Transportation Research Forum

Probabilistic Linkage Approach to Commercial Motor Vehicle and Carrier Datasets

Author(s): Jung-Taek Lee and Piyushimita Thakuriah (Vonu)

Source: *Journal of the Transportation Research Forum*, Vol. 43, No. 2 (Fall 2004), pp. 37-52
Published by: Transportation Research Forum
Stable URL: http://www.trforum.org/journal

Probabilistic Linkage Approach to Commercial Motor Vehicle and Carrier Datasets

In this paper, a probabilistic linkage method is explored in the context of linking databases in the Commercial Motor Vehicle and Carrier (CMVC) sector as a potential solution to overcome data quality problems. An application of this method is demonstrated by linking commercial motor vehicle inspection files kept by the Illinois State Police (ISP) and the inspection files available from the Illinois portion of the Motor Carrier Management Information System (MCMIS). Since one of the files to be matched is a subset of the other, the application allows us to validate the methodology. The results show 6,228 correct identifications of true matched record pairs out of 6,335 actual true matches (more than 99%) between the two files. The number of erroneously identified record pairs is 690 (about 11% of the actual true matched pairs.) Sensitivity analysis is conducted of error rates with respect to variations in the optimal thresholds for merging the databases. A simple analysis also shows how much of a clerical examination for unclear record pairs would have to be tolerated for a reduction in dollar expenditure.

by Jung-Taek Lee and Piyushimita Thakuriah (Vonu)

INTRODUCTION

Many policy developments on various transportation issues such as transportation safety require comprehensive analyses of large amounts of supporting data. These data are often not available from a single database, but rather from multiple databases that are sometimes owned and maintained by different agencies. Often these databases from various sources need to be merged to obtain necessary pieces of information. This calls for the linking or integration of diverse databases and establishing a new database that contains information relevant to study the problem at hand. However, it is not unusual for analysts to encounter problems such as lack of unique identifiers and data-quality issues in key identifiers.

In this paper, we address the issue of merging two data files when there are no unique identifiers. The application area in this study is the commercial motor vehicle and carrier (CMVC) sector. Safety-related databases in this sector are kept by various state and national regulatory, enforcement and commerce-related agencies. Linking these databases enables safety analysts to gain knowledge that would otherwise not be possible simply because the disparate pieces of information are in different databases. However, unique identifiers such as vehicle identification numbers (VIN), inspection numbers, or firm ID variables are not necessarily available or are incomplete in the files to be merged. Even if there are unique identifiers, these are often not unique enough to identify and correctly match two records (for example, identifiers of month, date, etc.) In addition, they have various data quality issues such as missing values and inconsistency across the files. All of these issues make the use of direct or deterministic data merge applications difficult. Probabilistic linkage can offer a way of overcoming this type of potential difficulty.

PREVIOUS APPLICATION OF PROBABILISTIC DATA LINKAGE

Probabilistic linkage has been extensively used in health-care record studies since its initial use (Dunn 1946; Fair 1999; Gill 1999; Newcombe et al. 1959; Newcombe 1988). The pioneering conceptual idea (Newcombe et al. 1959) was further extended by Fellegi and Sunter (1969) who produced optimal decision rules. Newcombe (1988) later developed a simpler concept of frequency ratio of record linkage in the application to a health data linkage. The most extensive study regarding record linkage in the transportation area is the Crash Outcome Data Evaluation Systems (CODES) program supported by the National Highway Traffic Safety Administration (NHTSA). The purpose of the CODES project is to link crash, hospital, EMS (Emergency Medical Services) claims, and long-term care data to conduct various analyses on health-care related issues. Johnson (1999) presented technical issues related to the linkage of population-based person-specific state crash and injury data. Utter's (1999) analysis of safety belt and helmet effectiveness resulted from the probabilistic linkage of crash and hospital data. Bigelow et al. (1999) employed probabilistic linkage to link Wisconsin CODES datasets, resulting in 70% matches between hospital and motor vehicle injury databases. Dean (2002) also presented probabilistic linkage between Utah crash and hospital data, but only linked about 12% of the data. The low linkage turned out to be the result of many missing values. Finison (2002) evaluated crashes involving older drivers in Maine by merging hospital, death certificate and crash data. It was found that urban and low-speed areas accounted for crashes among older drivers. These studies were restricted to only matched rates of two datasets and did not include resulting error rates that are required for validation purposes. These studies also were mostly conducted for health and medical-related issues, which could use personal identifiers (i.e., name, age, social security numbers and sex) recorded in hospital records, EMS, etc. Unfortunately, these unique identifiers are not usually available for the transportation safety studies that lead to law enforcement and policy decisions. There are very few studies that use the methodology in analyzing exclusively transportation-related datasets such as commercial motor vehicles and carriers.

DATA QUALITY ISSUES OF COMMERCIAL MOTOR VEHICLE AND CARRIER FILES

The CMVC safety sector requires data from many sources for complete analysis. In a previous study (Thakuriah et al. 2002), several data quality issues pertaining to the "cleanliness" of identifiers in this sector were described. In order to achieve a good success rate in matching different data files, two Data Quality Criteria (DQC) were suggested. These are data completeness (the degree to which identifier variables in the databases are free of missing values) and consistency (the degree of uniformity, standardization, and freedom from contradiction in data values in different files.) The authors considered three different databases in the CMVC sector. These are the Illinois State Police (ISP) Inspection Files, the Illinois portion of the Motor Carrier Management Information System (MCMIS) Crash and Inspection Files, and the Illinois Department of Transportation (IDOT) Crash Files. The authors found a large variation in the completeness in these databases. Approximately 0.5% of the records from the ISP inspection database did not have a value for the VIN variable. License state and license numbers of commercial motor vehicles inspected were complete for all records. On the contrary, the IDOT Crash Files showed that about 18.2% of VINs and 99.3% of the license state variables were missing. The MCMIS Crash File also turned out to have a large number of missing values; approximately 21% of VINs, and 13 to 17% of license state and license numbers. The MCMIS Inspection Files, on the other hand, had a high level of recorded values on license state and license numbers (0.03 to 1.9%); however, this system has large amounts of missing values for VINs (about 93%).

The files were also checked for completeness as defined above. Records in the MCMIS Crash File were compared with the ISP Inspection Files. For records in the MCMIS Crash File and the ISP Inspection File that had common VINs, about 98% of the license state values matched and about 88% of the license numbers were the same. However, the consistency check was not possible with the MCMIS Inspection File or the IDOT Crash Files because of the large numbers of missing values (of VIN in the MCMIS Inspection File and license state and license numbers in the IDOT Crash Files).

Given the scale of the data quality concerns in these illustrative databases, it is natural to speculate that exact or deterministic linkage of the different files on the basis of unique identifiers such as license number, license state and VINs would not be very successful. This reason motivates us to examine the potential of probabilistic linkage methods for integrating data in the CMVC sector.

PROBABILISTIC LINKAGE CONCEPT

The record linkage is to bring two files, file A and B, together using common identifiers such as sex, date of birth, address, etc. A probabilistic record linkage refers to the calculation of the likelihood of a correct linkage between two files. The theoretical background of this linkage methodology is discussed below based on the original works of Fellegi and Sunter (1969) and Newcombe (1988).

Matched and Unmatched Sets

Let *a* and *b* be values (elements of identifiers or data variables) of two data files *A* (searching file) and *B* (searched file), assuming some elements are common to *A* and *B*. The set of ordered pairs is then defined:

$A \times B = \{(a,b);\ a \in A,\ b \in B\}$. The set $A \times B$ represents all possible pairs of values (*a, b*) between file A and B, where element *a* belongs to A and *b* to B. The joint sets are a matched set that can be denoted by $M = \{(a,b);\ a = b,\ a \in A,\ b \in B\}$, and an unmatched set that can be denoted by $U = \{(a,b);\ a \neq b,\ a \in A,\ b \in B\}$ respectively. The *M* represents a set that assumes to hold exactly matched values (i.e., *a* and *b* are the same values) of record pairs from $A \times B$. The *U* represents a set of exactly unmatched values (*a* and *b* are not the same values) of record pairs from $A \times B$.

Linkage Weights

Let records corresponding to identifiers (data variables) of *A* be $\alpha(a)$ and records corresponding to identifiers of *B* be $\beta(b)$. Let $\gamma(a,b)$ or γ be a realization of an element pair of an identifier. Then the linkage operation is to observe γ and decide whether (a,b) is matched or not. That is,

$$(1)\ \begin{cases} (a,b) \in M \text{ (a positive link (i.e., matched))}: A_1 \\ (a,b) \in U \text{ (a positive non-link (i.e., unmatched))}: A_3 \end{cases}$$

A positive link (A_1) represents record pairs that indicate exactly the same identity or individual from files A and B. On the other hand, a positive non-link (A_3) represents record pairs that indicate exactly the different identifier or individual from files A and B. Uncertain record pairs are indicated as A_2. Linkage rule (*L*) can be defined as a mapping from Γ onto a set of random decision functions $D = \{d(\gamma)\}$, where Γ is the set of all possible realizations of γ and

$$(2)\ d(\gamma) = \{P(A_1 \mid \gamma), P(A_2 \mid \gamma), P(A_3 \mid \gamma)\};\ \gamma \in \Gamma$$

This then is $\sum_{i=1}^{3} P(A_i \mid \gamma) = 1$. In other words, the linkage rule is assigning the probabilities for taking each of the three possible actions on each observed value of γ.

Linkage Weights

The probability of agreement of values of an identifier in a positively linked file is represented by $m(\gamma)$ and is the conditional probability of γ, given that $(a,b) \in M$. That is,

$$(3) \quad \begin{aligned} m(\gamma) &= P\{\gamma[\alpha(a),\beta(b)] \mid (a,b) \in M\} \\ &= \sum_{(a,b)\in M} P\{\gamma[\alpha(a),\beta(b)]\} P\{(a,b) \mid M\} \end{aligned}$$

Similarly, the probability of agreement of values in a positively non-linked file is represented by $u(\gamma)$ and is the conditional probability of γ, given that $(a,b) \in U$. That is,

$$(4) \quad \begin{aligned} u(\gamma) &= P\{\gamma[\alpha(a),\beta(b)] \mid (a,b) \in U\} \\ &= \sum_{(a,b)\in U} P\{\gamma[\alpha(a),\beta(b)]\} P\{(a,b) \mid U\} \end{aligned}$$

Newcombe (1988) introduced a terminology called *frequency ratio* (or *odds*) that is defined as the ratio of the percentages of value-agreement frequencies in a linked file (i.e., $m(\gamma)$) to those in a non-link file (i.e., $u(\gamma)$). That is, the frequency ratio for an identifier γ is $\dfrac{m(\gamma)}{u(\gamma)}$.

The overall indication of the degree of assurance of a correct match for a given record pair is then obtained by multiplying all ratios of identifiers of a record pair. This is usually provided by first converting each of the frequency ratios into its base 2 logarithm that is called linkage weights. Therefore, the weight for an agreement of values of an identifier in a record pair is defined by $w(\gamma)^{agreed} = \log_2\left(\dfrac{m(\gamma)}{u(\gamma)}\right)$. In the same way, the weight for a disagreement of values of an identifier in a record pair is defined by $w(\gamma)^{disagreed} = \log_2\left(\dfrac{1-m(\gamma)}{1-u(\gamma)}\right)$. The sum of these weights is the total linkage weight (*overall odds* or *relative odds*). That is,

$$(5) \quad \textit{total weight for a record pair} = \sum_{\gamma} w(\gamma)$$

The base 2 logarithm is merely a tool to make the multiplication addable. The linkage weight of a value-agreement for an identifier is usually greater than that of the value-disagreed pair for an identifier.

The total linkage weight above represents the *relative* ranks of the matched pairs to one another in order of the assurance of correct links and is called *relative* odds. It is customary to convert relative odds to absolute odds to indicate the actual value of odds out of a total possible number of record pairs. The absolute total weight (*absolute odds*) for a record pair is then obtained by multiplying the odds of a random match between two files. That is,

$$(6) \quad \begin{aligned} &\textit{absolute odds} = \textit{relative odds} \\ &\times \frac{\text{number of linked records in searching file}}{\text{total number of records in searching file}} \\ &\times \frac{1}{\text{total number of records in file being searched}} \end{aligned}$$

The two fractions in the equation indicate the proportion or probability (p) of linked records out of all possible record pairs between two files A and B. This equation then can be rewritten as

$$(7) \quad \textit{total absolute linkage weight} = \sum_{\gamma} \log_2 w_\gamma \times p$$

where p is a probability of random true match between two files.

Blocking Files

Linking two files by examining all the possible pairs of records is time consuming and impractical. For example, the linkage of a file of 100 records to a file of 1000 records will create a total of 100,000 record pairs to be examined for the match-status. To avoid such a large number of record pairs, a blocking method is usually used. Blocking files is the process of dividing the files into pockets such as zip code, state, year, etc. This

method reduces the total number of all the possible record pairs because records are only compared with each other if they are from the same pockets. Estimates of $m(\gamma)$ and $u(\gamma)$ are usually obtained manually from the linked and the non-linked files that are reduced from the original files by the blocking method.

Thresholds

The frequency distributions of total weights of linked and non-linked record pairs are sorted in order to obtain upper and lower thresholds that separate all the record pairs from two files into an A_1 region for matched record pairs and an A_3 region for unmatched record pairs. The two regions have two types of record pairs from the linked file and the non-linked file, respectively. Region A_1 includes the record pairs from the linked file that are beyond the upper threshold (*a*) and wrongly identified unmatched record pairs as true matches (Type II error) from the non-linked file (*e*). Region A_3 also holds the record pairs from the non-linked file that are less than the lower threshold (*g*) and wrongly identified matched record pairs as truly unmatched (Type I error) from the linked file (*f*). The region A_2 contains all the record pairs excluded from the linked (*b*) and the non-linked files (*c*), respectively. The record pairs in this area require manual examination for their identities of matched or unmatched status. The optimum thresholds for A_1 and A_3 should be determined to maximize the correct identification of paired records from two files (*a* and *g*), minimize wrongly identified ones (*e* and *f*), and reduce record pairs that fall in uncertain areas (i.e., A_2.)

Simple Example

A simple example is provided to describe the concept that was discussed above. Let Table 1 be a collection of records of vehicles extracted from the two large files (*File A* and *File B*) to be merged. Each record consists of identifiers, ID number, make, year, and color. The identifier ID number is added in order to identify the same records easily. The same ID number exactly matches the same records from *File A* and *File B*, respectively, although it is assumed that there is an error in entering data for the purpose of this example. For instance, ID number 11 in *File A* represents a vehicle (make A, year 1997, color B) while the same ID number in *File B* represents the same vehicle with a data entrance error (i.e., the value of 1998 for year instead of true value of 1997.) Four records (ID numbers 11 through 14) from *File A* are assumed to be the same as the four corresponding records in *File B*.

The linked and non-linked files that are manually created by ID number for each identifier are shown in Table 2. All exactly

Table 1: Example of Probabilistic Linkage Method

File A (searching file)				File B (searched file)			
ID number	Make	Year	Color	ID number	Make	Year	Color
11	A	1997	B	11	A	1998	B
12	T	1999	R	12	T	1999	G
13	H	1990	R	13	H	1990	R
14	G	1999	Y	14	F	1999	Y
				15	H	1991	B
				16	T	1999	Y

Table 2: Linked, Non-linked Files and Frequency Ratios for Example

Identifiers							Agreed Ratio	Dis-agreed Ratio	$\frac{m(\gamma)}{u(\gamma)}$	$\frac{1-m(\gamma)}{1-u(\gamma)}$
Make	Linked	***AA***	***TT***	***HH***	GF		0.75	0.25	7.500	0.278
	Non-Linked	AT TA HA GA	AH TH HT GT	AH TH **HH** GH	AF TF HF GH	AT **TT** HT GT	0.10	0.90		
Year	Linked	97 98	***99 99***	***90 90***	***99 99***		0.75	0.25	3.750	0.312
	Non-Linked	97 99 99 98 90 98 99 98	97 90 99 90 90 99 ***99 99***	97 91 99 91 90 91 99 90	97 99 ***99 99*** 90 99 99 91	97 99 ***99 99*** 90 99 ***99 99***	0.20	0.80		
Color	Linked	**BB**	RG	**RR**	**YY**		0.75	0.25	5.000	0.294
	Non-Linked	BG RB RB YB	BR **RR** RG YG	**BB** RB RB YR	BY RY RY YB	BY RY RY **YY**	0.15	0.85		

agreed-values are indicated by bold-italic fonts in the table.

In the linked file for the identifier, make, these pairs indicate the same vehicles by ID number 1 through 4, indicating AA, TT, and HH. The pair of GF indicates manufacturer G and F, respectively, but G was assumed to be entered by mistake instead of F. The *Agreed* column displays the ratio of exact value-agreed pairs to the total pairs matched in the *Linked* and *Non-linked* files, respectively. For example, the value-agreed pairs (AA, TT, and HH) in the *Linked file* for the identifier, make, is 3 out of 4, or 0.75 (3/4). The ratio of *Disagreed* is simply 1-0.75, or 0.25. The last two columns display frequency ratios for each identifier for value-agreed and value-disagreed between the *Linked* and *Non-linked* files, respectively. For example, the value-agreed frequency ratio for make (i.e., $\frac{m(\text{Make})}{u(\text{Make})}$) is 0.75/0.10 or 7.5. Similarly, the value-disagreed frequency ratio for make (i.e., $\frac{1-m(\text{Make})}{1-u(\text{Make})}$) is 0.25/0.90 or 0.28.

Total weights (relative odds) for each record pair shown by ID number are listed in Table 3. These total weights are calculated by multiplying frequency ratios for each identifier for that record pair. For example, the total weight for the record pair of ID numbers 11 and 11 from files A and B is calculated by simply multiplying value-agreed frequency ratio for make, value-disagreed frequency ratio for year, and value-agreed frequency ratio for color, respectively. That is, frequency ratios for AA, 97 98, and BB (i.e., 7.500,

Table 3: Total Weights (Relative Odds) for Record Pairs, Absolute Odds, and Sorted Absolute Odds for Example

Record Pairs by ID number	Relative odds	Absolute odds	Sorted absolute odds
11 11	***11.71875***	***1.953125***	0.004255
11 12	0.025531	0.004255	0.004255
11 13	0.025531	0.004255	0.004255
11 14	0.025531	0.004255	0.004255
11 15	0.434028	0.072338	0.004255
11 16	0.025531	0.004255	0.004255
12 11	0.025531	0.004255	0.004255
12 12	***8.272059***	***1.378676***	0.004255
12 13	0.434028	0.072338	0.004255
12 14	5.208333	0.868056	0.004255
12 15	0.434028	0.072338	0.004255
12 16	8.272059	1.378676	0.004255
13 11	0.025531	0.004255	0.051062
13 12	0.025531	0.004255	0.072338
13 13	***8.272059***	***1.378676***	0.072338
13 14	0.025531	0.004255	0.072338
13 15	0.689338	0.11489	0.114890
13 16	0.025531	0.004255	***0.868056***
14 11	0.025531	0.004255	***0.868056***
14 12	0.306373	0.051062	***0.868056***
14 13	0.025531	0.004255	***1.378676***
14 14	***5.208333***	***0.868056***	***1.378676***
14 15	0.025531	0.004255	***1.378676***
14 16	5.208333	0.868056	***1.953125***

0.312, and 5.000), respectively, are multiplied, resulting in 11.719. The logarithms of each total weight and resulting addition were not made because this is a very simple calculation. Absolute odds for each record pair are also calculated by multiplying the probability of linked records out of all possible record pairs (see Equation 6.)

These absolute odds for each pair are then sorted in ascending order. Thresholds to determine A_1 and A_3, and A_2 regions are found by comparing absolute odds for exactly matched and unmatched pairs, respectively. In this example, it is obvious that absolute odds that are more than 0.8 will be in the A_1 region to include exact matched pairs. Other regions are simply determined by absolute odds that are less than 0.8. This is a very simplified example to describe the concept and procedure of this methodology. However, in a real application, there will be more blocked identifiers and number of records to be merged, resulting in significant computing time and more threshold options to choose.

APPLICATION TO COMMERCIAL MOTOR VEHICLE CARRIER FILES

The probabilistic record linkage is applied and tested by merging two inspection files described earlier: the ISP Inspection File and the MCMIS Inspection File. The MCMIS Inspection file contains data from state police inspection reports that are transmitted by SAFETYNET to MCMIS. The Illinois Inspec-tion File also contains information about carriers and inspection types. It should be noted that the MCMIS Inspection file is a subset of the ISP Inspection file, which means the contents of the MCMIS Inspection file

should be the same as the ISP Inspection file but not necessarily the exact same format. A month of 1996 files is used for creating the positively linked and non-linked files, which will provide $m(\gamma)$, $u(\gamma)$, and thresholds. These results will then be used to merge another month of the files for validation purposes.

There are two major reasons for applying this method to these two files. First, it is possible to obtain $m(\gamma)$ and $u(\gamma)$, and eventually the thresholds, accurately without a large effort. The same records from these two files are matched using the same unique identifier (i.e., Inspection Report Number.) It is possible to separate the merged file (that is, all the possible pairs between two files) into the linked and non-linked files, which allows us to obtain $m(\gamma)$ and $u(\gamma)$, total weights, and thresholds. This will avoid manual effort in obtaining these values. Second, by using the inspection report number it is possible to validate how well the probabilistic linkage merges these two files. Since matches and unmatches can be easily and accurately obtained by the inspection report number, the unavoidable errors of this methodology (i.e., wrongly identified record pairs as matched or unmatched as opposed to what they truly are) can also be easily obtained. As such, the application described in this paper is largely a methodology validation exercise.

Blocking Inspection Files

The number of records from 1996 for MCMIS and ISP Inspection files is 52,405 and 98,925, respectively. The permutation of these records creates about 5.2 billion record pairs. The blocking was done for the month of March 1996 to reduce the size of record pairs. The MCMIS Inspection file for March contains 4,497 records, while the ISP Inspection file contains 8,034 records. It is assumed again that those records in the MCMIS Inspection file should be a subset of the ISP Inspection file, resulting in matching record pairs close to 4,497.

Common Identifiers

The common identifiers between these two files are extracted from the data dictionaries. There are a total of 16 common identifiers for the MCMIS and ISP Inspection files, respectively. As described earlier, data quality issues plagued these files. Many of the identifiers in the MCMIS and ISP Inspection files have no values or data entrance. Some of these identifiers do not exist in the data files because of privacy issues (e.g., driver license number). Also some identifiers exist in one data file but not in the other file. For example, the identifier in the ISP Inspection file, state issuing Driver License, is in the ISP Inspection file but this identifier is not in MCMIS Inspection file. Driver license number, shipping paper number, and the name of the shipper are very unique identifiers that can provide a very high probability to match the same records. However, values of those identifiers were not provided for privacy reasons. Therefore, selecting useable common identifiers was based on the completeness of values of identifiers.

Nine identifiers that existed in both of the two files were selected. They are inspection report number, inspection year, inspection month, inspection day, beginning hour of inspection, beginning minute of inspection, ending hour of inspection, ending minute of inspection, and inspection levels. It should be noted that the identifier, inspection report number, is only used for separation between linked and non-linked files to get thresholds and validation of the probabilistic linkage methodology. Therefore, this identifier should not be used for this merging process. Also, identifiers such as year and month were not useable because the blocking was done for the month of March 1996 by using these two identifiers. Therefore, only six of these identifiers were used for the merging process. They are inspection day, beginning hour of inspection, beginning minute of inspection, ending hour of inspection, ending minute of inspection, and inspection levels (i.e., five levels of inspection; Level 1 – comprehensive

driver and vehicle inspection, Level 2 – comprehensive driver and vehicle walk around inspection, Level 3 – comprehensive driver inspection, Level 4 – special inspections, and Level 5 – detailed periodic vehicle inspection.) These identifiers are not unique because many vehicles could have the same values of these identifiers. However, the nature of these identifications is a good motive to evaluate the proposed merging methodology.

Total Weights and Optimal Thresholds

The values of $m(\gamma)$ and $u(\gamma)$ for the identifiers were obtained from the linked and non-linked files that were created by inspection report number as shown in Table 4. Linkage weights for the identifiers were obtained by the frequency ratios. The total linkage weights (*absolute odds*) for each possible record pairs of the linked and non-linked files were also obtained by multiplying the odds of random match. The total linkage weights were then multiplied by 10^6 for the convenience of avoiding values less than 1. There are six total weights for the linked and 65 total weights for the non-linked files as sorted in ascending order in Table 5. The number of record pairs is 4,497 for the linked file and 36,124,401 for the non-linked file.

Approach to Determine Thresholds

March 1996 threshold values in Table 6 were selected by examining frequencies of total weight values of the linked and non-linked files in Table 5 to adequately classify the matched and unmatched record pairs. For example, 1800 was selected to distinguish the frequencies of the total weight of 1849.602 from that of the total weight of 1548.865. It is obvious that 1800 distinguishes the most number of correct matched record pairs (3895 record pairs) while 1500 is marginal threshold that may be considered by adding 592 correct matched pairs but 759 record pairs to be manually reviewed. However, total weight for 700 is not optimum threshold because it only adds six more correct matched pairs while it adds 1532 record pairs to be manually reviewed.

Table 6 also presents the rates of identification of the true matches and corresponding total errors (Type I and Type II errors.) The second to the last column of Table 6 indicates the number of record pairs to be manually reviewed relative to the maximum number of pairs to be reviewed in the table (i.e., 67,077 pairs), which quantitatively shows the level of manual effort. The last column shows the dollar expenditure for the manual review of record pairs in the A_2 region.

Table 4: Values of $m(\gamma)$ and $u(\gamma)$ for March 1996 (%)

Identifiers	Agreed		Disagreed		$\frac{m(\gamma)}{u(\gamma)}$	$\frac{1-m(\gamma)}{1-u(\gamma)}$
	$m(\gamma)$	$u(\gamma)$	100-$m(\gamma)$	100-$u(\gamma)$		
Day of Inspection	99.98	3.99	0.02	96.01	25.05764	0.000208
Beginning Hour	99.86	7.00	0.14	93.00	14.26571	0.001505
Beginning Minute	99.89	6.28	0.11	93.72	15.90605	0.001174
Ending Hour	99.86	6.92	0.14	93.08	14.43064	0.001504
Ending Minute	99.77	6.55	0.23	93.45	15.23206	0.002461
Inspection Level	87.52	38.50	12.48	61.50	2.273247	0.202927

Table 5: Total Weights for Linked and Non-Linked Files

	Total Weight	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Linked	-722.082	1	0.02	1	0.02
	-421.345	2	0.04	3	0.07
	462.1706	1	0.02	4	0.09
	762.9073	6	0.13	10	0.22
	1548.8653	592	13.16	602	13.39
	1849.602	3895	86.61	4497	100
Non-Linked	-4459.09	17483031	48.39674	17483031	48.39674
(65 total weights)	-4158.36	10970284	30.36807	28453315	78.76481
	-3372.4	1189979	3.294114	29643294	82.05892
	:	:	:	:	:
	:	:	:	:	:
	-378.358	523	0.001448	36109688	99.95927
	-376.82	252	0.000698	36109940	99.95997
	92.8466	5729	0.015859	36115669	99.97583
	364.6126	984	0.002724	36116653	99.97855
	393.5833	4875	0.013495	36121528	99.99205
	407.6	79	0.000219	36121607	99.99227
	409.1377	10	2.77E-05	36121617	99.99229
	462.1706	947	0.002621	36122564	99.99491
	665.3493	731	0.002024	36123295	99.99694
	708.3367	37	0.000102	36123332	99.99704
	709.8744	11	3.05E-05	36123343	99.99707
	762.9073	719	0.00199	36124062	99.99906
	1548.865	167	0.000462	36124229	99.99952
	1849.602	172	0.000476	36124401	100

Table 6: Upper and Lower Thresholds and Resulting Frequency and Rates

UT	LT	A_1 (a+e)	A_3 (g+f)	Type I Error (f)	Type II Error (e)	A_2 (b+c)	True Match	Total Error (e+f)	Match Rate (%)	Total Error Rate (%)	Manual Effort Rate (%)	$Value
March 1996												
1800	1800	4067	36124831	602	172	0	3895	774	86.613	17.211	0.000	$0
1800	1500	4067	36124072	10	172	759	4487	182	99.778	4.047	1.132	$38
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
1800	-750	4067	36057754	0	172	67077	4497	172	100.000	3.825	100.000	$3,354
1500	1500	4826	36124072	10	339	0	4487	349	99.778	7.761	0.000	$0
1500	700	4826	36123299	4	339	773	4493	343	99.911	7.627	1.152	$39
April 1996												
1800	1800	6199	58968041	719	683	0	5516	1402	88.468	22.486	0	$0
1800	1500	6199	58966499	7	683	1542	6228	690	99.888	11.067	1.854	$77
1500	1500	7741	58966499	7	1513	0	6228	1520	99.888	24.379	0	$0
1500	700	7741	58963585	2	1513	2914	6233	1515	99.968	24.298	3.504	$146

UT: Upper Threshold; LT: Lower Threshold

Determination of the upper and lower threshold values is made by manual examination of the distributions of linked and non-linked files. This manual threshold selection is dependent on the interest of the study. For example, there can be three primary interests in determining the threshold values. One primary interest is to maximize the identification of the number of correctly matched or correctly unmatched pairs of records. This interest, however, may result in more falsely identified matched record pairs (Type II error) or false identified unmatched pairs (Type I error).

A second interest may be to minimize the number of incorrectly identified matched or unmatched record pairs (i.e., Type I and Type II errors.). However, this may reduce the number of correctly matched or correctly unmatched record pairs. Also, the number of record pairs that requires manual examination grows very sharply.

Third, the least number of record pairs to be manually examined may be the primary interest of the data linkage. Actually the least number of record pairs that requires manual examination is zero, from which only one threshold value is selected for both the upper and the lower thresholds. However, this method will result in the greatest error rate. In fact, the number of record pairs to be manually examined should be tolerated within some range of efforts because the number of erroneous record pairs can be substantially reduced. In addition, most record pairs in A_2 will be correctly identified manually depending on a clerk's skills. A manual effort may be converted to a dollar value by applying the assumed dollar cost of a clerk to review uncertain record pairs. This cost estimate for hiring a clerk may play a role in determining the optimal thresholds. For example, the cost estimates in the last column of Table 6 for hiring a clerk assume $10 an hour and shows the dollar expenditure for the manual reviews. In summary, it is necessary to find the upper and the lower thresholds that satisfy these interests. One approach is to determine the maximum number of tolerable record pairs to be manually examined first, based on resources such as a clerk and corresponding estimated dollar value to be used for this purpose. By determining this interest first, a large number of threshold ranges can be eliminated. The next step is to find which threshold ranges have the least number of Type I and II errors given the tolerable record pairs chosen to be reviewed manually.

Optimum Thresholds for Merged Inspection Files

The first part of Table 6 shows the threshold ranges and resultant frequencies, rates for true matches, total errors and record pairs to be manually reviewed for March. There are four threshold options available if the number of manual examinations is to be less than 1,000 record pairs due to resource costs. They are upper and lower thresholds of 1800 and 1800, 1800 and 1500, 1500 and 1500, and 1500 and 700, respectively.

The threshold range of 1800 and 1800 will result in the least number of identified true matches (3,895 of 4,067 record pairs or an 86.6% match rate), although there is no manual effort required for the record pair examination. This threshold range also generates a total of 774 errors (602 Type I and 172 Type II errors, respectively), resulting in about a 17.2% total error rate.

For the threshold range of 1800 and 1500, the number of record pairs is also 4,067. However, this range reduces the number of errors substantially to 182 (10 for Type I and 172 for Type II) at the cost of 759 of record pairs to be manually reviewed. Nevertheless, as many as 759 record pairs can be correctly identified manually, including 592 truly matched record pairs, which results in up to a total of 4,487 correctly identified true matches. This results in only 10 record pairs that are wrongly identified as truly unmatched records from the linked file, resulting in more than 99% of correct identification out of actual 4,497 true matched pairs in the linked file. This results in only about a 4% total error rate (10 Type I and 172 Type II errors out of 4,497 records in the linked file.)

The threshold range of 1500 and 1500 yields a total of 4,826 record pairs and zero pairs to be manually reviewed. However, this

range also yields 349 total errors (10 Type I and 339 Type II errors), resulting in 4,487 truly matched record pairs (99.8%). Although the number of truly matched records is the same as the threshold range of 1800 and 1500, the error rate in this range is almost two times more than that of the previous range (349 versus 182), resulting in a total error rate of about 8%.

The last threshold range to be considered is the range of 1500 and 700, which results in the highest frequency in identifying true matches (4,493 record pairs out of 4,826, including six manually identified pairs from A_2), resulting in more than a 99% matching rate. This range yields a total 773 record pairs to be manually examined (a 1.15% review rate), which is a little more than that of the range of 1800 and 1500. This range also generates 343 erroneous identifications (4 Type I and 339 Type II errors), which is about the same as the 1500-1500 range (about 8%).

The cost estimate in the last column also provides an insight for determining optimal threshold values. It can be seen that a sharp increase in the cost occurs from zero up to $3,354 (1800 for upper and -750 for lower thresholds, resulting in 67,077 pairs of records) as the number of uncertain record pairs increases although the range of 1800 and -750 yields 100% matching rate. This cost estimate may be considered as additional information to decision makers to determine optimal thresholds.

For the purpose of this test, though, the range of 1800 and 1500 would be the optimum thresholds primarily because of the least number of errors (182 wrongly identified record pairs), compared to the range of 1500 and 700 (343 wrongly identified pairs.) However, again, the selection of thresholds should be dependent on the interest of study. For example, if there is no resource available for the review of record pairs, the selection of 1500 for the upper and lower thresholds would make sense because this threshold still achieves high correct matched record pairs (4,487) and no record pairs to be examined. Of course, this simplification comes at the expense of the relatively high error rate (349 record pairs.)

Application of Thresholds to April 1996 Merge Procedure

This application is to validate how well pre-determined thresholds from the blocked March 1996 data identify correct record pairs and corresponding errors for merged April 1996 data. Again, the exact correct identification of match-status and errors are identified by Inspection Report Number.

To conduct the test, a very large merged file was created by combining records of MCMIS and ISP Inspection April files. The numbers of records of MCMIS and ISP April files are 6,240 and 9,451, respectively. The resultant merged file contains 58,974,240 record pairs (i.e., 6,240 x 9,451.) Out of this merged file, the number of truly matched record pairs should be 6,240 because, again, the Illinois portion of the MCMIS Inspection file is a subset of the ISP Inspection file. However, the linked file of April for the validation of the correct identification of the number of true matches for the pre-defined threshold values shows 6,235 record pairs, which is five record pairs less than 6,240. The remaining five record pairs appeared to be in the non-linked file. It is inferred that these five records in the MCMIS Inspection file may have missing values or errors in entering the inspection report number, resulting in unmatched pairs with the corresponding records in the ISP Inspection file.

The same values of $m(\gamma)$ and $u(\gamma)$ that were determined in the March data were applied to the record pairs in the merged file. Total linkage weights were also calculated by the same methodology. The values of the total linkage weights are the same as in the March merged file but frequencies for total linkage weights were different.

Correct identification of truly matched record pairs and corresponding Type I and Type II errors are shown in Table 5. The results are very similar to the performance of the selected thresholds from March-blocked data. The threshold range of 1800 and 1500 shows the very high capability for the identification of true matches and low error rates compared to other options as shown in March data. Within this range, correct

identification of the true matched record pairs is 6,228 (99.888% of actual 6,235 true matched records) including 712 manually identified record pairs from uncertain range of A_2. The total error rate is about 11% (7 from Type I and 683 from Type II errors out of actual 6,235 true matched record pairs.) The review effort is about 1.9%, which is the ratio of the number of manually reviewed pairs and the maximum number of record pairs (83,161) to be manually reviewed.

The range of 1500 and 700 provides the highest correct identifications of 6,233 (only two true matched are missed from 6,235.) However, this range generates 825 (1515 - 690) more errors and 1,372 (2914 - 1542) more record pairs to be manually reviewed, compared to the range of 1800 to 500. That is, the review effort rate is about 1.6% more than that of the range of 1800 and 500.

Sensitivity of Threshold Performance for March and April

Comparing results of correct identification, error rates, and effort rates between the March-blocked merged files and April merged file, it can be seen that the performance of these thresholds is consistent. For four threshold ranges, the order of ranks (from best to worst) are the same for these two months in the correct identification of matches. That is, the range of 1500 - 700 provides the highest correct identification, but the actual number of correctly identified difference is not great from that of 1800 - 1500 (only six more and five more correct identifications for March and April, respectively.) The order of ranks for error rate is also consistent. The range of 1800 - 1500 has the least error rates in both March and April, followed by the range of 1500 - 700. The manual effort review shows the same trend, in which the range of 1800 - 1500 has the least effort rate. The range of 1800 and 1500 has 14 less record pairs to be manually reviewed than that of the 1500 and 700 range for March, and 1,372 less for April. Therefore, it is sensible to conclude that the optimal thresholds from the blocked linked and non-linked files can be safely applied to the rest of the merged file.

DISCUSSION

The probabilistic linkage approach in merging two files in the CMVC sector has strong potential when there are no relational unique identifiers available. However, there are some issues with the approach that merit discussion. First, the number of identifiers (data entry variables) from the two files to be merged should be large enough to perform this methodology adequately. The more identifiers there are, the more accurate the total weight distribution will be. That is, there will be more total weight value threshold options that minimize manual examination requirements for record pairs in the uncertain area and related errors. The number of identifiers used in this research is only six. This result in 64 possible total weight values (i.e., 2^6=64) for paired records. One additional identifier results in a greatly increased number of options (i.e., 2^7=128) for choosing thresholds.

Second, the number of missing values of the identifiers that are actually used is an issue. While there are numerous missing values in the unique identifiers of license state and license numbers, there are very few missing values (actually zero missing values) in the more "non-unique" identifiers used in this study, so that a researcher may ignore them. The missing values in identifiers would reduce the total weights to a small value because they are eventually identified as the disagreed-valued identifiers for a record pair. For example, a missing value in beginning hour of inspec-tion in the MCMIS Inspection file will result in smaller total weight value because the value in the ISP Inspection file would not agree (i.e., no value vs. 4 PM). This may result in a smaller total weight value and force the record pair to belong to the region of non-matches, al-though they are actually truly matched.

Third, another issue is that the chosen identifiers are not very unique identifiers. Unique identifiers such as inspector ID numbers, driver license number and VIN (vehicle identification number) would increase the probability of correct matches. The common identifiers that were used in this study do not offer such detailed information.

The values, for instance, of identifiers such as start hour, start minute, inspection level, etc. may belong to many inspections. In addition, the methodology of correct identification of partially available data should be established to avoid incorrect identification. For example, an identifier of the 12-digit VIN (vehicle identification number) from file A may be incorrectly identified as disagreed-valued identifier by comparing another partial or complete 17-digit VIN identifier from the file B although they are actually the same.

Fourth, obtaining the $m(\gamma)$ and $u(\gamma)$ values is also an issue. In this research, these values were obtained easily without manual effort because the MCMIS Inspection file is a subset of the ISP Inspection file, and they have a relational unique identifier (Inspection report number) that was used for the calculation $m(\gamma)$ and $u(\gamma)$. However, probabilistic data merging is mostly used for two files without this type of unique identifier. The suggested way to obtain $m(\gamma)$ and $u(\gamma)$ was to reduce the size of two files by the blocking method. Nevertheless, these reduced record sizes of two files still generate very large merged linked and non-linked files. For example, blocked files A (100 records) and B (100 records) will still result in 10,000 record pairs for a clerk to examine for correct identifications for linked and non-linked files, which eventually are used to obtain $m(\gamma)$ and $u(\gamma)$. The automation of this procedure should be considered in future research.

CONCLUSIONS

This paper discussed and presented the application of the probabilistic linkage method to two inspection files (MCMIS and ISP Inspection files.) The feasibility of this linkage methodology is easily verified because the MCMIS Inspection file is a subset of the ISP Inspection file and there is a relational unique identifier (i.e., Inspection report number) that can be used to correctly identify true matches and related errors. In this application, total weights and resulting thresholds from the blocked linked and non-linked files (March data) were applied to other sets of files from the same sources to validate the performance of the methodology. The range of 1800 and 1500 as optimal thresholds resulted in 6,228 correct identifications of true matched record pairs out of actual 6,335 true matches (99.888%). The number of erroneous identi-fication of record pairs is 690 (about 11% of actual true matched pairs). This range also generates 1,540 record pairs to be manually reviewed for correct identification. The selection of thresholds is found to be very critical in identifying the correct match-status of record pairs. Maximizing the number of correct identifications of true matches brings up the issue of increasing the number of errors (Type II errors). Minimizing Type I and II errors also brings up the issue of increasing the number of record pairs that require manual review efforts. Minimizing the manual review efforts also raises the issue of increased number of errors. Therefore, the optimal threshold selection should be dependent on the purpose of the study and resources for manual examination of uncertain record pairs. An approach was described in this study, including a simple cost estimate for hiring a clerk to review record pairs in the uncertain region, to determine optimum thresholds to address these issues.

References

Bigelow, Wayne, Trudy Karlson, and Patricia Beutel. "Using Probabilistic Linkage to Merge Multiple Data Sources for Monitoring Population Health." *Association for Health Services Research Meetings*, June, 1999.

Dean, Michael. "Using CODES Linked Data to Evaluate Significance of Driver Medical Conditions in Crash Outcome in Utah, 1992-1996." *81st Annual Transportation Research Board Meeting*, Washington, DC, 2002.

Dunn, H. L. "Record linkage." *American Journal of Public Health* 36, (1946): 1412-1416.

Fair, Martha E. "Record Linkage in an Information age Society." *Record Linkage Techniques -1997: Proceedings of an International Workshop and Exposition.* Washington, DC, 1999.

Fellegi, Ivan P. and Alan B Sunter. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328), (1969): 1183-1210.

Finison, Karl. "Using CODES Linked Data to

Evaluate Main Crashes Involving Older Drivers," *81st Annual Transportation Research Board Meeting*, Washington, DC, 2002.

Gill, Leicester E. "OX LINK: The Oxford Medical Record Linkage System." *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition.* Washington, DC, 1999.

Johnson, Sandra. "Technical Issues Related to the Probabilistic Linkage of Population-Based Crash and Injury Data," *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition.* Washington, DC, 1999.

Newcombe, H.B., J.M. Kennedy, S.J. Axford, and A.P. James. "Automatic Linkage of Vital Records." *Science* 130 (3381), (1959): 954-959.

Newcombe, H.B. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business.* Oxford University Press, Oxford, 1988.

Thakuriah, Piyushimita , Jung-Taek Lee and Sanya Niumpridit. "Data Quality Evaluation of An Integrated Data Repository on Motor Carrier Safety." *81st Transportation Research Board Annual Meeting*, Washington, DC, 2002.

Utter, Dennis. "Use of Probabilistic Linkage for an Analysis of the Effectiveness of Safety Belts and Helmets." *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition.* Washington, DC, 1999.

Acknowledgements

The authors are solely responsible for the contents of this paper. The authors would like to thank the Bureau of Transportation Statistics, United States Department of Transportation for support for this study, BTS Grant DTTS-00-B003-IL.

Jung-Taek Lee *is a research assistant professor at the Urban Transportation Center, University of Illinois at Chicago. He has a B.S from the Dankook University, Seoul, Korea, a M.S. and a Ph.D. with a focus on the transportation engineering from the Department of Civil and Environmental Engineering, Michigan State University. His research interests include transportation safety, traffic operations, and Intelligent Transportation Systems.*

Piyushimita Thakuriah (Vonu) *is associate professor of urban planning and policy and associate director of the Urban Transportation Center, University of Illinois at Chicago (UIC). Her main research interests are Intelligent Transportation Systems, job accessibility planning and data quality. She teaches courses on transportation analysis and statistics in UIC.*