THOMAS	(01300)

1st Letter Table

A	1----
B	09---
CE	00---
CH	06---
CI	00---
CY	00---
C	07---
DG	07---
D	01---
E	1----
F	08---
GF	08---
GM	03---
GN	02---
G	07---
H	2----
I	1----
J	3----
KN	02---
K	07---
L	05---
M	03---
N	02---
O	1----
PF	08---
PH	08---
PN	02---
P	09---
Q	07---
R	04---
SCH	06---
SH	06---
S	00---
TSCH	06---
TSH	06---
TS	00---
T	01---
U	1----
V	08---
WR	04---
W	4----
X	07---
Y	5----
Z	00---

Basic Table

B	9
CE	0
CH	6
CI	0
CY	0
C	7
DG	7
D	1
F	8
G	7
J	6
K	7
L	5
M	3
N	2
PH	8
P	9
Q	7
R	4
SCH	6
SH	6
S	0
TSCH	6
TSH	6
TS	0
T	1
V	8
X	7
Z	0

COMPOSITION OF SURNAME CODE FROM EACH PROCEDURE THAT CONTAINS DAVIS

Lein:

Dubose
Dubs
Dubbs
Dopps
Doviak
Dubke
Dubus
Dubois
Duboise
Doubek
Defigh
Defazio
Davis
Debaca
Dabbs
Davies
Dubukey
Debus
Debose
Daves
Dipiazza
Dobbs
Dobak
Dobis
Dobish
Doepke
Divish
Dobosh
Dupois
Dufek
Duffek
Dupuis
Dupas
Devese
Devos
Deveaux
Devies

Roger Root:

Defouw
Davis
Dauphi
Defazio
Defay
Davy
Defee
Dayhoff
Davie
Davey
Davies
Daves
Deife
Dehoff
Devese
Devoe
Devee
Devies
Devos
Dafoe
Dove
Duff
Duffey
Duffie
Duffy
Duyava
Tafoya
Tevis
Tiffee
Tivis
Thevis
Tovey
Toeves
Tuffs

NYSIIS Eight Character:

Daves
Davies
Davis
Devies
Divish
Dove
Devese
Devies
Devos

Census Canada:

Daves
Davies
Davis
Devese
Devies
Devos

CIA Dictionary:

Davis
Davies

COMPOSITION OF SURNAME CODE FROM EACH PROCEDURE THAT CONTAINS SMITH

Lein:

Sand
Sandau
Sande
Sandia
Sando
Sandoe
Sandy
Santee
Santi
Santo
Send
Sennet
Shemoit
Shenot
Shumate
Simmet
Simot
Sineath
Sinnott
Sintay
Smead
Smeda
Smit
Smite
Smith
Smithey
Smithy
Smoot
Smyth
Snead
Sneath
Sneed
Snoddy
Sonday
Sunanday
Sund
Sunda
Sunday
Sundy
Swanda
Swenda
Swent
Swint
Synott

Census Canada:

Smathers
Smith
Smithart
Smithbower
Smitherman
Smithey
Smithgall
Smithingall
Smithmyer
Smithpeter
Smithson
Smithwick
Smithy
Smotherman
Smothers
Smyth

Roger Root:

Samotid
Simmet
Simot
Smead
Smeda
Smit
Smite
Smith
Smithe
Smithey
Smithson
Smithy
Smoot
Smyth
Szmodis
Zemaitis
Zmuda

CIA:

Smith

NYSIIS Eight Character & Modified:

Schmit
Schmitt
Schmitz
Schmoutz
Schnitt
Smit
Smite
Smith
Smits
Smoot
Smuts
Sneath
Smyth
Smithy
Smithey