EXACT MATCHING LISTS OF BUSINESSES: BLOCKING, SUBFIELD IDENTIFICATION, AND INFORMATION THEORY

William E. Winkler, Energy Information Administration

1. INTRODUCTION

The purpose of this paper is to present an evaluation of matching strategies for name and address files of businesses. In evaluating matching methods, we wish to minimize erroneous matches and nonmatches and the amount of manual review.

This work and previous work by various authors (Newcombe, Kennedy, Axford, and James, 1959; Newcombe and Kennedy, 1962; Newcombe, Smith, Howe, Mingay, Strugnell, and Abbatt, 1983; Coulter, 1977; Coulter and Mergerson, 1977; Rogot, Schwartz, O'Conor, and Olsen, 1983; Kelley, 1985) rely on matching strategies based on a theory of record linkage formalized by Fellegi and Sunter (1969) and first considered by Newcombe et al. (1959). The Fellegi-Sunter model provides an optimal means of obtaining weights associated with the quality of a match for pairs of records. Linked pairs (designated matches) and nonlinked pairs (designated nonmatches) receive high and low weights, respectively. Pairs designated for further manual followup receive weights between the sets of high and low weights.

Early work by Newcombe et al. (1959, 1962) showed the potential improvement (lower rates of erroneous matches and nonmatches and of manual followup) when weights were computed using surname and date of birth in comparison to when weights were computed using surname only. Coulter (1977) provided an example of the decrease in discriminating power as the probability of identifiers (such as surnames, first names, middle names, and place names) being misreported (transcribed inaccurately) and/or pairs of identifiers associated with individuals being different but accurately reported increases.

While the applied work referenced above involved files of individuals only, this paper provides an evaluation involving files of businesses. Matching using files of businesses is different from matching files of individuals because business files lack universally available and locatable identifiers such as surnames.

Matching consists of two stages. In the blocking stage, sort keys, such as SOUNDEX abbreviation of surname, are defined and used to create a subset of all pairs of records from files A and B that are to be merged. Records having the same sort key are in the same block and are considered during further review. Records outside blocks are designated as nonmatches. In the discrimination stage, surnames and other identifying characteristics are used in assigning a weight to each pair of records identified during the blocking stage.

With the exception of Newcombe et al. (1959, 1962), little work has been performed in evaluating how many erroneous nonmatches arise due to a given blocking strategy. The chief reason that little work has been performed is that identifying erroneous nonmatches due to blocking and accurately estimating error rates is difficult (Fellegi and Sunter, 1969; Winkler, 1984a,b).

The key to identifying difficulties in blocking files of businesses is having a data base in which all matches are identified and which is representative of problems in many business files. In section 2, the construction of such a data base from 11 Energy Information Administration (EIA) and 47 State and industry files is described. Section 2 also contains a summary of the Fellegi-Sunter model and the criteria used in evaluating competing matching strategies.

Section 3 is divided into two parts. The first part contains results obtained by multiple blocking strategies using a procedure in which the numbers of erroneous nonmatches and matches are minimized under a predetermined bound on the number of pairs to be passed on to the discrimination stage (for more details see Winkler, 1985b; for related work see Kelley, 1985). The results are related to results obtained during the discrimination stage and build on earlier work of Winkler (1984a, 1984b).

In the second part, the main results of the discrimination stage are presented. The effects of improved spelling standardization procedures and identification of additional comparative subfields are highlighted.

The second part also contains results on the variation of cutoff weights and misclassification and nonclassification rates during the discrimination stage. The results are based on small samples used for calibration and obtained using multiple imputation (Rubin, 1978; Herzog and Rubin, 1983) and bootstrap imputation (Efron, 1979; Efron and Gong, 1983). Fellegi and Sunter (1969, p. 1191) indicate that results based on samples are unreliable.

Finally, the second part presents results addressing the strong independence assumptions necessary under the Fellegi-Sunter model and conditioning techniques that can be used in improving matching performance in some situations when direct application of the Fellegi-Sunter model yields high misclassification and/or nonclassification rates. The investigation of independence uses the hierarchical approach of contingency table analysis (Bishop, Fienberg, and Holland, 1975). The conditioning argument uses a steepest ascent approach (Cochran and Cox, 1957).

Section 4 contains a summary.

2. EMPIRICAL DATA BASE, METHODS, AND EVALUATION CRITERIA

This paper's approach to developing more effective matching strategies involves:

1. constructing an empirical data base for testing procedures;
2. employing the Fellegi-Sunter model of record linkage;
3. defining evaluation criteria; and
4. refining procedures in response to empirical results.

2.1. Creation of a Suitable Empirical Data Base

The empirical data base consists of 66,000 records of sellers of petroleum products. It was constructed from 11 EIA lists and 47 State and industry lists containing 176,000 records. Easily identified duplicates having essentially similar NAME and ADDRESS fields were deleted when the melded file was reduced from 176,000 to 66,000 records.

The data base contains 54,850 records identified as headquarters or parents (records used for mailing purposes); 3,050 records identified as duplicates (records having names and addresses similar to their parents'); and 8,511 records identified as associates (records such as subsidiaries and branches that have names and/or addresses different from their parents').

Duplicates were identified primarily through elementary computer-assisted techniques (see Winkler, 1984a); associates were identified through surveying and call-backs. Our evaluation will only consider how well various strategies perform in matching duplicates with headquarters. The presence of unidentified associates, however, can cause falsely higher error rates.

2.2. Methods

2.2.1. The Formal Probabilistic Model

The Fellegi-Sunter model (1969) uses an information-theoretic approach embodying principles first used in practice by Newcombe (Newcombe et al., 1959). In practice, specific binit weights of agreement (or disagreement) are computed by,

$$W = \log_2 A/B$$

where

A= the proportion of a particular agreement (or disagreement) defined as specifically as one wishes among matched pairs, and

B= the corresponding proportion of the same agreement (or disagreement) among pairs that are rejected as matches.

The following table will help us to understand more specifically the computation of weights.

Counts of True State of Affairs

Specified Characteristic	Match	Nonmatch
Agree	a	b
Disagree	c	d

If we wish to compute the weight associated with agreement on a specified characteristic, then we take A=a/(a+c) and B=b/(b+d); for disagreement, we take A=c/(a+c) and B=d/(b+d).

For each detailed comparison of a pair of records, the weights for appropriate agreements and disagreements are added together, and the total weight, TWT, is used to indicate the degree of assurance that the pair relates to the same entity. The procedure assumes that weights associated with individual agreements or disagreements are uncorrelated with each other (at least conditionally, see e.g., Fellegi and Sunter, 1969, p. 1190).

Cutoffs UPPER and LOWER are chosen (using empirical knowledge or educated guesses) and the following decision rule is used:

If TWT > UPPER, then designate pair as a match.

If LOWER < = TWT < = UPPER, then hold for manual review.

If TWT < LOWER, then designate pair as a nonmatch.

Given fixed upper bounds on the percentages of erroneous nonmatches having TWT < LOWER and of erroneous matches having TWT > UPPER, Fellegi and Sunter (1969, p. 1187) show that their procedure is optimal in the sense that it minimizes the size of the manual review region.

In some cases, either looking at disjoint subsets of the set of blocked pairs and/or increasing or decreasing individual weights used in computing the total weight, TWT, can improve the efficacy of the above decision rule. For instance, among a set of records that are blocked into pairs using the first six characters of the STREET field, individual weights associated with agreements and disagreements on characteristics of the NAME field might be increased and decreased, respectively.

A procedure that uses individual weights, that have been varied in order to achieve greater accuracy in the set of pairs designated as matches and nonmatches and/or a reduction in the set of records held for manual review, will be referred to as a modified information-theoretic procedure. An unmodified procedure will be referred to as the basic information-theoretic procedure.

2.2.2. Specific Weight Computation

Weights are computed for comparisons of the the following subfields of the STREET field:

HOUSE NUMBER, PREFIX (direction words), STREET NAME, SUFFIX (words such as ST and RD), UNIT DESIGNATOR.

Weights are computed for subfields of the NAME field:

KEYWORD1 (largest word),
KEYWORD2 (2nd largest, tie broken by alpha sort), and
CO (concatenation of initials). The following subfields were also used in computing individual weights:

Field	Subfield Columns	Designated
NAME	1-4,5-10,11-20,21-30	N1,N2,N3,N4
STREET	1-6,7-15,16-30	S1,S2,S3
ZIP	1-3,4-5	Z1,Z2
CITY	1-5,6-10,11-15	C1,C2,C3
STATE	1-2	
TELEPHONE	1-3,4-6,7-10	T1,T2,T3
WL-NAME1/	1-4,5-10,11-20,21-30	W1,W2,W3,W4

1/ Sort words in NAME field by decreasing order of word length. Break ties with alpha sort.

2.2.3. Variances

As the truth and falsehood of matches in the set of blocked pairs were known for the evaluation files, estimated error rates and their variances were obtained using multiple samples.

The basic procedure was to draw samples of equal size, compute cutoff weights using each sample (based on at most 2 percent of non-matches being classified as matches and at most 3 percent of matches being classified as nonmatches), use each pair of cutoff weights on the entire data base to determine overall error rates, and compute the variances of the cutoff weights and the overall error rates over the set of samples.

The multiple imputation procedure of Rubin (1978) has been used for evaluating the effects of different methods of imputing for missing data but is applicable in our situation. Multiple imputation entails obtaining several estimates using different samples and then computing the mean and variance over samples. In using Rubin's procedure, we sample without replacement.

The key difference from Efron's bootstrap is that sampling is performed with replacement. Our application corresponds to the first example in the paper of Efron and Gong (1983).

2.2.4. The Independence Assumption

Fellegi and Sunter (1969, pp. 1189-90) state that the independence assumption for the comparisons of information contained in different subfields is crucial to their theory but that the independence assumption may not be crucial in practice. They note that obtaining total weights having a probabilistic interpretation only necessitates that comparisons be conditionally independent. The conditioning must be consistent with the way total weights are computed.

Even if dependencies occur, it may be possible to vary weights associated with individual comparisons (i.e., steepest ascent, see e.g., Cochran and Cox, 1957, pp. 357-369) to determine whether the efficacy of the overall weighting procedures can be improved. Our specific steepest ascent method generally involved choosing a few individual weights in disjoint subsets determined by blocking criteria (sections 3.1 and 3.2) and varying them by +/- 0.5.

It is important to note that modifications to individual weights may be heavily dependent on the subsets determined by the blocking criteria.

2.3. Criteria for Evaluation

A Type I error is an erroneous nonmatch and a Type II error is an erroneous match. The Type I error rate is U/D*100 where U is the number of erroneous nonmatches and D is the number of matches. The Type II error rate is F/M*100 where M is the number of pairs designated as matches and F is the number of erroneous matches.

3. RESULTS USING THE EMPIRICAL DATA BASE

Results of the empirical analyses for the blocking stage and the discrimination stage are presented in sections 3.1 and 3.2 respectively.

3.1. Comparison of Sets of Blocking Strategies

The following four criteria were used for blocking files into sets of linked pairs used in the discrimination stage. The set of four criteria was developed by comparing a large number of criteria. Detailed reasons for their adoption are given in Winkler (1985b).

BLOCKING CRITERIA

1. 3 digits ZIP, 4 characters NAME
2. 5 digits ZIP, 6 characters STREET
3. 10 digits TELEPHONE
4. Word length sort NAME field, then use 1.*

*This criterion also has a deletion stage which prevents matching on commonly occurring words such as 'OIL,' 'FUEL,' 'CORP,' and 'DISTRIBUTOR.'

Blocking 3050 duplicates with 54,850 parents using the set of blocking criteria yielded 4485 pairs (2991 matches and 1494 nonmatches) for consideration during the discrimination stage.

It is important to note that there are 39 matches that are not identified during the blocking stage. They are never again considered.

3.2. Discrimination

The discrimination stage was divided into two parts: (1) a part in which 2240 pairs were designated as matches using an ad hoc decision rule and (2) a discrimination stage in which the remaining 2245 pairs were designated as either matches, erroneous matches, or candidates for manual review.

The ad hoc decision rule generally consisted of designating those pairs as matches that had been connected by two or more blocking criteria. The exceptions were records connected by 1 and 4, only (NAME and WL-NAME), and 2 and 3, only (STREET and TELEPHONE). Slightly more than 98 percent of the 2240 records designated as matches were actually matches.

Prior to use in the information-theoretic discrimination procedure, the 2245 remaining pairs were further divided into four mutually exclusive classes using the blocking criteria:

Class 1 (1021 records):	Linked by 1, only, and by 1 and 4, only.
Class 2 (624 records):	Linked by 2, only, and by 2 and 3, only.
Class 3 (256 records):	Linked by 3, only.
Class 4 (344 records):	Linked by 4, only.

3.2.1. Overall Results

Table 1 presents a summary of results obtained during the discrimination stage. It shows that 2148 (96 percent) of 2245 records are classified as matches or nonmatches and that only 3 percent (68/2148) of the classified records are misclassified. Results are based on using the entire data set for calibration (i.e., obtaining cutoff weights) and evaluation. Variance results (section 3.2.2) based on 25 different samples used for calibration yield cutoff weights and error rates that are consistent with results in Table 1.

Two observations are that the cutoff weights vary substantially across classes and that 100 percent of the records in classes 2 and 4 can be classified. The varying cutoff weights indicate that cutoff weights may vary with different types of address lists. Thus, new calibration information may be needed for each new file encounted. Calibration information is based on knowing the actual truth and falsehood of matches within a representative set of blocked pairs.

The largest group of misclassified records are those erroneous matches that have the same address and phone number as the headquarters' records. For example:

(a) Apex Oil 222 Columbia St NE
Salem OR 97303 503/588-0455
Jones Co 222 Columbia St N E
Salem OR 97303 503/588-0455
(b) A A Oil Main St
Smallsville TX 77103 713/643-2121
Smith J K Co Main St
Smallsville TX 77103 713/643-2121

Example (a) represents two different companies located in the same office building. Example (b) represents two different fuel oil dealers, one of which has gone out-of-business.

Misclassified matches (erroneous nonmatches) generally had typographical differences or missing data in a number of subfields, as in the examples below:

(c) Smith Oil W 31st St N Church St
Hardsburg PA 18207 713/643-2121
Smith J K N Church St
Hardsburg PA 18207 missing
(d) Mcneely R 3312-14 Harris Ave
MPLS MN 55246 612/929-6677
R Mcden Neely 3312 Harris Ave
St Louis Par MN 55246 612/929-6677

Example (c) has a minor variation in the NAME field, a major variation in the STREET field, and a missing TELEPHONE field. Example (d) has major variations in the NAME field and CITY fields and a minor variation in the STREET field.

3.2.2. Variances

Tables 2, 3, and 4 present estimates and their coefficients of variation obtained using 25 calibration samples and Rubin's multiple imputation technique. For each calibration sample, the sample sizes in Classes 1, 2, 3, and 4 were 240, 200, 120, and 160, respectively. Cutoff weights and misclassification rates were obtained for each sample. Estimates are the average cutoff weights and average misclassification rates over 25 replications (samples). Variances of the estimates are over 25 replications.

Overall, the results indicate that the estimated cutoff weights and misclassification rates vary significantly from calibration sample to calibration sample. The variances are functions of both the sample sizes on each replication and the number of replications. When the number of replications was held at 25 and the sample sizes decreased to 120, 100, 80, and 90 for the four classes, estimated coefficients of variation over 25 replications were approximately 30 percent higher on the average for misclassified matches and about the same for misclassified nonmatches.

The fact that the coefficients of variation decrease substantially as sample sizes increase indicates that calibration samples should be as large as possible. As the total number of records considered in these analyses was quite small, taking substantially larger samples was not practicable.

Examination of Table 2 shows that the estimated coefficients of variation associated with the cutoff weights using the modified information-theoretic procedure range from 15.3 percent to 99.5 percent; and from 14.3 percent to 115.4 percent with the basic information-theoretic procedure. The cutoff weights are consistent with the cutoff weights given in Table 1. Results in Tables 1 were obtained using the entire data set instead of samples.

Examination of Tables 3 and 4 show that the misclassification and nonclassification rates can vary significantly. Coefficients of variation of the estimated misclassification rates for the modified information-theoretic procedure vary from 33.2 to 109.9; for the basic procedure from 33.8 to 112.9.

Comparison of the modified and basic weighting procedures shows that the modified procedure is able to classify accurately significantly more records, particularly in classes 2 and 4, than the basic procedure. The results are consistent with those presented in Table 1.

Results obtained using Efron's bootstrap imputation with 25, 100, 200, and 500 replications are consistent with the results in Tables 2, 3 and 4.

3.2.3. The Independence Assumption

Independence of comparisons does not hold. This is shown by the significant variation of the lower and upper cutoff weights across Classes 1 thru 4 in Tables 1 and 2. If the comparisons were independent, then individual weights and cutoffs for the total weights would be reasonably consistent across classes. Individual weights (not shown) vary more than the cutoff weights across classes.

Independence of interactions within classes is illustrated by Table 5. It shows the two-way independence of the interactions of some of the subfields given in section 2.2.2.

In over half the entries in Table 5 the two-way interactions are independent unconditionally at the 95 percent confidence level and the hierarchical principle (Bishop, Fienberg, and Holland, 1975) assures that all such two-way interactions are always conditionally independent. In all cases in which two-way interactions are not unconditionally independent, a third variable was found so that the two-way interactions were independent at the 95 percent confidence level given the third variable (see also Winkler, 1985b).

It is important to note two points. First, some of the interaction of variables (not presented in the tables) such as H and S1 or W1 and K11 are often not independent unconditionally and it seems likely that they will generally not be independent conditionally. Second, building a precise model, by mutually exclusive class, in which only the minimal set of variables necessary for effective discrimination is included, and which precisely models the conditional relationships, is likely to be difficult and heavily dependent on the empirical data base used.

4. SUMMARY

The results of this paper imply that the keys to delineating matches and nonmatches accurately are: (1) good spelling standardization and (2) accurate identification of corresponding subfields. They also imply that the independence assumption, required by the information-theoretic model of Fellegi and Sunter (1969), is not critical in practical applications of the type performed in this paper.

ACKNOWLEDGEMENT

The author wishes to thank Dr. Yvonne M. Bishop for a number of helpful comments and conversations.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), Discrete Multivariate Analysis, MIT Press, Cambridge, MA.

Coulter, R.W. (1977), "An Application of a Theory for Record Linkage," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Coulter, R.W. and Mergerson, J.W. (1977), "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Cochran, W.G. and Cox, G.M. (1957) Experimental Designs, J. Wiley and Sons, New York. Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," Ann. Stat., 7, 1-26.

Efron, B. and Gong, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," The American Statistician, 37, 36-48.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," JASA, 40, 1183-1210.

Herzog, T. and Rubin, D. (1983), "Using Multiple Imputations to Handle Nonresponse," in Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies, edited by Madow, W.G., Olkin, I., and Rubin, D.B. Academic Press, New York, 210-245.

Kelley, R. P. (1985), "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Invited paper presented at the Workshop on Exact Matching Methodologies in Rosslyn, VA, on May 9-10, 1985.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," Science, 130, 954-959.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM, 5, 563-566.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J., Strugnell, A., and Abbatt, J.D. (1983), "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers," Comput. Biol. Med., 13, 157-169.

Rubin, D. (1978), "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," ASA 1978 Proceedings of the Section on Survey Research Methods, 20-28.

Rogot, E., Schwartz, S., O'Conor, K., and Olsen, C. (1983), "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index." ASA 1983 Proceedings of the Section on Survey Research Methods, 319-324.

Winkler, W. E. (1984a), "Issues in Developing Frame Matching Procedures: Exact Matching Using Elementary Techniques." Presented to the ASA Energy Statistics Committee in April 1984.

Winkler, W. E. (1984b), "Exact Matching Using Elementary Techniques." ASA 1984 Proceedings of the Section on Survey Research Methods, 237-242.

Winkler, W. E. (1985a), "Preprocessing of Lists and String Comparison," Invited paper presented at the Workshop on Exact Matching Methodologies in Rosslyn, VA, on May 9-10, 1985.

Winkler, W. E. (1985b), "Exact Matching Lists of Businesses: Blocking, Subfield Comparison, and Information Theory," to appear in Record Linkage Techniques, Proceedings of the Workshop on Exact Matching, Statistics of Income Division, Internal Revenue Service.

Table 1: Results from Using a Modified Information-Theoretic Model for Delineating Matches and Erroneous Matches (3 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total Classed	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	4.5	7.5	28	8	692	274	966	1021
2	2.5	2.5	5	3	379	245	624	624
3	-0.5	4.5	5	6	104	110	214	256
4	8.5	8.5	9	4	266	78	344	344
Totals			47	21	1441	707	2148	2245

Table 2: Estimated Cutoff Weights and Their Variances 25 Replications, With and Without Conditioning

Class	Status 1/	Estimated Cutoff Weights		Variance of Estimated Cutoff Weights		CVs of Estimated Cutoff Weights	
		LOWER	UPPER	LOWER	UPPER	LOWER	UPPER
1	C	2.66	7.72	7.02	2.05	99.5	18.5
2	C	1.44	1.44	0.62	0.62	54.9	54.9
3	C	-3.39	5.82	8.74	2.08	87.2	24.8
4	C	6.89	11.92	1.11	7.57	15.3	23.1
1	WC	-1.92	8.05	4.90	1.50	115.4	15.2
2	WC	-5.04	4.56	0.52	1.41	14.3	26.1
3	WC	-6.38	6.82	1.46	1.66	18.9	18.9
4	WC	1.71	12.13	3.11	7.56	102.9	22.7

1/ C-Conditioning, WC-Without Conditioning.

Table 3: Estimated Counts and Rates of Misclassification and Nonclassification 25 Replications, With and Without Conditioning

Class	Status 1/	Total Records	Misclassed as		Not Classed	Correctly Classed as		Proportion Misclassed as	
			Match	Non-Match		Match	Non-Match	Match	Non-Match
1	C	1021	10.4	27.4	75.2	260.7	647.2	.038	.041
2	C	624	9.7	3.0	0.0	244.0	367.3	.038	.008
3	C	256	3.0	3.5	94.2	85.2	70.0	.034	.048
4	C	344	1.4	10.2	23.5	54.3	254.6	.026	.039
Total		2245	24.5	44.1	192.9	644.2	1338.1	.037	.032
1	WC	1021	8.9	26.2	145.4	237.1	603.3	.036	.042
2	WC	624	3.8	3.9	450.6	89.4	76.3	.040	.048
3	WC	256	1.6	2.3	178.8	38.1	35.1	.041	.062
4	WC	344	1.3	9.6	57.7	38.8	236.6	.032	.039
Total		2245	15.6	42.0	832.5	403.4	951.3	.037	.042

1/ C-Conditioning, WC-Without Conditioning.

Table 4: Coefficients of Variation of Estimated Counts of Misclassification and Nonclassification 1/, 25 Replications, With and Without Conditioning

Class	Status 2/	Total Records	Misclassed as		Not Classed
			Match	Non-Match	
1	C	1021	69.5	47.4	54.7
2	C	624	64.6	81.1	0.0
3	C	256	96.6	84.1	40.9
4	C	344	109.9	33.2	60.8
Total		2245			
1	WC	1021	62.3	42.3	34.0
2	WC	624	112.9	96.2	9.0
3	WC	256	106.9	65.5	8.1
4	WC	344	99.6	33.8	34.3
Total		2245			

1/ Units are percentages.

2/ C-Conditioning, WC-Without Conditioning.

Table 5: Independence of Two-Way Interactions for Selected Subfields that are Generally Not Connected with Blocking Characteristics, By Class 1/

Class	K11/H	K22/H	K11/SN	K22/SN
1	yes	yes	no 2/	no 2/
2	NA	NA	yes	yes
3	no 4/	no 3/	no 2/	yes
4	yes	yes	yes	yes

NA- not applicable because one of two variables is basically the same as a blocking characteristic due to small sample size.

1/ Kii compares the ith KEYWORDs for i=1, 2; H compares HOUSE NUMBER; and SN compares STREET NAME.

2/ Independent when H is included in a 3-way contingency table analysis.

3/ Independent when K11 is included.

4/ Independent when K22 is included.