INTERNATIONAL JOURNAL OF OCCUPATIONAL
SAFETY AND ERGONOMICS (*JOSE*) 2002, VOL. 8, NO. 3, 353–363

Computer-Intensive Methods in Traffic Safety Research

Harold Stanislaw

California State University, Stanislaus, USA

The analysis of traffic safety data archives has improved markedly with the development of procedures that are heavily dependent upon computers. Three such procedures are described here. The first procedure involves using computers to assist in the identification and correction of invalid data. The second procedure makes greater computational demands, and involves using computerized algorithms to fill in the ''gaps'' that typically occur in archival data when information regarding key variables is not available. The third and most computer-intensive procedure involves using data mining techniques to search archives for interesting and important relationships between variables. These procedures are illustrated using examples from data archives that describe the characteristics of traffic accidents in the USA and Australia.

data analysis errors screening missing values unknown values imputation data mining

1. INTRODUCTION

Traffic safety research often involves the analysis of extensive data archives. The data archives typically document motor vehicle accidents and variables that characterize them, such as the age and gender of each person involved in the accident, the number of cars, motorcycles, trucks, and pedestrians involved in the accident, the date and time of the accident, and

Portions of this paper were presented at the 5th International Conference on Computer-Aided Ergonomics and Safety, in Maui, HI, 2001.

Correspondence and requests for offprints should be sent to Harold Stanislaw, Department of Psychology, California State University, Stanislaus, Turlock, CA 95382, USA. E-mail: <hstanisl@athena.csustan.edu>.

the lighting and weather conditions. Such archives are commonly used to identify types of drivers, vehicles, or roadways that seem to be involved in an unusually high number of accidents. Data archives are also used to examine the effects of economic, legislative, and environmental factors on accident frequency.

In order to analyze these data archives efficiently and thoroughly, researchers must increasingly rely upon computer-intensive methods. The present paper describes three such methods. The first and simplest of these methods is data screening, which involves identifying invalid data entries, so they can be corrected or deleted. The second method is imputation, which is more complex than data screening and involves determining values for missing data. The third and most sophisticated method is data mining, which is used to search for interesting and important relationships in an archive.

Each method is illustrated with examples that involve the analysis of three large data archives. These archives are summarized in Table 1. The Fatal Accident Reporting System (FARS) and the General Estimates System (GES) describe motor vehicle accidents that occurred in the USA. The third archive describes motor vehicle accidents occurring in the Australian state of New South Wales (NSW). The FARS archive describes all accidents involving a fatality, but excludes non-fatal accidents. The GES archive includes non-fatal accidents, but describes only a randomly selected portion of accidents occurring in the USA. The NSW archive describes all accidents, regardless of their severity. The NSW archive is not publicly available, but the FARS and GES archives are both available online (ftp://ftp.nhtsa.dot.gov).

TABLE 1. Characteristics of the Fatal Accident Reporting System (FARS), General Estimates System (GES), and New South Wales (NSW) Traffic Safety Data Archives

Characteristic	FARS	GES	NSW
Country	USA	USA	Australia
Severity of included accidents	Fatal	Major property damage, injury, fatal	Major property damage, injury, fatal
Number of included accidents	All	Random selection	All
Variables per accident (minimum)	181	147	99
Approximate accidents per year	35,000	54,000	50,000
Years included in archive	1975–present	1988–present	1976–1992
Number of data entries	> 160 million	> 100 million	> 120 million

2. DATA SCREENING

Traffic safety data archives frequently contain errors. Typographical mistakes made during data entry account for some of these errors, but other factors contribute as well. Accident reports are completed by police officers and other personnel who sometimes have little understanding of, or interest in, the subsequent analysis of the data they provide. These individuals may assign a low priority to creating reports that are accurate, complete, and legible. Compounding this is the problem that reports are often written under suboptimal conditions (e.g., in a cramped vehicle alongside a road at night, or at a police station several hours after leaving the accident scene). Additional problems arise because data entry operators are rarely involved in the actual analysis of data, and thus may not fully appreciate the need for high quality data. Indeed, many traffic safety agencies routinely assign data entry to outside contractors.

Clearly, data archives should be screened for errors before they are analyzed. One possibility is to check for entries that fall outside the range of allowable or plausible values. For example, the NSW data archive lists 29 accidents that occurred in 1986 and involved vehicle operators with blood alcohol concentrations (BACs) of 0.998 gm/dl. These values are impossibly high; BACs exceeding 0.3 gm/dl are usually lethal (Garriott, 1983). The suspect BACs in the NSW archive probably resulted when data entry operators attempted to key in 0.999, which is the numerical code used for unknown BACs. The NSW archive also lists 36 accidents that occurred between 2 and 3 a.m. on days when the introduction of daylight savings time caused the 2 a.m. hour to be skipped. Presumably, these errors resulted from improperly completed accident forms, rather than typographical errors. Invalid or implausible data entries are generally easy to identify, and data entry programs can be written to automatically reject attempts to enter out of range values. Data entry operators can then re-enter the data (if a typographical error was made), examine the accident report more closely (if an illegible entry was encountered), enter a missing value (if the accident report contains an invalid entry), or otherwise correct the entry (such as adding an hour to the time of the accident for the daylight savings time error).

A more sophisticated form of data screening involves comparing several different entries, to ensure that the data are internally consistent. For example, the NSW archive often lists a vehicle operator's age and gender twice: once when describing the vehicles that are involved in accidents, and

again when describing the people who are injured in accidents. The NSW archive lists 44 accidents in which the two age and gender entries conflict with each other. Another 29 accidents describe an adult as using a child restraint. In most of these latter accidents, a child in the same vehicle as the adult is described as using an adult seat belt, suggesting that the data for the adult and the child were reversed.

In contrast to out of range data entries, it is extremely difficult to prevent operators from keying in errors that can be detected only by examining several data entries simultaneously. Thus, the errors tend to surface only after the data have been entered. Because data entry is often performed off-site, the original accident reports cannot be retrieved to correct the erroneous entries; instead, the erroneous entries must be deleted.

Combinations of data entries that are not particularly unusual when considered individually may be highly suspect when they are considered in the context of the entire archive. For example, the NSW archive sometimes lists two BAC measurements for a vehicle operator. One measurement is made at the scene of the accident, whereas the second is made when the operator is hospitalized. The two readings need not match precisely, as they are obtained at different points in time, and usually rely upon different technologies (breath testing at the scene of the accident, and blood sampling at the hospital). Even so, the two measurements should agree reasonably well. In most years, this is the case; the correlation between the two measurements is typically .75 after 1983. However, the correlation is much lower in earlier years, averaging only .05 before 1980. This suggests that many of the BAC entries prior to 1983 are incorrect. One possibility is that a programming error was made in 1986, when the archive format was changed and entries from earlier years were revised to match the new format.

Few data analysis programs can be configured to automatically detect errors that are evident only from a simultaneous examination of several entries. To detect these sorts of errors, analysts must determine which combinations of data are impossible or logically implausible, and then write their own computer programs to search for these combinations. An example of this is provided by a procedure used to verify the accident times listed in the NSW archive. The NSW archive originally recorded accident times using a 12-hr clock, with a.m. or p.m. entered in a separate data field. This is the format that most Australians use to describe time. In 1986, the archive switched to the 24-hr clock commonly used throughout Europe and in military settings. This switch might be expected to have triggered some

confusion, with data entry operators occasionally forgetting to add 12 hrs to the time listed on an accident report, when the accident occurred after noon.

A FORTRAN program, summarized in Figure 1, checks for this error. The check is not made for all accidents; if the accident is listed as occurring at 1300 hrs or later, the accident time was clearly entered using a 24-hr clock. However, for accident times before 1300 hrs, the FORTRAN program examines a data field that lists the natural lighting conditions at the time of the accident. This field provides a rough indicator of the time of the accident, as the possible entries are dawn, daylight, dusk, and darkness. The program then uses the date and location of the accident to estimate the position of the sun at the accident time listed in the archive. A comparison is then made between the sun position and the lighting condition listed in the archive. When a discrepancy is found (e.g., the accident occurred 3 hrs after sunrise but the lighting is described as "darkness"), the program determines if adding 12 hrs to the entry for the accident time would eliminate the discrepancy. If so, the program checks one additional data

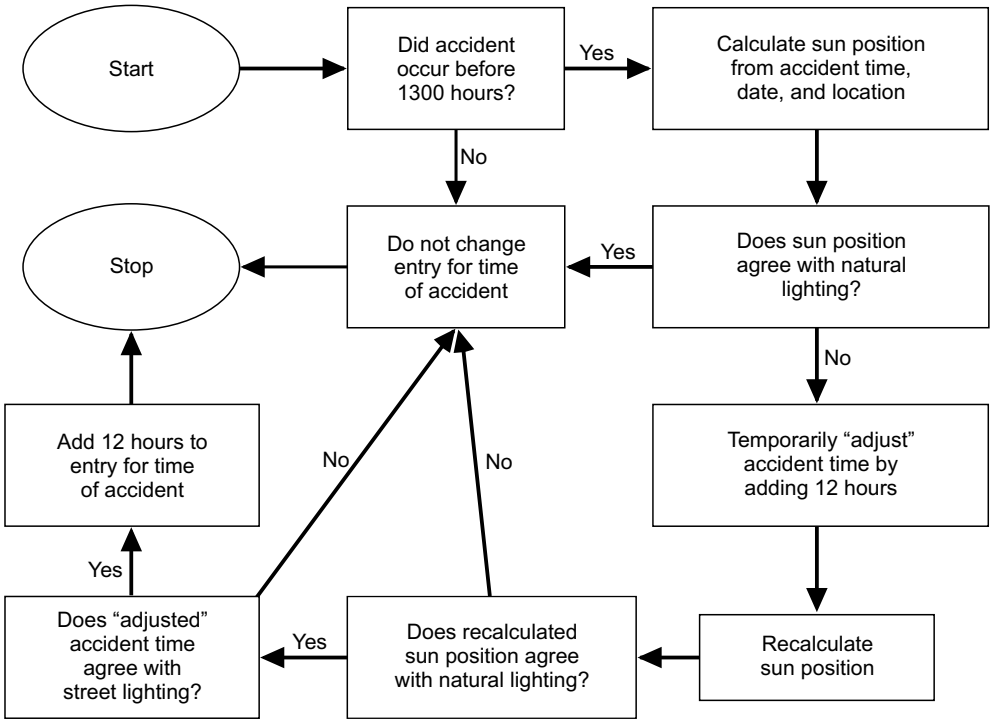

Figure 1. Flow chart for determining whether or not to add 12 hrs to accident times in the New South Wales, Australia, archive.

field before it actually changes the entry for the time of the accident: The status of the street lights (i.e., whether the lights are on or off) must be inconsistent with the accident time entered in the archive, but consistent with the time that results from adding 12 hrs to the entered accident time. This check is made to ensure that the most probable error in the data is the accident time, rather than the lighting conditions at the time of the accident.

The FORTRAN program summarized in Figure 1 found 3,185 accident times in the NSW archive that conflicted with the lighting information but could be "corrected" by adding 12 hrs to the listed time of the accident (i.e., by assuming that the accidents occurred after noon, rather than before noon). On average, 562 corrections were required per year after the switch to the 24-hr clock, compared with only 32 corrections per year before the switch. This supports the notion that adoption of the 24-hr clock triggered a large number of data entry errors.

3. IMPUTATION

The preceding example demonstrates that questionable data entries can sometimes be replaced with values that are likely to be more accurate. However, in most cases data analysts cannot reasonably determine the values they should use when incorrect entries are found; the only option is to delete the suspect data entries, thereby adding missing values to the archive. Missing values also arise when accident reports are incomplete, or when a vehicle operator leaves the scene of an accident before the police arrive to gather data.

Missing values may not be of concern if the archive is extensive, so that adequate statistical power remains after incomplete records are excluded from the analysis. However, even if the statistical consequences are negligible, missing values can have dramatic effects on the validity of research findings. For example, the FARS archive lists BACs for only about one quarter of the vehicle operators who survive an accident. Little confidence can be placed in research that excludes three quarters of the available data.

BAC data are more comprehensive in NSW, where BAC testing is legally mandated whenever an accident occurs. However, even in the NSW archive approximately 20% of the vehicle operators have unknown BACs. Furthermore, evidence from other data fields suggests that operators with unknown BACs are approximately 4 times more likely to be intoxicated at the time of the accident than operators with known BACs (Stanislaw, 1996).

This suggests that, in many accidents with missing BACs, the operator was intoxicated and left the accident scene to avoid detection. Failure to account for this bias during data analysis would lead to severe underestimation of the frequency of alcohol involvement in traffic accidents. Thus, analysts should not simply ignore records with missing values.

Numerous approaches exist for the treatment of missing values (Little & Rubin, 1987). The approach that will be discussed here is imputation, which involves replacing missing values with entries that are thought to be reasonably accurate estimates of the values that are missing.

A variety of methods are used to impute values for the GES archive (Shelton, 1993). One method, called univariate imputation, involves using records with non-missing values to provide estimates for records with missing values. For example, if 60% of the vehicle operators whose gender is known are male and 40% are female, 60% of the operators whose gender is not known are selected at random and assumed to be male. The remaining operators with missing gender information are assumed to be female. Because the operators who are assigned a particular gender are selected at random, this imputation technique is said to be stochastic.

Hot decking is a modified version of univariate imputation. This technique involves identifying groups of records in the archive that share identical values for several of the non-missing variables, including the year of the accident. The criteria that are used to define each group are selected in an iterative manner to ensure that the groups are as homogeneous as possible. Within each group, the complete records are used to impute estimates for the missing values, using the same stochastic procedures that are employed for univariate imputation. Cold decking is similar to hot decking, but uses complete records from previous years to stochastically impute missing values for accidents in subsequent years.

Stochastic imputation helps to ensure the stability of relative frequency counts. For example, the ratio of male to female operators is about the same before and after stochastic imputation. This is desirable, but stochastic imputation also introduces errors into the data. For example, suppose that hot decking suggests 60% of the vehicle operators whose gender is missing are males, and that this percentage is, in fact, correct. With stochastic imputation, 60% of the male operators will be correctly imputed to be male; the remaining 40% of males will be imputed to be females. Similarly, 40% of the female operators will have their gender correctly imputed. Overall, only 52% of the operators ($60\% \times 60\% + 40\% \times 40\%$) will have their gender correctly identified. More operators would have their gender correctly

imputed if they were all assumed to be male; this non-stochastic strategy would correctly identify gender in 60% of the operators (60% × 100% + 40% × 0%). However, the non-stochastic strategy overestimates the number of male operators, and underestimates the number of female operators.

In light of these problems, it should be no surprise that imputed GES data fail to pass tests that are regularly used to screen the GES archive for data entry errors (Shelton, 1993). For example, imputation of the GES age data produces a suspiciously large number of young truck drivers.

A different approach is used to impute BAC data in the FARS archive (Klein, 1986). The goal here is to determine the probability that the vehicle operator's BAC is 0.10 mg/dl or greater (above the legal limit for intoxication in most jurisdictions), the probability that the operator's BAC is between 0.01 and 0.09 mg/dl, and the probability that the operator's BAC is 0. If the BAC is known, a 1 is entered for the "probability" corresponding to the operator's actual level, and a 0 is entered for the two remaining levels. However, if the BAC is not known, a discriminant function is used in conjunction with other, non-missing variables to estimate probabilities for each of the three BAC levels. This approach seems to provide a good solution to the problems introduced by using stochastic imputation techniques.

In order to avoid tautologies, variables that are used to impute missing values should not enter into analyses that involve imputed data. For example, age should not be used to impute missing BAC values in studies that investigate age-related trends in accidents that involve alcohol. Thus, researchers may wish to develop several different strategies for imputing the same variable, and vary the strategy according to the analysis being conducted.

All imputation techniques risk introducing erroneous data into the archive. Thus, it may be desirable to conduct analyses of the variables that are candidates for generating imputed values, rather than the imputed values themselves (Stanislaw, 1998). This approach was used to study the impact of random breath testing on alcohol-related accident rates in NSW (Stanislaw, 1996). BACs were unknown for a large number of vehicle operators. However, analyses of the known BACs suggested that operators were rarely intoxicated in daytime accidents, but were frequently intoxicated in late evening accidents. Accidents were therefore divided into those that occurred at times of day when alcohol was likely to be involved, and those that occurred at times of day when alcohol was unlikely to be involved. The analysis then focused on the time of the accident (which was almost always known), rather than the operator's BAC (which was often missing).

4. DATA MINING

Traffic safety archives are rich sources of information. However, their extensive size and the large number of variables they track can actually hinder the discovery of important relationships. Data mining techniques are computer-intensive procedures that have been developed to help circumvent this problem.

An overview of data mining techniques is provided by Westphal and Blaxton (1998). Several of the techniques focus on multivariate graphical presentation. The analyst can use these techniques to visualize relationships between several variables simultaneously and search for patterns of interest. Other techniques attempt to automate the search for interesting data patterns.

A common goal in many studies is to develop a method for predicting a variable of interest. If this variable is continuous with a range of possible values, analysts can develop prediction models by using such familiar analytic techniques as stepwise multiple regression. The prediction of categorical dependent variables, such as whether or not a vehicle operator is intoxicated, requires different analytic techniques. One possibility is to present the computer with examples of operators who are and who are not intoxicated. These examples can be used to "train" a neural network to determine, on the basis of other variables in the archive, which drivers are most likely to be intoxicated. Alternately, examples of intoxicated and non-intoxicated operators can be used to derive a decision tree that describes a series of binary decisions that can be used to predict intoxication. Decision tree methodology is described in texts such as Breiman, Friedman, Olshen, and Stone (1984). Two popular software packages that implement the methods are CART (www.salfordsystems.com) and AnswerTree (spss.com/answertree). A summary of decision tree software can be found at www.recursive-partitioning.com, which also lists user comments and ratings.

Data mining techniques have not yet been widely used to analyze traffic safety archives. One potentially useful application of data mining is in deriving methods for imputing missing values. For example, when BAC measures are not available, a large number of other variables can be examined to determine the likelihood that the accident involved alcohol (see Table 2). One of the best predictors of alcohol involvement is the time of the accident (Stanislaw, 1996, 1998): In the NSW archive, alcohol is involved in about 30% of the accidents that occur between midnight and

4 a.m., but fewer than 1% of the accidents that occur between 8 a.m. and noon. Decision trees can be developed to determine how best to combine information about the time of the accident with other predictors of alcohol involvement, such as the number of vehicles involved in the accident and the licensing status of the operators. Data mining techniques can also automatically determine the particular hours of the day in which alcohol involvement is extremely likely, the hours in which alcohol involvement is extremely unlikely, and the hours that have intermediate levels of alcohol involvement.

TABLE 2. Variables That Predict Alcohol Involvement in Motor Vehicle Accidents (Stanislaw, 1998)

Variable	Predictive accuracy (%)
Time of accident	86
Licensing status of vehicle operator	75
Number of vehicles involved	71
Gender of vehicle operator	66
Accident severity	61
Age of motor vehicle operator	58

Conventional analytic approaches tend to be confirmatory in nature, and are used to test specific hypotheses that are usually specified before the data are examined. Data mining techniques, by contrast, are exploratory. They are frequently atheoretical, and seek relationships within an archive that are of interest but that would otherwise be difficult or impossible to discover. Because the techniques are exploratory, any findings that emerge should be considered tentative. Stepwise multiple regression, for example, is notorious for "discovering" relationships that do not accurately reflect the true nature of the data (e.g., Derksen & Keselman, 1992; Thompson, 1995).

In light of these problems, findings made by data mining techniques should always be cross validated in subsequent, confirmatory analyses. Data mining software packages often conduct such analyses automatically. Traffic safety data archives are particularly well suited for cross validation. Most archives are so large that a substantial amount of data can be set aside for cross validation purposes, without compromising the exploratory analysis.

REFERENCES

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Pacific Grove, CA, USA: Wadsworth.

Derksen, S., & Keselman, H.J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45*, 265–282.

Garriott, J.C. (1983). Forensic aspects of ethyl alcohol. *Clinics in Laboratory Medicine, 3*, 385–396.

Klein, T.M. (1986). *A method for estimating posterior BAC distributions for persons involved in fatal traffic accidents* (NHTSA Technical Report No. DOT HS 807 094). Washington, DC, USA: U.S. Government Printing Office.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York, NY, USA: Wiley.

Shelton, T.S.T. (1993). *Imputation in the NASS General Estimates System* (NHTSA Technical Report No. DOT HS 807 985). Washington, DC, USA: U.S. Government Printing Office.

Stanislaw, H. (1996). *Long-term effects of the random breath testing program on drink-driving crashes in New South Wales* (Research Note No. RN 1/96). Sydney, Australia: Roads and Traffic Authority of New South Wales, Road Safety and Traffic Management Directorate.

Stanislaw, H. (1998). *Surrogate measures of alcohol involvement in traffic accidents: Better than blood or breath tests?* Paper presented at the 24th International Congress of Applied Psychology, San Francisco, CA, USA.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*, 525–534.

Westphal, C., & Blaxton, T. (1998). *Data mining solutions*. New York, NY, USA: Wiley.