A METHOD OF IDENTIFYING DUPLICATIONS IN LARGE LISTS OF NAMES

by

R. Paul Moore

Research and Development Branch
Standards and Research Division
Statistical Reporting Service

June 1970

A METHOD OF IDENTIFYING

DUPLICATIONS IN LARGE LISTS OF NAMES

Introduction

Identification of duplications within a list of names is necessary when sampling from a list if it contains duplication. The approach used to identify duplicate names in this research study was to assign to each surname a numeric code based upon phonetics. Then, certain numeric and alphabetic information was compared for all pairs of names and addresses having the same phonetic code. Pairs which met certain requirements were printed out as possible duplications for further investigation. A sample of the large number of possible duplicates was studied using supplemental information sources, including telephone contacts. All pairs in the sample were classified as "duplicates" or "nonduplicates." Empirical results of this study and suggested uses for the procedure are summarized in this report.

Purpose

The purpose of the project was to develop an automated procedure for identifying duplications within a list of names.

Procedures

A computer program was written to examine each surname in a list and assign a numerical code to that listing. The Furst code technique, a procedure that groups names together which sound alike, was used. The Furst code procedure examines each letter of the surname, in sequence, assigning numeric codes as follows:

Letters	Code Assigned
S, Z	0
D, T	1
N	2
M	3
R	4
L	5
J	6
C, G, K, Q	7
F, V	8
B, P	9

The letters A, E, I, O, U, H, W, X and Y are not given codes. These are simply skipped when they appear. Exceptions to the above scheme for certain permutations of letters are:

Letters	Code Assigned
CH	6
DSY	6
TSY	6
GE	6
SY	6
SH	6
SCH	6
ZY	6
PH	8

A double letter such as SS or DD receives only a single code. Also, JR and SR are not given any code but are passed over. The program generates a 12-digit numerical code for each surname. When an entire name has been coded without using 12 digits, the remainder of the code is filled with zeros. Only the first 12 digits are recorded for names requiring longer codes.

A second computer program was developed. It grouped all names with the same Furst code together and made further comparisons. Every pair of names within each Furst code group was checked to see whether (1) the first three letters of the first name matched, (2) the first four letters of the surname matched, and (3) the first three digits of the zip code matched. If all of the above items were identical, the pair of names was printed out as a possible duplicate. An exception to criterion (1) above was made when one name lists a first name and the other lists only initials. Only the first letter of the first name field had to be identical in this situation. This caused pairs such as C. W. Stephenson and Charles Stephenson to qualify as possible duplicates.

When one of the names in a pair was a company, corporation, partnership or other nonindividual name and the other name was an individual, then criterion (1) (which compares the first names) was bypassed. This makes Start Brothers, William Start and Start Nursery and Orchard meet the qualifications of being possible duplicates.

Many of the pairs which are printed out are not duplicates since one can easily think of pairs of names and addresses such as William Jones, Wilbur Jones, and Wilson Jones which represent different persons while still meeting the three requirements stated above. There are also cases where actual duplications of names would not meet the above criteria such as William Smith and Bill Smith. If the criteria for possible duplications are made less restrictive, then the number of possible duplicates which must be checked can become very large. For example, 140,000 names from a state assessor's census were examined with a match on the 12-digit name code the only requirement for determining possible duplications. The result was that 100,000 of the names were called possible duplicates. Following up on 100,000 names with reasonable accuracy was considered an impossible task. Thus, the three restrictions stated previously were used. Using these restrictions on a later input of 28,503 names from the state assessor's census, 2,152 name and address records appeared in the output as possible duplicates. This means we would on the average need to look at about 8 percent of the names as possible duplicates instead of 71.4 percent without the additional restrictions.

It will always be necessary to visually investigate all the pairs of names classified as possible duplicates regardless of the criteria used. The methods described here can only be used to eliminate a total visual check of all possible duplicates. Additional information such as visual inspection, farm directories, telephone directories, telephone calls and personal interviews may be used in identifying bona fide duplicates.

Empirical Results

The 1964 Illinois State Farm Census was studied using some of the procedures described in the previous section. When the 140,000 names had been coded, it was apparent that further machine work was needed to reduce the amount of manual checking required. Using duplication programs, efforts were made to eliminate manual checking for pairs of names which had no similarity to each other. The first duplication program used the following criteria for selecting possible duplicate pairs:

(a) Furst code based on surname must match.
(b) First four letters of surname must match.
(c) First four letters of first name must match.
(d) Five digit zip code must match.

Test runs indicated these criteria were too restrictive since a number of apparent duplications did not meet these criteria.

The criteria used in the duplication program were modified as follows:

(a) Furst code based on surname must match.
(b) First four letters of surname must match.
(c) First three letters of first name must match.
(d) First three digits of zip code must match.

Tests were made using these criteria. A review of the output revealed that most of the actual duplications were probably identified by this procedure. Limited visual checking indicated that relaxing the criteria, to require a match on only the first two letters of the first name, identified as possible duplicates only additional pairs of names which visually appeared to be nonduplicates.

The first three digits of the zip code designate a sectional center of the mail delivery system. Most states contain from 10 to 30 sectional centers. In rural areas, each sectional center identifies an area equivalent in size to several counties. A five digit zip code, in most rural areas, defines a particular town. Thus it is desirable to make comparisons for checking duplications using only the first three letters of the zip code.

The 1967 State Farm Census list for seven counties in Central Illinois (See Figure 1) was used in an intensive study of the problems involved in determining which possible duplications were actual duplicate pairs. The 8,854 names listed in the census for the seven counties were coded and 206 pairs of names met the criteria for being possible duplicates as stated on page 2 of this report. The following sequential process was used in determining which pairs were duplicates and nonduplicates:

(a) Visual Checking - Some of the pairs listed were obviously nonduplicates although they met the criteria of the program. For example, Carl L. Thomas with Carl B. Thomas and Wiley Moore with William J. Moore. Pairs of this sort were called nonduplicates.

(b) Farm Directories - When both names of the pair were listed in the farm directories, the pair was called a nonduplicate pair.

(c) Telephone Directories - When both members of a pair were found in telephone directories, the pair was called a nonduplicate pair.

(d) Telephone Calls - Pairs remaining after the above tests were investigated further. When a telephone number was found for only one member of a pair, the individual was questioned by telephone and the determination of duplicate or nonduplicate was made. Usually the person interviewed knew if there was another individual with the same or similar name nearby. Often they had received mail intended for the other person in the past. The Mailing addresses listed for the two entries in the State Farm Census record book were also helpful.

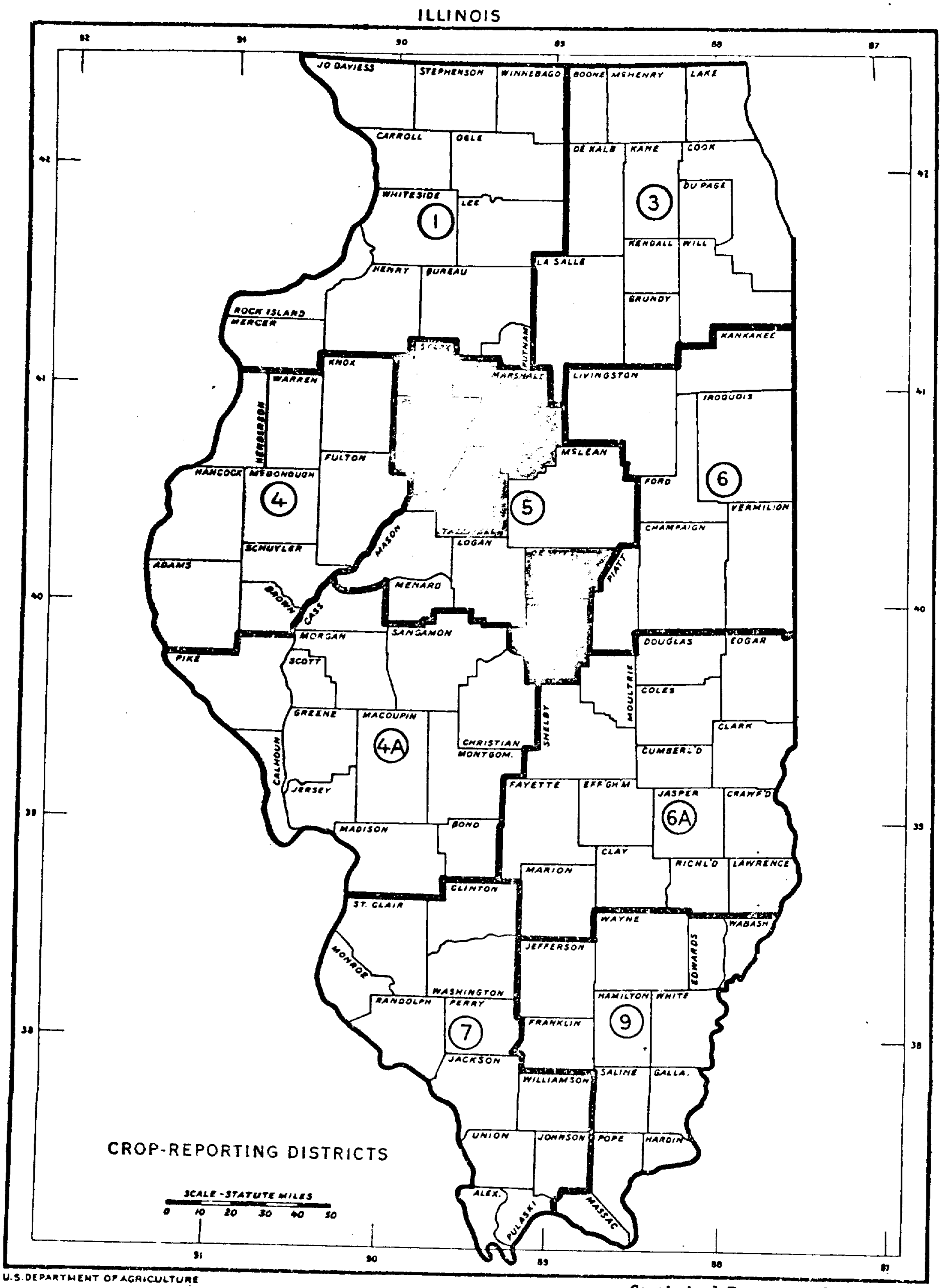
(e) Judgment - There were a few pairs for which no telephone numbers were found. Personal interviews could have been used in following up on these pairs of names. These determinations were made with the personal judgment of the investigator since time and money for personal interviews was not available.

The number of pairs classed as duplicates and nonduplicates by each of the above methods is shown in Table 1. Other methods such as personal interviews could have been used or some of the methods used in this study could have been eliminated. For instance, farm directories are not available for some areas so more pairs would have to be checked out with telephone directories and telephone calls.

Table 1.- Study of names duplicated in the 1967 Illinois State Farm Census, by methods used to determine duplications, by type of determination made, seven counties.

Methods used in sequential order	Determination made		
	Duplicate pairs	Nonduplicate pairs	Total number of pairs
	Number	Number	Number
Visual inspection........	0	53	53
Both names listed in farm directories.......	0	44	44
Both names listed in telephone directories..	0	25	25
Resolved by telephone calls.................	36	39	75
Resolved by judgment.....	8	1	9
Total number of pairs....	44	162	206

Figure 1 - Study Area for Checking Out Possible Duplications, 1967 State Farm Census.



U.S. DEPARTMENT OF AGRICULTURE

Statistical Reporting Service

Other characteristics of the possible duplicate pairs were studied. The pairs were classified as to how alike the two names were; exactly the same, nearly the same, somewhat different, and definitely different. "Nearly the same" included (a) slight differences in spelling, (b) same except one lists a middle initial, and (c) same except one lists a Jr. or Sr. "Somewhat different" included (a) first initial and middle name vs. first name, (b) first name vs. initials only, and (c) apparent partnerships vs. individual names. "Definitely different" were those pairs which could be classed as nonduplicates by visual inspection. Table 2 shows that the more alike the two names in a pair are, the higher the probability that the pair is an actual duplication. This classification method, admittedly subjective, appears to provide an indication of duplication status. The chi-square test is significant at the one percent level. This indicates a lack of independence for the pairs.

Table 2.- Number of pairs of possible duplicates, by how alike the names were, by duplication status, Illinois State Farm Census, 1967. 1/

Comparison of two names in a pair	Final determination		Total
	Duplicates	Nonduplicates	
	Number	Number	Number
Exactly same...........	27	31	58
Nearly same............	12	45	57
Somewhat different.....	5	33	38
Definitely different...	0	53	53
Total..................	44	162	206

1/ Computed chi-square (3 d.f.) = 37.83. Significant at the one percent level.

The 206 pairs of possible duplicates were also classified by (a) geographic location of the two names in the State Farm Census and (b) whether the towns listed in the mailing address were the same or different. Names are located in the State Farm Census book by townships within counties. Tables 3 and 4 show the results of these two comparisons. The chi-square tests of independence were significant at the five percent level (Table 3) and the one percent level (Table 4). There is some relationships between the duplication status and all three of the classification variables: type of names, location in census books and towns listed. The strongest indicators, towns listed and type of names, were independent of each other based on a chi-square test. The other indicator, location in census books, was related more to type of names and to towns listed (both significant chi-squares at one percent level) than to duplication status. Based on this analysis, the location in census books indicator did not appear to be as useful as type of name or town listed.

Table 3.- Number of pairs of possible duplicates, by location in census books, by duplication status, Illinois State Farm Census, 1967. 1/

Location in SFC books	Final determination		Total
	Duplicates	Nonduplicates	
	Number	Number	Number
Same township	9	42	51
Same county, different township	31	76	107
Different county	4	44	48
Total	44	162	206

1/ Computed chi-square (2 df) = 8.96. Significant at the five percent level.

Table 4.- Number of pairs of possible duplicates, by towns listed, by duplication status, Illinois State Farm Census, 1967. 1/

Towns listed	Final determination		Total
	Duplicates	Nonduplicates	
	Number	Number	Number
Towns same.............	33	57	90
Towns different........	11	105	116
Total..................	44	162	206

1/ Computed chi-square (1 df) = 22.30. Significant at the one percent level.

Recommendations

A suggestion for operational use of these procedures is as follows:

1. Process list frame through name coding and duplication programs.

2. Eliminate obvious nonduplicate pairs from the listing of possible duplicates by visual check.

3. Determine the actual duplication status for each pair of the remaining possible duplicates. Follow the procedure outlined on pages 3 and 4.

Conclusion

The job of checking all of the possible duplicates for a particular list is formidable but a requisite for reducing nonsampling error in the list frame.

Other list sources such as ASCS, tax assessors, processors, market dealers, county agents, etc., will probably contain considerably more duplication than the Illinois State Farm Census list used in this study. This problem is particularly acute if several list sources are combined into a master list frame. Failure to identify and remove duplication from these list sources could result in a serious upward bias in the estimates generated from survey data from a sample selected from the list.