SURNAME BLOCKING FOR RECORD LINKAGE

F. Quiaoit, Cancer Research Center of Hawaii, and
M.P. Mi, University of Hawaii, Honolulu

In the linkage between two documentary sources, each record from one source is compared with all the records in the other source. For one-file linkage involving a single source, each record is compared with all other records except itself. In either case, the number of such pair-wise comparisons becomes extremely large even if the size of the documentary source is moderate. The fact that only a small fraction of these comparisons are meaningful emphasizes the need for the grouping of records based on one or more selected items of identifying information. This is known as blocking. Once blocks are formed, the comparison of records is only made between the two corresponding blocks for two-file linkage or within the block for one-file linkage.

In principle, any identifier may be used as a blocking criterion. Surname is often selected for this purpose. Blocking may be made on the whole or part of the surname configuration. The use of a phonetic code on the surname for blocking has become popular in many applications. The objective of the present study was to evaluate the performance of several blocking methods based on prevalent name patterns in various racial groups in a multi-ethnic population, and to test the effects of blocking on linked pairs in which one or both records had known reporting or recording errors in the surname field.

MATERIALS AND METHODS

Data on surnames from the complete 1942-43 Population Registration in Hawaii were used. There were a total of 439,601 individuals registered and fingerprinted under martial law. Eight major racial groups were selected including Caucasian, Portuguese, Hawaiian, Chinese, Filipino, Japanese, Puerto Rican, and Korean. All recorded surname configurations for male subjects were analyzed in the present study. Two methods, namely: the New York State Identification and Intelligence System (NYSIIS) and the Russell's Soundex system were chosen to pre-code surnames phonetically. Under each method, records were blocked with the same code. These two systems were compared specifically to the other five methods of blocking, namely, by the whole surname, first character of surname, first two, three, or four characters of surname, respectively. Criteria such as the total number of blocks formed, distribution of block size, and surname information in matching were used for evaluation.

A set of known linked record pairs was obtained from the linkage project between the 1942 Population Registration file and the death file (1942-79) in Hawaii. It consisted of all male subjects aged 60 and over in the 1942 population who died during the 38-year period from 1942 to 1979. A total of 11,367 linked pairs were established by computer as well as by manual search (Mi et al., 1983). Pairs, in which recorded surname and first name were switched, were excluded. There were 672 pairs with various error conditions in surname. The concordance rate of each method, which is the percentage of record pairs that were properly placed in the same block regardless of these errors, was used for comparison.

RESULTS AND DISCUSSION

The number of male subjects in the 1942 Population Registration is shown for each racial group in Table 1. The total number of recorded configurations for surname varied greatly among racial groups ranging from only 241 in the Korean group to 14,374 among the Filipino. The average number of individuals possessing the same surname varied from 2.6 for the Caucasian group to 29.5 for Chinese men. The value for each racial group was also the average block size when blocking was based on the whole surname of twelve characters. Most of the surname configurations were unique, having only a single representation in the population. These unique configurations included rare spelling variations, and errors in reporting and recording. When a part of the surname was used for blocking, records having the same leading characters in their surname fields were grouped together. As shown in Table 1, the number of blocks increased from an initial maximum of 26, based on the first character of the surname, to several hundreds or thousands using more leading characters for blocking. However, the magnitude of increase was not linear for each additional character used, and varied from one race to another. The distribution of blocks by size also changed. When the whole surname was used for blocking, most blocks were small with 10 or less records. If blocking was based on the first character of surname, the block size increased tremendously. If more leading characters were used, the number of records in each block decreased as expected. The performance of the first four characters of surname for blocking was comparable to the NYSIIS and Soundex method in the percentage distribution of blocks by size in all groups except the Chinese and Koreans. The NYSIIS and Soundex method produced a much higher percentage of large blocks of over 50 records in the Chinese and Korean groups. This was because almost all the Chinese and Korean surnames were five characters or less in length.

It should be emphasized that block size is an important consideration in the choice of a blocking method for linkage. Since the number of pair-wise comparisons is equal to the product of the size of two corresponding blocks in two-file linkage and to the product of the block size and block size minus one in one-file

linkage, a larger block size will greatly affect the cost of a linkage.

The other criterion which deserves attention is the loss of surname information in matching by blocking. Suppose that there is no blocking and the whole documentary source or file is used as a giant block for pair-wise comparison. The amount of information provided by surname in matching is approximately $1 - \Sigma p_i^2$ where p_i is the relative frequency of the i^{th} surname configuration and $\Sigma p_i = 1$. The squared term represents the probability of chance match on the i^{th} configuration. When summed over all configurations, the squared term gives the total probability of chance match in surname. The exact probability of chance match is $1 - \Sigma p_i p_i'$ in the two file linkage where p_i' is the relative frequency of the ith configuration in the second source. If all individuals have the same surname, that is, $p_i = 1$, every record pair must agree on surname and the total probability of chance match reaches the maximum of 1. Under this special condition, surname clearly provides no information. On the other hand, if each individual record has a different surname, the probability of chance match is minimal and the amount of information provided by surname reaches the maximum. When blocking is made based on surname (a part or whole), the newly structured block consists of records of one or more surnames, each with the relative frequency of p_{ij}, the j^{th} surname within the i^{th} block. The relative frequency of the i^{th} block is q_i, and the probability of chance match for records with the i^{th} blocking criterion is q_i^2. The probability of chance match on surname within newly structured blocks is $\Sigma\Sigma p_{ij}^2/\Sigma q_i^2$, and the amount of information of surname in matching is estimated by $1 - \Sigma\Sigma p_{ij}^2/\Sigma q_i^2$. Suppose that the whole surname is used for blocking. Because each block is characterized by a different surname, obviously $\Sigma\Sigma p_{ij}^2/\Sigma q_i^2 = 1$, therefore surname is no longer informative and provides no discrimination among records within any block in which pair-wise comparisons are made.

The average and maximum number of surnames per block and the estimates of surname information in matching under various blocking methods are given in Table 2. When blocking is based on the first character, the amount of surname information was generally high except for the Korean group. The probability of chance match on surname was estimated to be 0.085, the highest among the eight racial groups studied (Kagawa and Mi, 1985). The amount of information decreased rapidly, particularly in the Chinese group, as the number of leading characters for blocking increased. When blocking is based on the NYSIIS and Soundex codes, the amount of information was close to those estimates derived from the blocking based on the first four characters in several racial groups. These phonetic coding methods seemed to be desirable especially for the Chinese and Korean groups, but not for the Japanese. The concordant rate was defined as the percentage of total pairs in which both members were blocked concordantly by a given method. Table 3 gives the estimates of the concordant rate for the four selected methods. The rate over all racial groups was 56.7, 43.9, 56.4, and 64.9 percent, respectively, for blocking based on the first three characters, first four characters, NYSIIS code, and Soundex code of surname. Both NYSIIS and Soundex methods consistently produced a high concordant rate in all racial groups. Because Chinese and Korean surnames are generally short (composed of three to five characters), errors would have to occur in the first few characters. It was anticipated that blocking based on the first three and four characters would not be highly desirable. Among the 672 linked pairs, 176 linked pairs were found to be concordant by all four methods. Erroneous conditions at the end of the surname were not detected even by the modified NYSIIS system. There were 87, 106, 98, 86 and 119 record pairs in which errors occurred in the first, second, third, fourth, and between the fifth and eighth positions, respectively. Therefore, it may be concluded that in a population where spelling variations or errors in reporting and recording usually occur after the fourth position of the surname, these four methods would perform equally well for blocking. Otherwise, NYSIIS and Soundex should be more promising than methods which are based on the use of leading characters.

REFERENCES

Mi, M.P., J.T. Kagawa, and M.E. Earle. 1983. An operational approach to record linkage. Meth. Inform. Med. 22:77-82.

Kagawa, J.T. and M.P. Mi. 1985. On matching with personal names, pp. 269-273 in this volume.

Table 1. Block Characteristics by Methods

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
Number of Male Subjects	34566	15970	7752	16118	40323	84298	4372	3786
Blocking by Complete Surname								
Number of Blocks	13286	1595	2071	546	14374	5137	924	241
Block Size Distribution, %								
1 - 10	96.7	85.1	93.4	77.5	96.6	73.8	92.3	80.1
11 - 50	3.0	10.5	6.4	14.6	3.0	19.9	6.5	13.7
51 - 100	0.2	2.6	0.1	2.0	0.2	3.1	0.8	4.6
101 - 500	0.1	1.6	0.0	5.5	0.1	3.1	0.4	0.8
501 - 1000	0.0	0.2	0.0	1.1	0.0	0.2	0.0	0.8
> 1000	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0
Average Size	3	10	4	30	3	16	5	16
Maximum Size	397	550	97	1313	289	1022	288	848
Blocking by First Character of Surname								
Number of Blocks	26	26	23	24	26	25	24	22
Block Size Distribution, %								
1 - 10	3.9	11.5	17.4	12.5	3.9	16.0	8.3	31.8
11 - 50	3.9	19.2	26.1	12.5	3.9	4.0	25.0	27.3
51 - 100	3.9	3.9	21.7	0.0	3.9	8.0	8.3	9.1
101 - 500	15.4	15.4	17.4	45.8	23.1	12.0	50.0	18.2
501 - 1000	15.4	23.1	13.0	16.7	15.4	8.0	8.3	9.1
> 1000	57.7	26.9	4.4	12.5	50.0	52.0	0.0	4.6
Average Size	1329	614	337	672	1551	3372	182	172
Maximum Size	3474	1922	4214	4157	4539	11229	811	1055
Blocking by First 2 Characters of Surname								
Number of Blocks	280	155	142	113	232	178	144	82
Block Size Distribution, %								
1 - 10	34.3	36.1	62.0	39.8	35.8	32.6	58.3	65.9
11 - 50	21.8	26.4	24.7	27.4	17.2	18.0	24.3	15.9
51 - 100	10.0	12.3	4.2	8.0	12.1	10.1	9.7	12.2
101 - 500	28.6	18.7	7.8	18.6	26.3	18.5	7.6	2.4
501 - 1000	5.0	5.8	0.7	3.5	4.7	6.7	0.0	3.7
> 1000	0.4	0.7	0.7	2.7	3.9	14.0	0.0	0.0
Average Size	123	103	54	143	174	474	30	46
Maximum Size	1008	1128	2869	4153	2809	6321	422	872

See note at the end of the table.

Table 1. Block Characteristics by Methods (Continued)

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
Blocking by First 3 Characters of Surname								
Number of Blocks	2212	655	491	354	1880	835	471	179
Block Size Distribution, %								
1 - 10	68.6	68.8	75.6	68.1	66.5	50.1	84.1	77.1
11 - 50	24.5	19.1	18.3	19.5	23.7	24.9	12.3	14.5
51 - 100	3.8	6.6	3.1	3.1	4.9	7.3	2.3	5.6
101 - 500	3.1	4.9	3.1	6.8	4.6	12.7	1.3	1.7
501 - 1000	0.0	0.6	0.0	2.5	0.2	2.9	0.0	1.1
> 1000	0.0	0.0	0.0	0.9	0.0	2.2	0.0	0.0
Average Size	16	24	16	46	21	101	9	21
Maximum Size	471	575	487	1378	740	3879	300	849
Blocking by First 4 Characters of Surname								
Number of Blocks	6941	1112	974	490	5719	1818	709	229
Block Size Distribution, %								
1 - 10	90.6	79.9	82.3	75.9	85.9	61.1	89.0	79.0
11 - 50	8.2	13.1	15.4	13.9	11.9	24.5	9.0	14.9
51 - 100	0.9	4.1	1.4	2.7	1.5	5.9	1.4	4.4
101 - 500	0.3	2.6	0.8	5.9	0.6	6.9	0.6	0.9
501 - 1000	0.0	0.3	0.0	1.0	0.0	0.7	0.0	0.9
> 1000	0.0	0.0	0.0	0.6	0.0	0.8	0.0	0.0
Average Size	5	14	9	33	7	46	6	17
Maximum Size	401	554	255	1322	422	3838	300	848
Blocking by NYSIIS								
Number of Blocks	7293	1025	631	209	6526	1922	649	89
Block Size Distribution, %								
1 - 10	91.7	79.4	80.0	71.8	87.6	55.8	88.4	68.5
11 - 50	7.1	12.5	13.8	12.4	10.7	26.4	9.2	14.6
51 - 100	0.8	4.6	4.3	3.3	1.2	6.8	1.5	10.1
101 - 500	0.4	3.2	1.9	7.7	0.6	10.0	0.8	4.5
501 - 1000	0.0	0.3	0.0	2.9	0.0	0.8	0.0	2.3
> 1000	0.0	0.0	0.0	1.9	0.0	0.2	0.0	0.0
Average Size	5	16	13	77	6	44	7	43
Maximum Size	414	586	406	2311	366	1114	300	965

See note at the end of the table.

Table 1. Block Characteristics by Methods (Continued)

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
	Blocking by Soundex							
Number of Blocks	2864	813	441	161	2779	948	555	86
Block Size Distribution, %								
1 - 10	72.9	73.8	77.1	60.9	66.8	43.1	85.8	62.8
11 - 50	22.1	16.0	15.7	16.2	26.8	26.9	11.5	16.3
51 - 100	3.6	5.8	3.6	4.4	4.8	9.5	1.6	12.8
101 - 500	1.5	4.1	3.0	13.0	1.6	15.5	1.1	5.8
501 - 1000	0.0	0.4	0.7	3.7	0.0	4.3	0.0	2.3
> 1000	0.0	0.0	0.0	1.9	0.0	0.6	0.0	0.0
Average Size	12	20	18	100	15	89	8	44
Maximum Size	449	587	774	2275	352	1395	300	885

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean.

Table 2. Surname Characteristics within Blocks

Blocking Criterion	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
	Average Number of Surnames Per Block							
First character	511	61	90	23	553	206	39	11
First 2-characters	48	10	15	5	62	29	6	3
First 3-characters	6	2	4	2	8	6	2	2
First 4-characters	2	1	2	1	3	3	1	1
NYSIIS	2	2	3	3	2	3	1	1
Soundex	5	2	5	3	5	5	2	2
	Maximum Number of Surnames Per Block							
First character	1407	184	961	73	1553	834	113	31
First 2-characters	352	100	632	53	962	376	48	22
First 3-characters	178	31	118	12	269	210	23	23
First 4-characters	37	10	60	8	117	89	10	10
NYSIIS	51	13	71	39	52	70	9	
Soundex	68	16	136	24	74	71	15	15
	Surname Information in Matching							
First character	0.99	0.89	0.99	0.81	0.99	0.98	0.86	0.47
First 2-characters	0.94	0.70	0.99	0.70	0.97	0.94	0.63	0.29
First 3-characters	0.75	0.32	0.93	0.20	0.85	0.84	0.34	0.08
First 4-characters	0.40	0.14	0.78	0.07	0.57	0.79	0.18	0.02
NYSIIS	0.48	0.17	0.90	0.57	0.46	0.43	0.20	0.25
Soundex	0.64	0.20	0.95	0.54	0.61	0.64	0.27	0.14

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean.

Table 3. Concordant Rate of Blocking

Blocking Method	Racial Groups[1]								
	Total	CAU	HAW	CHI	FIL	JAP	PUR	KOR	OTH
	Number of Linked Pairs with Errors in Surname								
	672	167	77	28	78	222	54	10	36
	Concordant Rate (%)								
First 3-characters	56.7	56.3	62.3	32.1	48.7	54.5	79.6	50.0	63.9
First 4-characters	43.9	50.3	52.0	14.3	32.1	41.4	59.3	20.0	44.4
NYSIIS	56.4	60.5	57.1	57.1	59.0	51.4	70.4	40.0	44.4
Soundex	64.9	66.5	53.3	71.4	71.8	65.3	75.9	50.0	44.4

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean; OTH = All Others.