US Department
of Transportation

National Highway
Traffic Safety
Administration

DOT HS 807 094
Final Report

July 1986

A Method for Estimating Posterior BAC Distributions For Persons Involved in Fatal Traffic Accidents

This document is available to the public from the National Technical Information Service, Springfield, Virginia 22161.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear only because they are considered essential to the object of this report.

Technical Report Documentation Page

1. Report No.
DOT HS 807 094

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
A Method for Estimating Posterior BAC Distributions
For Persons Involved in Fatal Traffic Accidents

5. Report Date
July 1986

6. Performing Organization Code
NRD-31

7. Author(s)
Terry M. Klein

8. Performing Organization Report No.

9. Performing Organization Name and Address
Sigmastat, Inc.
7015 Palamar Terrace
Seabrook, MD 20706

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DTNH22-86-P-07202

12. Sponsoring Agency Name and Address
U.S. Dept. of Transportation
National Highway Traffic Safety Administration
400 Seventh Street, SW
Washington, DC 20590

13. Type of Report and Period Covered

14. Sponsoring Agency Code
NHTSA

15. Supplementary Notes

16. Abstract

A new method is proposed for estimating BAC distributions for persons with unknown BAC test results on the Fatal Accident Reporting System (FARS) files. The method utilizes discriminant analysis to form linear combinations of variables associated with alcohol involvement in drivers and nonoccupants, and uses these linear functions to estimate posterior BAC distributions based on various person, vehicle and accident attributes. Accident-level BAC distributions can be computed directly from the person-level BACs as the joint probability distribution of all drivers and nonoccupants involved in each accident.

The FARS database of drivers and nonoccupants is stratified by vehicle body type, known vs. unknown police-reported alcohol involvement, and driver age, resulting in twenty separate model strata for estimating BAC distributions. Variables found to be most useful in estimating BAC are: police-reported alcohol involvement, accident hour, person age, vehicle role, injury severity, weekday/weekend, use of occupant restraint, driver license status, number of entries on driver record, person sex, location of nonoccupant in relation to roadway, and whether or not the driver could drink legally (minimum drinking age in accident state).

Validation tests are conducted using cases with known BAC test results from the 1984 and 1985 FARS. Results of these tests are presented and discussed.

17. Key Words
Blood Alcohol Concentration, BAC,
alcohol, discriminant analysis, FARS

18. Distribution Statement

19. Security Classif. (of this report)

20. Security Classif. (of this page)

21. No. of Pages
49

22. Price

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Table No. Page No.

Appendix Tables

A METHOD FOR ESTIMATING POSTERIOR BAC DISTRIBUTIONS FOR PERSONS INVOLVED IN FATAL TRAFFIC ACCIDENTS

by Terry M. Klein

INTRODUCTION

It is widely recognized that alcohol is a major contributing factor in the occurrence of traffic accidents and the resulting severity of injuries. Alcohol has been found to be more prevalent in fatal accidents than in personal injury and property-damage-only accidents, and is more closely associated with nighttime than daytime fatal crashes, presumably due to the drinking habits of American society, in which most drinking occurs during the evening hours. In the past, various methods of estimating alcohol involvement in fatal accidents have been proposed; these estimates have been based on a variety of data ranging from small scale in-depth accident investigations, to more recent analyses of large statewide and national databases.

Most previous methods of estimating alcohol involvement have been cross-sectional in nature; that is, a method of estimation was proposed and applied to a single year of data without considering the possibility of tracking alcohol involvement over time. In addition, most if not all methods were aimed at estimating alcohol involvement in the aggregate, producing estimates over a restricted range of associated variables of interest, such as by time of day, driver age, etc.

The need remained for a method of estimating alcohol involvement in fatal accidents which could be applied consistently over time, is amenable to automated computation of estimates, and produces estimates that can be disaggregated across a wide range of variables of interest. The benefits of such an estimation procedure would be the ability to track changes in alcohol involvement over time, and the ability to focus on a large number of associated variables with a view toward examining behavioral relationships and hypothesizing and evaluating the effectiveness of drunk-driving countermeasures.

The development and implementation of this modeling effort was conducted by the author as a consultant with Lawrence Johnson & Associates, Inc., through a contract funded by the National Highway Traffic Safety Administration.

SOME PREVIOUS MODELS FOR ESTIMATING ALCOHOL INVOLVEMENT

The development of a new approach for estimating alcohol involvement in fatal accidents began with a review of previous efforts, from which a number of key points were identified.

One of the earlier methods of estimating alcohol involvement in fatal accidents utilized a sample of fifteen states for which a high percentage of fatally-injured drivers had known Blood Alcohol Concentration (BAC) test results on the Fatal Accident Reporting System (FARS) (1). The fifteen states are: California, Colorado, Delaware, Hawaii, Nevada, New Hampshire, New Jersey, New Mexico, Oregon, Rhode Island, Vermont, Virginia, Washington, Wisconsin and the District of Columbia. BAC test results for fatally-injured drivers were used to produce relative BAC distributions. These distributions were applied to national counts of fatally-injured drivers, inflating the fifteen-state experience to produce national estimates of alcohol involvement in fatally-injured drivers. This approach had several shortcomings: (1) estimates of alcohol involvement were available only for fatally-injured drivers, (2) estimates of alcohol involvement at the accident level were not readily available, (3) the sample was heavily weighted by the fatally injured drivers from a single state (fatally-injured drivers in California accounted for forty-five percent of the sample), and (4) the sample of states changes over time. As the rate of BAC testing has increased over time, the sample of states providing high BAC reporting has increased to approximately thirty. This poses certain problems in that it would be desirable to make maximum use of the high reporting states, yet estimates over time can be based only on those states with historically high reporting. As BAC reporting increases, the additional new information must be ignored; the new estimates would not be comparable with the previous estimates, since they would be based on different groups of states. Thus, this approach would not be able to capitalize on increases in BAC reporting, even if all drivers were eventually tested for BAC.

A second method of estimating the prevalence of alcohol in fatal accidents involved the partitioning of all drivers with known BAC test results into a large number of mutually exclusive cells, the structure of which was determined by combinations of variables associated with alcohol involvement, such as time of day, driver age and sex, accident type (single- or multiple-vehicle or nonoccupant accident), etc. (2). The underlying assumptions of this approach were that, within each cell, the driver BAC distribution for the known cases was representative of those drivers on FARS with unknown BAC, and that the cell estimates could be inflated from the known drivers to all drivers, producing national estimates. This approach also suffered from several problems: (1) accident-level estimates could not be developed from these results (except for single-vehicle accidents, in which there was only one driver and no nonoccupants), (2) the estimation procedure is computationally cumbersome, and not particularly amenable to automated classification of new cases with unknown BAC, and (3) the assumption that drivers with unknown BAC were similar, with respect to BAC, to those drivers with known BAC was rather strong, and not necessarily supported by the police-reported alcohol involvement variable presented in Table 1.

Table 1
Police-reported Alcohol Involvement
Known BAC Cases vs. Unknown BAC Cases
FARS 1982

	No. of Cases	Not Alcohol Involved	Alcohol Involved	Not Reported	Pol.-Rep. Unknown
Known BAC	18,489	26%	45%	9%	21%
Unknown BAC	37,540	55%	14%	22%	9%

As can be seen in Table 1, the police-reported alcohol involvement variable indicates that the majority of drivers with unknown BAC were thought to be not alcohol involved, providing little evidence to support, at least in the aggregate, the proportional allocation of unknown BAC cases according to the distribution of known BAC cases.

A third approach to estimating alcohol involvement utilized discriminant analysis to develop linear functions of variables associated with alcohol involvement for classifying accidents into one of two BAC groups (BAC<0.05, BAC≥0.05) (3). Discriminant analysis is a multivariate statistical technique for estimating linear functions of variables, and using these linear functions, on a case-by-case basis, to calculate the (posterior) probability that the case "belongs" to each of several mutually exclusive groups. While this approach seemed to be the most promising for developing estimates of alcohol involvement and for classifying new cases with unknown BAC, the initial effort fell short on several counts: (1) the selected BAC groups did not permit estimation of accidents involving totally sober (BAC=0.00) or legally drunk (BAC≥0.10) drivers, (2) no estimates were available at the person level (driver or nonoccupant), such as driver alcohol involvement by age or sex, and (3) one set of linear classification functions was developed and applied to all accidents as one homogeneous group, regardless of the participants involved.

The current effort attempted to incorporate the finer points of the previous models while enhancing the utility of the resulting estimates. A modified version of the previous linear discriminant analysis was selected as the methodology to be employed. While the previous effort aimed at producing accident-level estimates, the current application focused on estimating BAC distributions at the person level (drivers and nonoccupants).

MODELING CONSIDERATIONS

In designing an approach to estimating the prevalence of alcohol in fatal accidents, a number of constraints, or requirements, were imposed and a number of issues required resolution. A first requirement specified that the derived models have the ability to generate estimates of alcohol involvement within defined BAC levels. This involved the selection of cutpoints for BAC groups, across which classification functions, and hence probabilities, were to be estimated. A high BAC group was defined as a BAC greater than or equal to 0.10, since this is generally considered to be the level of illegal intoxication while driving in the U.S., and would provide a measure of compliance with the intent of the law. A second level to be defined was a BAC of 0.00, that is, no presence of alcohol in the blood. The remaining group, $0.01 \leq BAC \leq 0.09$, represents the population of drivers who drink, but remain within the legal limits. For each driver and nonoccupant with unknown BAC on the FARS, the model will estimate three numbers: the probabilities that the person had BAC in each of the three groups.

A second requirement was that BAC distributions be developed for each person on the FARS file, who was "actively involved" in a fatal accident. Specifically, each driver and nonoccupant with unknown BAC on FARS should have his/her own BAC probability distribution (the three numbers to be estimated, representing the probability that the person had BAC=0.00, the probability that the person had BAC in the 0.01-0.09 range, and the probability that the person had $BAC \geq 0.10$). The previous discriminant analysis focused on classifying accidents into BAC groups, which provided no estimated BAC distributions for any persons involved. The new procedure, producing an estimated BAC distribution for each person, permits the investigation of alcohol involvement over the complete set of person-level characteristics (e.g., age, sex, prior violations, etc.) using the BAC probability distribution as weights (relative frequencies) within each respective BAC group.

Another important issue to be considered was the selection of the set of known BAC data to be used in the modeling effort. For a number of years, one of the more closely-followed subsets of FARS data has been the set of states that have had a consistently high level of BAC reporting for fatally-injured drivers; at least eighty-five percent of the fatally-injured drivers were tested and have known BAC test results on the FARS file since 1982. The fifteen states used in 1982 were: California, Colorado, Delaware, Hawaii, Nevada, New Hampshire, New Jersey, New Mexico, Oregon, Rhode Island, Vermont, Virginia, Washington, Wisconsin and the District of Columbia. The major reason for using a subset of states with consistently high reporting was to avoid selection bias. This occurs, for example, when police choose to test only those drivers suspected of being legally drunk, and hence, yielding inflated estimates of driver alcohol involvement. Over time, BAC testing rates have increased, and the number of states currently testing at least eighty-five percent of the fatally-injured drivers is approximately thirty. Thus the set of "good" states is constantly changing.

The alternative to using the fifteen-state sample was to use all known BAC test results on the FARS file. The high rate of testing (of fatally-injured drivers) in the fifteen states avoids the criticism of selection bias, which proposes that states with lower rates of testing generally select the more drunken drivers for testing, resulting in biased estimates of the true rate of alcohol involvement in fatal accidents. This argument, which was probably more appropriate in the early days of BAC testing, was not supported by the data presented in Table 2, that compares BAC distributions for fatally-injured drivers in the fifteen states with those in all fifty states (plus the District of Columbia).

Table 2
Comparison of Known BAC Distributions of Fatally-Injured Drivers
Fifteen High-reporting States vs. Fifty States
FARS 1982

	No. of Cases	BAC = 0.00	0.01-0.09	BAC > 0.10
50 States	13,396	38%	11%	52%
15 States	5,137	41%	11%	48%

There are only small differences between the BAC distributions of the two samples. It should be noted that the fifteen states account for thirty-eight percent of all fatally-injured drivers with known BAC cases on the FARS. In addition, there is nothing "sacred" about the fifteen states with high BAC testing rates; there is no reason why all states should have the same driver BAC distributions. Some of the difference between the fifteen and fifty state samples might be attributable to actual differences in driver BAC distributions among the states.

In addition to the similarity of the BAC distributions of fatally-injured drivers, the testing of surviving drivers in the fifteen states is not nearly so high as for fatally-injured drivers, and thus, any advantage in using only the fifteen state sample for developing estimates for surviving drivers is diminished.

Lastly, since estimates were to be developed for drivers with unknown BAC in ALL states, using only data from the fifteen-state sample would be "throwing away" valuable data regarding the remaining thirty-five states.

Based on this, it was decided that all cases with known BAC test results would be used for the modeling effort, including fatally-injured and surviving drivers in all fifty states plus the District of Columbia. This is an important point to consider because the BAC distributions found in the known set of cases form the basis for estimating the "prior" probabilities of each case "belonging" to each of the three BAC groups. One distinguishes here between "prior" probability (the probability that a driver picked at random from the population (of drivers on FARS) belongs to a specific BAC group, without knowing any additional information, i.e., before the fact) and "posterior" probability (the probability that a driver picked at random from the population (of drivers on FARS) has a specific BAC, after having observed the various accident, vehicle, and person characteristics, i.e., after the fact). For example, a driver on FARS has a prior probability of twenty-eight percent of having a BAC$\geq$0.10; however, if this accident were known to have occurred between midnight and 6 a.m., all other things remaining equal, the posterior probability of a driver on FARS having BAC$\geq$0.10 would be fifty-three percent.

The statistical technique selected to estimate BAC probability distributions was discriminant analysis. Discriminant analysis is a multivariate statistical technique with two general goals: (1) separating mutually exclusive sets of objects or observations, and (2) allocating or classifying new observations to previously defined groups. The first goal, separation, is exploratory in nature, and is often employed to investigate observed differences among groups of objects when relationships are not well understood. One attempts to find "discriminants" (variables) whose values are such that the groups or collections of objects are separated as much as possible. The second goal, classification, is less exploratory in nature in the sense that classification procedures generally result in well-defined rules to optimally assign a new object to one of several predefined groups. With this goal in mind, it is clear that classification generally requires more problem structure than does separation. In practice, however, these two goals frequently overlap, and many times, the distinction between discrimination and classification becomes blurred (4).

In developing rules to classify new objects into predefined groups, it may be that one group has a greater likelihood of occurrence than the other group(s) because one of the populations is relatively much larger than the other(s). For example, a randomly-selected person would be very unlikely to have a rare disease (by definition). An optimal classification rule should take these "prior" probabilities of occurrence into account, and should classify this randomly-selected person as healthy unless the diagnostic data overwhelmingly indicate to the contrary.

On the other hand, another aspect of classification is cost, that is, the cost of misclassifying an object. In the example above, while the prior probability of having the specific rare disease is low, this can be offset by assigning a relatively high cost associated with misclassifying the person as healthy when he/she really has the disease. This misclassification is clearly more costly than concluding that the disease is present when, in fact, it is not, since indicating a potential health problem will invariably lead to further diagnostic tests that will eventually reveal the true situation. On the other hand, concluding that the person is healthy will leave the disease undetected. An optimal classification procedure should, when possible, account for misclassification costs.

In the current effort to estimate BAC probability distributions, the costs of misclassification have been assumed to be equal. The goals of separation and classification lose their distinction in the case of equal misclassification costs and equal prior probabilities, since the derived allocation rules involve functions designed to maximally separate populations. However, for this analysis the prior probabilities have been estimated using the distributions of known BAC data on the FARS files; this practice is generally referred to as the use of "proportional priors". Thus, this effort remains in the realm of classification/allocation.

A good classification rule should result in few misclassifications. For example, one should be critical of a derived classification rule that correctly allocates new objects into one of two possible groups only fifty percent of the time, since this rate could be achieved using a rule based on the flip of a coin. While this is an extreme example, one should always consider how well the derived classification rules perform compared with the random assignment of new objects to groups.

Several measures are available for evaluating the performance of the estimated classification functions. However, the real measure of interest is how well the classification function will perform in allocating future samples. One measure of performance, called the apparent error rate is defined as the fraction of observations in the training sample (the sample used to develop the estimates) that are misclassified. While this measure is intuitively appealing and easy to calculate, unfortunately, it tends to underestimate the actual error rate, which is the theoretical error rate that could be achieved with all prior probabilities, density functions and costs known. This problem stems from the fact that the data used to build the classification functions were also used to evaluate them, although the problem tends to disappear as the sample sizes become large. This phenomenon was observed in the present analyses, especially for those samples that were particularly large (e.g., drivers of passenger cars, light trucks/vans, and motorcycles).

One possible remedy to this situation would be to randomly partition the original training sample into two parts: a new training sample and a validation sample. The error rate observed in classifying the validation sample can be used to estimate the actual error rate. This approach was used in the analysis of drivers of passenger cars and light trucks/vans, since these groups had very large numbers of known BAC cases. The original samples were randomly partitioned into four subsamples, and separate classification functions were estimated for each subsample, providing information as to which variables consistently entered into each of the four respective sets of classification functions.

Although this method tends to overcome the bias problem by not using the same set of data to both build and evaluate the classification functions, it suffers from two main deficiencies: (1) it requires large samples, and (2) the function evaluated is not necessarily the function of interest. Ultimately, almost all of the data will be used to construct the classification functions; if not, some valuable information may be lost. Thus, after having observed the model structure for each of the four subsamples, in the end, all four subsamples were combined in order to estimate the final sets of classification functions.

Another approach to estimating the performance of classification functions that has worked well is called Lachenbruch's "holdout" procedure (4) sometimes referred to as "jackknifing". The procedure involves developing classification functions using all observations, and then estimating the misclassification rates by omitting one observation at a time, recalculating the classification functions without the contribution of the "held-out" observation, and then classifying this "new" observation. With the use of computers, this can be accomplished fairly easily for even large samples (a matrix identity permits quick recalculation of the discriminant functions without the contribution of the held-out observation), and provides a very nearly unbiased estimate of the expected actual error rate. This jackknifing option is available in many "canned" discriminant analysis computer packages, and was used in this effort. Although classifying cases (persons) into BAC groups was not the ultimate goal of this modeling effort, these misclassification rates provided a handy and reliable statistic for comparing the accuracy of several candidate models.

One difficultly encountered in this effort was that persons were rarely classified as belonging to the low BAC group ($0.01 \leq BAC \leq 0.09$), persons were generally classified as either totally sober or legally drunk, based on the maximum of the three estimated posterior probabilities. This was due to two factors: (1) the prior probability of a person having a low BAC was approximately ten percent, compared with the remaining probabilities of a zero BAC and a high BAC, each in the range of thirty-to-sixty percent, and (2) the characteristics of this group (daytime/nighttime, weekday/weekend, etc.) generally resembled those of the legally drunk drivers, presumably due to the drinking patterns in American society. However, since classification of persons into groups was not the ultimate goal of this effort, and since the low BAC group is relatively small compared with the sober and legally drunk drivers, this was not considered to be a significant problem. The generation of aggregate BAC distributions, when compared to actual BAC test results, would provide the best measure of the performance of the final models.

The main objective of this effort was to use discriminant analysis to derive rules (functions) for classifying new unknown BAC cases, and to use these functions to estimate, for each driver and nonoccupant with unknown BAC on the FARS file, posterior probabilities that the person had a BAC in each of the three groups. The posterior probabilities are to be retained and used as weights for estimating alcohol involvement across various person-level variables. Aggregate estimates of alcohol involvement in drivers, nonoccupants and accidents ultimately will be produced. The true test of performance is the comparison of the actual vs. estimated probabilities derived from the validation sample.

The posterior probabilities are computed directly from the classification functions. As each case is "run through" the model, the three classification functions (one for each BAC group) are evaluated using the case's attributes, producing three numbers, referred to as "discriminant scores" [S(0), S(1), S(2)]. The probability that a case belongs to the zero BAC group, P(BAC=0.00), is calculated as:

$$\frac{\exp[S(0)]}{\exp[S(0)]+\exp[S(1)]+\exp[S(2)]} .$$

The remaining probabilities are computed in a similar manner. The transformation from linear discriminant scores to a probability space is more easily justified under the assumption of normally distributed discriminant scores. Examination of the normal plots (also called Q-Q plots) of the cumulative distribution of discriminant scores for each respective BAC group vs. the cumulative normal distribution did not reveal any information that would lead one to reject this assumption. Retaining the posterior probabilities to be used as weights seemed to be a novel approach, and one which held great promise for describing alcohol involvement at the person level.

A second reason for retaining the person-level posterior probabilities was that they could be used to compute accident-level posterior probabilites, which are simply the joint probability distribution of all drivers and nonoccupants involved in each accident. This permitted the generation of the same types of estimates of alcohol involvement across the various accident-level variables, and did not require additional discriminant analyses to be conducted at the accident level.

Accident-level BACs are defined as follows: An accident is considered to be at zero BAC if ALL persons involved had zero BAC (the product of the individual probabilities that the persons involved had zero BAC); an accident is considered to be at $0.00 \leq \text{BAC} \leq 0.09$, if at least one person had a positive BAC, but no person was legally drunk ($\text{BAC} \geq 0.10$); an accident was considered to be at $\text{BAC} \geq 0.10$ if at least one person had $\text{BAC} \geq 0.10$ (the complement of the probability that no person had $\text{BAC} \geq 0.10$). The accident BAC may be considered to be the highest BAC of any driver or nonoccupant involved.

The analysis was conducted using the BMDP statistical software package (5) program P7M - Stepwise Discriminant Analysis. This program provides great flexibility in exploring the data and directing the steps of the analysis through the use of various hypothesis contrasts (similar to their use in ANOVA), forward and backward stepping, methods for controlling the order in which variables enter, and setting F-to-enter and F-to-remove criteria. In addition, the program output includes estimated misclassification rates based on the aforementioned jackknife procedure.

MODELING APPROACH

The data used to develop classification functions represent all drivers and nonoccupants with known BAC test results on FARS in 1982 and 1983. Comparable data for 1984 and 1985 also were available. The overall modeling approach involved the development of a set of classification functions based on the 1982-1983 data, and the use of known BAC cases for 1984 and 1985 as two validation samples. As long as the model performs well from year-to-year (with only minor modifications) on cases with known BAC, it should not be necessary to estimate new classification functions (these modifications will be discussed in the section titled "Model Validation and Maintenance"). This approach relies on the assumption that while the overall prevalence of alcohol in fatal accidents might change, the relative associations between alcohol involvement and the discriminant variables would remain intact. Of course this must be investigated each year. If in fact the rate of alcohol involvement does change from year-to-year, but there is no corresponding change in the relative distributions of variables associated with alcohol involvement (e.g., time of day, single- vs. multiple-vehicle accident, weekday-weekend, etc.) then one might question the presumed causal effect of alcohol in fatal accident occurrence. For example, one would expect a real reduction in alcohol involvement to be manifest as a reduction in the percentage of single-vehicle fatal accidents, or a reduction in the percentage of nighttime fatal accidents, etc.

All previous attempts to estimate alcohol involvement in fatal accidents had at least one feature in common: The methods treated all objects (drivers or accidents) as if they belonged to a single homogeneous group, and all had the same underlying marginal BAC distribution. For example, in the "cell" method, in each individual cell drivers of motorcycles were treated in the same manner, and assigned the same BAC distribution as drivers of medium and heavy trucks, assuming their actual BACs were unknown. Examination of the BAC distributions, for cases with known BAC, reveals that medium and heavy truck drivers have a much lower rate of alcohol involvement than do drivers of motorcycles. These data are presented in Table 3.

Table 3
BAC Distributions for Drivers of Various Vehicle Body Types
Known BAC Cases
FARS 1982

	No. of Cases	BAC=0.00	0.01-0.09	BAC >0.10
Passenger cars	23,611	39%	12%	49%
Utility vehicles	654	28%	13%	59%
Motorcycles	4,460	36%	16%	49%
Buses and large limousines	40	95%	3%	3%
Light trucks and vans	6,640	32%	12%	56%
Medium and heavy trucks	1,543	80%	7%	13%
Vehicles towing motorhomes	51	69%	2%	29%
Miscellaneous vehicles	279	48%	10%	42%
Nonoccupants	6,589	48%	9%	43%
Total	43,867			

Thus, it appears that a significant enhancement to the estimation of unknown BACs can be achieved through stratification of the model by various vehicle classes, that is, the development of individual classification functions for drivers of each of the various classes of vehicles. This additional accuracy results from the abilty to use different sets of discriminant variables in the classification functions for each of the model strata, and the use of more appropriate prior probabilities for each respective stratum, rather than the use of one overall prior BAC probability distribution. This stratification accounts for the observed differences in the BAC distributions of drivers of different vehicle types; for example, drivers of medium and heavy trucks tend to be working when driving their vehicles, which probably explains their low rate of alcohol involvement compared to drivers of motorcycles, vehicles more closely associated with recreational driving.

Data for 1982 and 1983 were combined in order to achieve large enough samples of known BAC cases in most of the model strata, especially medium and heavy trucks, utility vehicles, and miscellaneous vehicle body types. It was also desirable to develop models which could be expected to perform well over time, in order to avoid the re-estimation of new models each year. If the results of the validation runs for each successive year (1984, 1985, etc.) were to exhibit sufficiently large divergences between known BAC cases and their computed estimates, this would lead to investigations regarding potential modifications to the classification functions (variable selection and parameter estimation).

The 1982-1983 FARS files contain 128,007 drivers and nonoccupants involved in fatal accidents, of which 43,867 (thirty-four percent) have known BAC test results on file. Table 4 presents the number of cases with known BAC within each of the model strata, and the FARS codes for the stratification by the variable BODY_TYP.

Table 4
Sample Sizes for Model Strata
Initial Stratification

BODY_TYP Codes		Known BAC	Unk. BAC
01-09	Passenger cars	23,611	43,044
10-12	Utility vehicles	654	824
20-29	Motorcycles	4,460	4,313
30-39,13	Buses and large limousines	40	552
40-41,48-51,53-69	Light trucks and vans	6,640	14,324
70-72,74-76,78-79	Medium and heavy trucks	1,543	8,138
42,52,73,77	Vehicles towing motorhomes	51	178
80-89	Miscellaneous vehicles	279	2,029
N/A	Nonoccupants	6,589	10,738
	Totals	43,867	84,140

One of the most important variables in the estimation of driver and nonoccupant alcohol involvement was the police-reported alcohol involvement (the FARS variable name is DRINKING). This variable has four possible responses: (1) no (alcohol not involved), (2) yes (alcohol involved), (3) not reported, and (4) unknown (police reported). In theory, this assessment is made by the police officer before any BAC test is administered, and whether or not any BAC test is administered. The first two responses provide a definitive judgment regarding alcohol involvement as rendered by the police, and this information proved to be quite significant in classifying persons and estimating the posterior probabilities of alcohol involvement. However, the remaining two responses (not reported and police-reported unknown) offered no definitive indication of alcohol involvement, and required special treatment in the modeling process. Since this is a dummy (yes/no) variable, unknown cases could not be treated in the same fashion as variables such as accident hour or person age, wherein unknown values were replaced by estimated mean values based on the remaining known cases within each respective model stratum.

The cases with nonreported or unknown police-reported alcohol involvement were found not only for those cases with unknown BAC; on the contrary, there were numerous cases with known BAC for which no definitive police-reported alcohol involvement was available. Since these unknowns could not be replaced by estimates, and since there were cases with known BAC available for analysis, a second level of stratification was introduced for definitive (yes/no) vs. nonreported or unknown police-reported alcohol involvement. Thus, the initial number of models to be estimated doubled from the original nine to eighteen, representing a nine-by-two factorial stratification scheme. Table 5 presents the sample sizes within each of the new strata.

Table 5
Sample Sizes for Model Strata
Second Level of Stratification

	Known Police-Rept'd Alcohol Involvement		Unknown Pol.-Rept'd Alcohol Involvement	
	Known BAC	Unk. BAC	Known BAC	Unk. BAC
Passenger cars	16,223	30,769	7,388	12,275
Utility vehicles	478	589	176	235
Motorcycles	2,615	2,585	1,845	1,733
Buses and large limousines	37	449	3	103
Light trucks and vans	4,852	10,103	1,788	4,221
Medium and heavy trucks	1,035	5,939	508	2,199
Vehicles towing motorhomes	35	126	16	52
Miscellaneous vehicles	182	621	97	1,408
Nonoccupants	3,653	7,024	2,936	3,714
Totals	29,110	58,205	14,757	25,940

The models for each of the eighteen strata were estimated using the known BAC cases and applied to those cases with unknown BAC. For example, data representing 16,223 drivers of passengers with known BAC were used to estimate classification functions which were applied to data representing 30,769 drivers of passenger cars, on a case-by-case basis, to produce the estimated BAC probability distributions.

MODEL ESTIMATION

The BMDP statistical package, program P7M - Stepwise Discriminant Analysis, was used to estimate classification functions for each of the respective model strata. Since "stepwise" programs such as this and stepwise regression analysis are exploratory in nature, variables that did not enter the classification functions were recursively dropped from the analysis, and new estimates derived. Since cases with unknown values in the potential discriminant variables were omitted from the analysis, this procedure allowed the maximum utilization of known cases on which the final classification functions were based. However, the great majority of the discriminant variables in the final model enjoy a very high rate of reporting.

In reviewing the estimation results, it is important to remember that discriminant analysis produces classification functions that are associative in nature, rather than causal. While the coefficients of variables within a single model can serve to corroborate certain hypotheses or observations regarding associations between alcohol involvement and the discriminant variables, the presence of a particular variable in one model and its absence in another can only imply that the variable either did or did not make a significant marginal improvement in classifying known cases, given the sets of variables that had already entered, in stepwise manner, into each respective model.

As is the case with most multivariate analysis techniques, the judgment of the analyst plays a role in the selection of the final models. Since the BMDP discriminant analysis program permits a great deal of flexibility in directing the analysis, there were occasions that required selecting the best model from several reasonable models. The model with the greatest percentage of cases correctly classified was generally selected, although in a few instances, where the rates of correct classification were almost equal among the candidate models, the simpler model was chosen, i.e., the model with the fewest number of discriminant variables.

Models were estimated for the various combinations of vehicle classes vs. definitive/unknown police-reported alcohol involvement. These models were combined into one large estimation program, and applied to 1984 FARS cases with known BAC test results. This test was used as a validation of the accuracy of the proposed models in estimating aggregate alcohol involvement. Validations were run against various driver subsets, most notably, fatally-injured vs. surviving drivers, and drivers across age groups. During this latter validation, a larger than expected difference appeared between the actual vs. estimated rates of alcohol involvement for persons under age 21; however, the models fit well for the older age groups. After a number of investigations into this result, it seemed clear that the aggregate relationship between alcohol involvement and age is, to the greatest extent, influenced by the attributes of drivers age 21 and older, who comprise eighty percent of the drivers with known BAC on FARS, and that this alcohol-age relationship seems to change for drivers under age 21:

The reasons for this difference in the alcohol-age relationship can only be hypothesized; one might speculate that (1) all drivers age 21 and older can drink legally in the U.S., while many, if not most drivers under age 21 cannot, (2) the driving patterns of persons under age 21 are generally different from those of older drivers. For example, a greater proportion of the driving by persons under age 21 occurs at night so the high association between accident hour and alcohol involvement tends to overestimate the expected rate of alcohol involvement for these younger drivers.

In light of this, it seemed the best solution would be to estimate new models for drivers under age 21, specifically for drivers of passenger cars, light trucks and vans, and motorcycles. These three vehicle body types account for ninety-six percent of the fatal accident involvements by drivers under age 21. The final model structure is presented in Table 6, accompanied by the sample sizes (numbers in parentheses are vehicle class subtotals).

Table 6
Sample Sizes for Model Strata
Final Stratification Scheme

	Known Police-Rept'd Alcohol Involvement		Unknown Pol.-Rept'd Alcohol Involvement	
	Known BAC	Unk. BAC	Known BAC	Unk. BAC
Passenger cars	(16,223)	(30,769)	(7,388)	(12,275)
< 21	3,537	6,002	1,523	2,453
> 21	12,686	24,767	5,865	9,822
Utility vehicles	478	589	176	235
Motorcycles	(2,615)	(2,585)	(1,845)	(1,733)
< 21	612	793	409	469
> 21	2,003	1,792	1,436	1,264
Buses and large limousines	37	449	3	103
Light trucks and vans	(4,852)	(10,103)	(1,788)	(4,221)
< 21	806	1,582	255	687
> 21	4,046	8,521	1,533	3,534
Medium and heavy trucks	1,035	5,939	508	2,199
Vehicles towing motorhomes	35	126	16	52
Miscellaneous vehicles	182	621	97	1,408
Nonoccupants	3,653	7,024	2,936	3,714

For all of the driver strata, the same variables were used as candidates for entry into the discriminant functions at the initial step (step zero), except in the case of the police-reported alcohol involvement variable. The following variables initially were available for inclusion in both driver and nonoccupant models as appropriate:

- police-reported alcohol involvement,
- person age,
- vehicle role (Single vehicle, Multiple vehicle striking/struck),
- injury severity (fatally injured or surviving),
- accident hour (6 am = 01 through 5 am = 24),
- accident day (Mon = 1, Sun = 7),
- time of day-day of week interaction (accident hour x accident day),
- weekday/weekend (weekday = Mon-Fri, weekend = Sat-Sun),
- use of occupant restraint, including motorcycle helmets,
- driver license status (valid or not for this vehicle),
- number of entries on driver record (previous accidents, DWI convictions, speeding citations, other citations),
- whether or not driver could drink legally,
- person sex,
- location of nonoccupant in relation to the roadway,
- location of nonoccupant in relation to the intersection,
- whether or not the driver was involved in a nonoccupant accident,
- percent of drivers tested for BAC by state in which accident occurred,
- factor representing state "size" (population, vehicle miles of travel, square miles of land), and
- factor representing exposure to driving risk at the state-level (per capita vehicle miles of travel, number of driver licenses per registered vehicle).

The last two factors were developed using the BMDP Factor Analysis program, in an attempt to incorporate variables in the discriminant analysis, which might account for state-to-state differences. The percent of drivers tested for BAC by each state was included to account for any possible bias which might exist due to the various levels of BAC testing, as discussed earlier. None of these variables entered any of the discriminant functions.

Classification functions were developed for all strata except drivers of buses and large limousines, and drivers of vehicles towing motorhomes, due to the small sample sizes in these strata. The BAC probability distributions for drivers in these strata with unknown BAC were estimated by proportional allocation based on the known cases AND the police-reported alcohol involvement, where a definitive yes/no was available. Table 7 presents the variables that entered each of the discriminant functions.

Table 7
Variables in the Final Discriminant/Classification Functions

Known Pol-Rep't Alc Inv	DRK	HR	AGE	SSS	SEV	WK	RES	LIC	REC	LDA	DAY	SEX	RWY	PCT
Passenger cars														
< 21	X	X	X	X	X		X	X						73%
> 21	X	X	X	X	X		X	X	X					80%
Utility vehicles	X	X			X			X						79%
Motorcycles														
< 21	X	X		X										73%
> 21	X	X	X	X	X	X								73%
Buses and large limousines														
Light trucks and vans														
< 21	X	X	X	X	X		X			X	X			77%
> 21	X	X	X	X	X			X						81%
Medium and heavy trucks	X	X		X	X			X						87%
Vehicles towing motorhomes														
Miscellaneous vehicles	X	X			X	X								83%
Nonoccupants	X	X	X								X	X	X	81%
Unknown Pol-Rep't Alc Inv		HR	AGE	SSS	SEV	WK	RES	LIC	REC	LDA	DAY	SEX	RWY	
Passenger cars														
< 21		X	X	X	X	X			X	X				67%
> 21		X	X	X	X		X		X					69%
Utility vehicles		X		X	X									68%
Motorcycles														
< 21			X	X		X	X			X				63%
> 21		X		X		X	X		X					64%
Buses and large limousines														
Light trucks and vans														
< 21		X	X	X	X	X								68%
> 21		X		X	X	X								67%
Medium and heavy trucks		X		X										85%
Vehicles towing motorhomes														
Miscellaneous vehicles		X												68%
Nonoccupants		X	X			X						X	X	65%

The variables found to be most useful in classifying cases are (not necessarily in order of importance):

- police-reported alcohol involvement (DRK),
- accident hour (6 am = 01 through 5 am = 24) (HR),
- person age (AGE),
- vehicle role (Single vehicle, Multiple vehicle striking/struck) (SSS),
- injury severity (fatally injured or surviving) (SEV),
- weekday/weekend (weekday = Mon-Fri, weekend = Sat-Sun) (WK),
- use of occupant restraint, including motorcycle helmets (RES),
- driver license status (valid or not for this vehicle) (LIC),
- number of entries on driver record (previous accidents, DWI convictions, speeding citations, other citations) (REC),
- whether or not person could drink legally (LDA),
- accident day (Mon = 1, Sun = 7) (DAY),
- person sex (SEX), and
- location of nonoccupant in relation to roadway (RWY).

Most of these variables are quite familiar in alcohol traffic safety research; their close association with alcohol involvement has been recognized for many years. Not all of these variables entered each set of classification functions, but various combinations were observed, as noted above in Table 7. The column labeled PCT represents the percentage of cases that were correctly classified with regard to BAC group, using the jackknife procedure. As can be seen, the known police-reported alcohol involvement models correctly classified cases between seventy and eighty percent of the time, while the unknown police-reported alcohol involvement models correctly classified cases between sixty and seventy percent of the time. This compares favorably with the random assignment of cases to groups, which would result in correct classifications approximately thirty-three percent of the time.

The police-reported alcohol involvement variable formed the basis for one level of model stratification. When this variable is known (that is, yes or no) it serves, in the aggregate, as a second prior probability. This variable was the single most influential variable in classifying cases.

The variable representing the Minimum Legal Drinking Age (LDA) was a dummy (0-1) variable used in the analysis of drivers under the age of twenty-one, and represents whether or not the person could legally drink, based on the person's age. The remaining variables are self-explanatory.

For the purpose of illustration, Table 8 presents the estimated classification functions for drivers of passenger cars, twenty-one years of age and older, with known police-reported alcohol involvement. Classification functions for all of the model strata can be found in the Appendix.

Table 8
Estimated Classification Functions
Passenger Cars, Known PRAI, 21 years of age and older

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Drinking	2.32158	10.21068	11.50245
Hour	.55378	.66990	.70592
Age	.19033	.16911	.17937
SSS	2.05068	1.66487	1.34295
Severity	2.84507	3.24652	4.10704
Restraint	1.80014	1.31890	1.00636
Lic Stat	12.03953	11.81247	11.22057
Dr Record	.96757	1.03303	1.09779
Constant	-16.14600	-21.33804	-21.83783

Each case's attributes are evaluated using all three classification functions, producing the three discriminant scores, which are in turn transformed into posterior probabilities. The constant term is a function of the number of cases, the number of (BAC) groups, the respective group means for each variable in the classification functions, and the prior probability for each respective group.

Several observations about the variables can be made from Table 8. Note that the following variables' coefficients increase in magnitude between the BAC=0.00 group and the BAC≥0.10 group: DRINKING, HOUR, INJURY SEVERITY, AND DRIVER RECORD. This indicates, for example, that the likelihood that a fatal accident-involved driver is drunk increases as the accident hour increases, from 6 am (HOUR=1) to 5 am (HOUR=24); or, as the number of previous entries on the driver record (DRIVER RECORD) increases, so does the likelihood of alcohol involvement; or, a surviving driver (SEVERITY) is less likely to be legally drunk than a fatally-injured driver (all other things being equal). The same types of observations can be made regarding other variables, such as RESTRAINT use: a driver who was restrained in the accident is more likely to be sober or less than legally drunk, than an unrestrained driver. Thus, a sense of model "face validity" can be gained from inspecting these estimated coefficients, that is, the models seem to make sense.

The variables selected for the final discriminant functions entered each respective model in stepwise order based on an evaluation of an F statistic. At step zero (the initial step) the F-to-enter for a variable corresponds to the F statistic computed from a one-way analysis of variance on the variable for the groups used in the analysis (BAC groups). In subsequent steps, the F-to-enter for each variable not in the discriminant function is equal to the F statistic corresponding to the one-way analysis of variance on the residuals of the variable; i.e., at each step the F-to-enter is computed from a one-way analysis of covariance where the covariates are the previously entered variables (5).

This procedure can be altered through the use of user-defined hypothesis contrasts, the aim of which is to emphasize separation between specific groups which are "closer together" than others. For example, investigation of the group means of the candidate discriminant variables indicated smaller pairwise differences between the means of the $0.01\leq \mathrm{BAC}\leq 0.09$ group and the $\mathrm{BAC}\geq 0.10$ group compared with the BAC=0.00 group. The contrasts are structured in the same manner as in analysis of variance, with the above contrast represented by the triplet (0,-1,1) representing the three BAC groups. The use of contrasts affects the entry order of the variables since the calculated F statistics, upon which variable entry depends, is now based on a test of the specified contrast, rather than on the hypothesis of the equality of all group means. However, the coefficients of the classification functions, the classification matrix and the posterior probabilities are not directly affected; they depend only on the set of variables selected and not on the values of the contrast (5).

The relative strength of the variables in a regression equation can be investigated by computing the standardized regression coefficients, which represent the coefficients that result from a regression analysis based upon all standardized variables. There is no analogous standardization for variables in the discriminant functions; discriminant analysis is associative in nature compared with regression analysis; the regression coefficients represent the estimated change in the dependent variable resulting from a unit change in the independent variables. A very loose measure of explanatory power may be available from investigation of the stepwise order in which variables entered the discriminant functions. Table 9 lists the order in which variables entered each respective stratum's discriminant functions.

Table 9
Order of Entry for Variables in the Final Discriminant Functions

Known Pol-Rep't Alc Inv	DRK	HR	AGE	SSS	SEV	WK	RES	LIC	REC	LDA	DAY	SEX	RWY
Passenger cars													
< 21	2	4	3	5	1		7	6					
> 21	2	5	6	3	1		8	4	7				
Utility vehicles	1	4			2			3					
Motorcycles													
< 21	1	2		3									
> 21	1	3	5	2	4	6							
Buses and large limousines													
Light trucks and vans													
< 21	1	3	4	5	2		7			6	8		
> 21	2	4	6	3	1			5					
Medium and heavy trucks	1	5		4	2			3					
Vehicles towing motorhomes													
Miscellaneous vehicles	1	2			3	4							
Nonoccupants	1	2	6								5	4	3
Average Entry Position	1.3	3.4	5.0	3.6	2.0	5.0	7.3	4.2	7.0	6.0	6.5	4.0	3.0
Number of Appearances	10	10	6	7	8	2	3	5	1	1	2	1	1

Unknown Pol-Rep't Alc Inv		HR	AGE	SSS	SEV	WK	RES	LIC	REC	LDA	DAY	SEX	RWY
Passenger cars													
< 21		1	3	2	4	6			7	5			
> 21		5	2	3	4		6		1				
Utility vehicles		1		2	3								
Motorcycles													
< 21			3	1		2	4			5			
> 21		1		2		3	4		5				
Buses and large limousines													
Light trucks and vans													
< 21		1	5	2	4	3							
> 21		1		2	3	4							
Medium and heavy trucks		2		1									
Vehicles towing motorhomes													
Miscellaneous vehicles		1											
Nonoccupants		1	2			5						4	3
Average Entry Position	N/A	1.6	3.0	1.9	3.6	3.8	4.7	N/A	4.3	5.0	N/A	4.0	3.0
Number of Appearances	0	9	5	8	5	6	3	0	3	2	0	1	1

Table 9 presents the order in which variables entered each respective stratum's discriminant function, in addition to the average entry position and the number of times the variable was entered. For the models with definitive (yes/no) police-reported alcohol involvement, this variable (DRK) was clearly the most significant in terms of correctly classifying cases into the three BAC groups, always entering either first or second, and appearing in all ten respective models. The second most significant varaible was injury severity (SEV=fatally-injured/surviving), with an average entry position of 2.0 and eight appearances in models, followed closely by accident hour (HR) and vehicle role (SSS=single vehicle, multiple-vehicle striking/struck).

For the strata with unknown police-reported alcohol involvement, the most significant variables were accident hour (HR) and vehicle role (SSS), both generally entering the discriminant functions as either the first or second variables, and appearing in almost all models. This was followed by person age (AGE) and injury severity (SEV).

As mentioned previously, one method of investigating model performance is to divide the total sample into several random subsamples, and develop separate discriminant functions for each subsample. From these results one can review the performance of the derived discriminant functions, noting the percentages of cases correctly classified and the specific variables that entered each respective subsample's model. As an example, Table 10 presents the results of this approach applied to the estimation of the model for drivers of passengers cars, age 21 years and older, with known police-reported alcohol involvement.

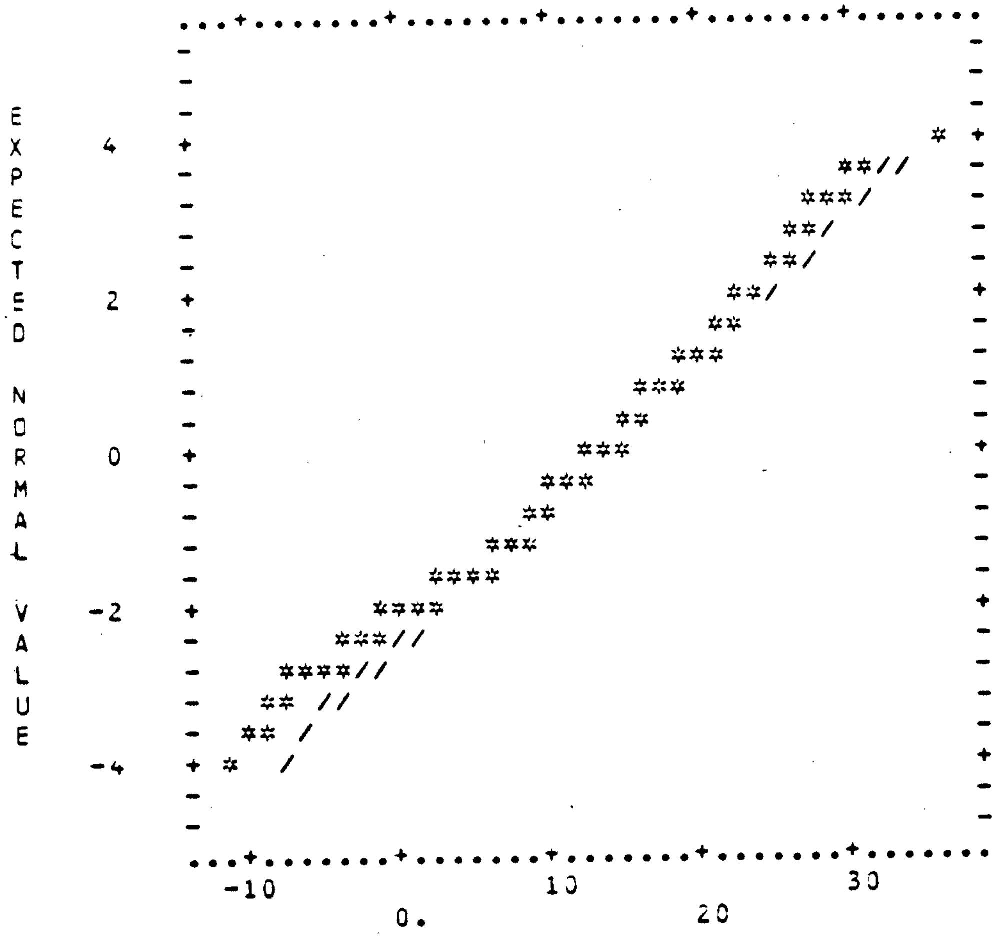
Table 10
Estimated Classification Functions
Passenger Cars, Known PRAI, 21 years of age and older
Total Sample and Four Random Subsamples

Variables	Total Sample	Subsample No. 1	Subsample No. 2	Subsample No. 3	Subsample No. 4
SEV	X	X	X	X	X
DRK	X	X	X	X	X
LIC	X		X	X	X
HR	X	X	X	X	X
SSS	X	X		X	X
AGE	X	X			X
REC	X	X			X
RES	X				
Percent Correctly Classified	80%	80%	80%	81%	79%

The total sample consisted of approximately 12,700 cases with known BAC; each systematic subsample consisted of approximately 3,200 cases. The percentage of cases correctly classified is very close for all four subsamples and the total sample, indicating a highly consistent accuracy in classifying cases into the "correct" BAC group. In addition, four variables entered all of the subsample models: police-reported alcohol involvement, injury severity, accident hour, and the driver's license status (valid/not valid). The only variable in the total-sample model that did not enter into any of the subsample models was restraint use (RES). This situation was identified earlier as one disadvantage of this approach; i.e., there is a possibility of missing the contribution of a variable if the final model is "selected" only from the subsample models. The variable restraint use (RES) entered into the total-sample model because its inclusion made a significant contribution to correctly classifying cases, given the variables already in the model, probably due to the larger sample size of the total sample.

As mentioned in an earlier section, the transformation of discriminant scores to posterior probabilities is more justified under the assumption of normally distributed discriminant scores. Most discussions of discrimanant analysis assume that the classification variables have natural units of measurement, that is, the variables can assume the value of any real number. In the current application, most of the variables are categorical in nature, and some can assume only two values (0,1) signifying the absence/presence of some condition (such as driver alcohol involvement). In this situation the assumption of the multivariate normality of the discriminant variables is not a sensible one. However, since the discriminants themselves are linear combinations of a large number of variables, they will often be nearly normal (4). This can be investigated using the normal plot of the discriminant scores. Figure 1 is a plot of the discriminant scores for the BAC=0.00 group, for the drivers of passenger cars, age 21 years and older, with known police-reported alcohol involvement.

Figure 1
Normal Plot of Discriminant Scores
Passenger Cars, Known PRAI, 21 years of age and older
Scores for the BAC=0.00 Group



In this normal probability plot, the observed values are plotted along the horizontal axis. The data values are ordered before plotting. The vertical axis corresponds to the expected normal value based on the rank (quantile) of the observation. The plotted points represent the set of points (x(i),q(i)) where the x(i) are the actual observations after ordering (i.e., x(1) is the point with the least magnitude and x(n) and the largest value) and q(i) is the standard normal quantile with probability level (i-1/2)/n. When the points lie very nearly along a straight line, the normality assumption remains tenable. This is the case in Figure 1.

MODEL VALIDATION AND MAINTENANCE

After having estimated the set of classification functions, the issue arises regarding how well these derived rules will perform in estimating BAC probability distributions for "new" persons with unknown BAC. One measure of this performance has already been observed, that is, the jackknife estimates of the misclassification rates available during the estimation stage. A better test of performance would be the use of data not used in estimating these models; the 1984 and 1985 FARS data were available for this purpose.

Two modifications to each year's models should be made before estimating BAC distributions for the unknown cases: (1) the mean values for each discriminant variable should be computed and substituted for the unknowns, and (2) the prior probabilities for each respective stratum should be calculated for the known BAC cases, and used explicitly in each respective stratum's models.

The coefficients of the variables in the final discriminant functions were estimated using only cases with known BAC and known values for the potential discriminant variables. While the rate of reporting is quite high for these variables, some cases for which BAC will be estimated will contain missing data, such as unknown person age, unknown driver license status, etc. For these cases, the unknown data is estimated using the mean for all persons in the respective stratum (e.g., drivers of passenger cars, age 21 years and older, with known police-reported alcohol involvement, etc.). Since the discriminant function coefficients and the constants for all BAC groups are functions of the group means of the variables, this substitution "pushes" the person's estimated BAC distribution toward that of a person with average age for the respective stratum, while the remaining "known" variables still contribute their information in estimating the BAC distribution for the unknown case.

The second facet of model maintenance, the use of each year's respective prior probabilities, is aimed at making maximum utilization of the known BAC data. One can expect changes in the overall rate of alcohol involvement in addition to changes in the relative prevalence of alcohol across the various discriminant variables. The updated prior probabilities, derived from each year's known BAC cases, can be explicity represented in the discriminant functions through a simple modification of the constant term for each BAC group. The constant term is a linear combination of the natural logarithm of the prior probability and the group centroid (multivariate mean). The modification is the subtraction of the logarithm of the previous year's prior probability and the addition of the logarithm of the current year's prior probability. This modification is made to each BAC group's constant term in each of the respective model stratum. The validation tests are conducted after having made both of these modifications to all of the various models.

The validation tests were conducted using the sets of persons with known BAC in 1984 and 1985 separately, and evaluating these persons' attributes using the derived classification functions (assuming that these persons had unknown BAC). The estimated BAC distributions were compared with the distributions obtained from the actual BAC test results of these persons.

While many such tests of model performance are possible, the focus was placed on investigating performance by vehicle class (the first level of stratification), by age group (the last level of stratification). In addition, a number of validations were conducted on specific aggregate statistics of general interest:

- o all drivers in fatal accidents,
- o all fatally-injured drivers,
- o all fatally-injured drivers in single-vehicle accidents,
- o all fatally-injured drivers in multiple-vehicle accidents, and
- o all surviving drivers in multiple-vehicle accidents.

The first set of tests compared the actual vs. estimated BAC distributions for all drivers in the above subsets. The results for 1984 and 1985 are presented in Tables 11-15 (percentages may not add to 100 percent due to rounding).

Table 11
Actual vs. Estimated BAC Distributions
All Drivers
FARS 1984 and 1985

	BAC=0.00	0.01-0.09	0.10+	N
1984				
Actual	.42	.13	.45	21,985
(Estimate)	(.43)	(.12)	(.45)	
1985				
Actual	.47	.12	.41	23,787
(Estimate)	(.48)	(.11)	(.41)	

Table 12
Actual vs. Estimated BAC Distributions
All Fatally-Injured Drivers
FARS 1984 and 1985

	BAC=0.00	0.01-0.09	0.10+	N
1984				
Actual	.43	.11	.46	16,113
(Estimate)	(.46)	(.09)	(.45)	
1985				
Actual	.46	.10	.44	16,774
(Estimate)	(.48)	(.08)	(.43)	

Table 13
Actual vs. Estimated BAC Distributions
Fatally-Injured Drivers in Single-Vehicle Accidents
FARS 1984 and 1985

	BAC=0.00	0.01-0.09	0.10+	N
1984 Actual (Estimate)	.28 (.31)	.10 (.09)	.62 (.60)	8,310
1985 Actual (Estimate)	.30 (.33)	.11 (.09)	.60 (.58)	8,453

Table 14
Actual vs. Estimated BAC Distributions
Fatally-Injured Drivers in Multiple-Vehicle Accidents
FARS 1984 and 1985

	BAC=0.00	0.01-0.09	0.10+	N
1984 Actual (Estimate)	.60 (.61)	.11 (.09)	.29 (.30)	7,803
1985 Actual (Estimate)	.62 (.64)	.10 (.08)	.28 (.28)	8,321

Table 15
Actual vs. Estimated BAC Distributions
Surviving Drivers in Multiple-Vehicle Accidents
FARS 1984 and 1985

	BAC=0.00	0.01-0.09	0.10+	N
1984 Actual (Estimate)	.45 (.43)	.17 (.20)	.38 (.38)	3,390
1985 Actual (Estimate)	.57 (.55)	.12 (.15)	.31 (.30)	4,299

As can be seen in Tables 11-15, the estimated BAC distributions are a close approximation to the actual values. The differences between actual and estimated percentages are in the range of one-to-three percentage points, with most estimates within two percentage points of the actual BACs. The sample sizes for these driver subsets are quite large, over 3,000 persons up to almost 25,000. In addition, while there are only two years of data, the differences between actual and estimated BAC appear to be consistent from 1984 to 1985, not exhibiting any systematic divergences.

Table 16 presents the estimated BAC distributions for all drivers by driver age.

Table 16
Actual vs. Estimated BAC Distributions
All Drivers by Age
FARS 1984

	BAC=0.00	0.01-0.09	0.10+	N
All Drivers Actual (Estimate)	.42 (.43)	.13 (.12)	.45 (.45)	21,985
Under 21	.43 (.42)	.18 (.17)	.40 (.41)	4,216
21-44	.35 (.36)	.13 (.12)	.53 (.52)	13,208
Over 44	.65 (.64)	.08 (.06)	.27 (.29)	4,558
Unknown Age	1.00 (.80)	0.00 (.10)	0.00 (.09)	3

Table 17
Actual vs. Estimated BAC Distributions
All Drivers by Age
FARS 1985

	BAC=0.00	0.01-0.09	0.10+	N
All Drivers Actual (Estimate)	.47 (.48)	.12 (.11)	.41 (.41)	23,787
Under 21	.49 (.50)	.17 (.15)	.34 (.35)	4,338
21-44	.39 (.41)	.11 (.11)	.50 (.48)	14,454
Over 44	.69 (.68)	.07 (.05)	.24 (.26)	4,987
Unknown Age	.50 (.64)	.13 (.06)	.38 (.29)	8

As can be seen in Tables 16 and 17, the results for the individual age groups, while not as close as the overall estimate, are also within "tolerable" limits. It should be emphasized that these tests are conducted on the known BAC cases to investigate the performance of the classification rules, which are to be applied to the cases with unknown BAC. When final estimates of alcohol involvement are produced, the estimated BAC distributions for cases with unknown BAC will be combined with those cases with known BAC to produce the final estimates. Some degree of error will always be expected, however, the relative error will be smaller for those subsets containing a greater percentage of known BAC cases.

Fatally-injured drivers are tested far more often than are surviving drivers, and might represent a less biased sample of BACs against which validation tests can be conducted. Tables 18 and 19 present the results of validations for fatally-injured drivers with known BAC for 1984 and 1985, respectively.

Table 18
Actual vs. Estimated BAC Distributions
Fatally-Injured Drivers by Age
FARS - 1984

	BAC=0.00	0.01-0.09	0.10+	N
All Drivers				
Actual	.43	.11	.46	15,982
(Estimate)	(.45)	(.09)	(.46)	
Under 21	.44	.14	.41	2,885
	(.45)	(.11)	(.44)	
21-44	.34	.11	.55	9,417
	(.37)	(.10)	(.53)	
Over 44	.66	.07	.27	3,678
	(.66)	(.05)	(.29)	
Unknown Age	1.00	0.00	0.00	2
	(.72)	(.14)	(.14)	

Table 19
Actual vs. Estimated BAC Distributions
Fatally-Injured Drivers by Age
FARS - 1985

	BAC=0.00	0.01-0.09	0.10+	N
All Drivers Actual (Estimate)	.46 (.48)	.10 (.09)	.44 (.43)	16,774
Under 21	.49 (.51)	.15 (.12)	.36 (.37)	2,881
21-44	.36 (.39)	.10 (.09)	.54 (.52)	9,950
Over 44	.68 (.68)	.07 (.05)	.25 (.27)	3,939
Unknown Age	0.00 (.40)	.25 (.09)	.75 (.51)	4

Again, the comparisons of actual vs. estimated BAC distributions gives a favorable impression, with the overall estimate for all drivers being very close to the actual value.

In addition to investigating the performance of the model across various driver age groups, estimates for the various vehicle body classes were computed and compared with the actual values. These data are presented for 1984 and 1985 in Tables 20 and 21, respectively. It should be remembered that the BACs for drivers of the vehicle body types BUSES AND LARGE LIMOUSINES and VEHICLES TOWING MOTORHOMES were estimated by proportion allocation based on the known cases, since no meaningful discriminant functions were developed.

Table 20
Actual vs. Estimated BAC Distributions
All Drivers by Vehicle Body Type
FARS - 1984

	BAC=0.00	0.01-0.09	0.10+	N
All Drivers				
Actual	.42	.13	.45	21,985
(Estimate)	(.43)	(.12)	(.46)	
Passenger cars	.43	.13	.45	13,728
	(.43)	(.12)	(.45)	
Utility vehicles	.28	.14	.58	333
	(.33)	(.12)	(.55)	
Motorcycles	.39	.15	.46	2,773
	(.40)	(.14)	(.47)	
Buses and large limousines	.88	.04	.08	25
	(.88)	(.04)	(.08)	
Light trucks and vans	.36	.13	.52	3,937
	(.35)	(.11)	(.54)	
Medium and heavy trucks	.80	.07	.13	1,002
	(.79)	(.06)	(.15)	
Vehicles towing motorhomes	.65	.06	.29	17
	(.65)	(.06)	(.29)	
Miscellaneous vehicles	.49	.06	.44	170
	(.56)	(.08)	(.36)	

Table 21
Actual vs. Estimated BAC Distributions
All Drivers by Vehicle Body Type
FARS - 1985

	BAC=0.00	0.01-0.09	0.10+	N
All Drivers Actual (Estimate)	.47 (.48)	.12 (.11)	.41 (.41)	23,787
Passenger cars	.47 (.49)	.11 (.11)	.42 (.40)	14,483
Utility vehicles	.29 (.31)	.13 (.12)	.58 (.57)	367
Motorcycles	.41 (.42)	.15 (.13)	.44 (.45)	2,914
Buses and large limousines	.93 (.93)	.02 (.02)	.05 (.05)	41
Light trucks and vans	.41 (.39)	.11 (.11)	.47 (.50)	4,470
Medium and heavy trucks	.86 (.85)	.04 (.04)	.10 (.11)	1,250
Vehicles towing motorhomes	.77 (.77)	.08 (.08)	.15 (.15)	26
Miscellaneous vehicles	.50 (.52)	.11 (.10)	.39 (.38)	236

Inspection of Tables 20 and 21 shows close agreement between the estimated and actual BAC distributions, especially for the vehicle body types with larger sample sizes. Thus, the validation tests show that the estimated BAC distributions are in very close to the actual BAC distributions, and can be used with confidence for estimating the cases with unknown BAC.

DISCUSSION

The use of linear discriminant analysis has produced a set of classification functions which have been used to estimate posterior BAC distributions for persons involved in fatal traffic accidents, who do not have a known BAC test result on the FARS files. These estimated BAC distributions have been shown to provide estimates which are "reasonably" close to actual values from several validation samples (1984 and 1985). The word "reasonably" has been used intentionally. Thusfar, point estimates have been provided, with no discussion of the standard error, or variability, of these estimates about the actual value.

From the many validation test results presented, one might surmise that the estimated BAC distributions were consistently within three percentage points of the actual values, with most estimates within two percentage points. If this is any measure of the accuracy of the model estimates, then one would require larger year-to-year differences in the estimated BAC distributions in order to infer statistically significant trends. However, since the final estimates of alcohol involvement are a combination of known BACs and estimated BACs, one might expect greater accuracy in these combined estimates of the rate of alcohol involvement, than that observed in the validation tests. This stems from the fact that there is no error of estimation in the known BAC cases, and the final percentages of alcohol involvement, being a function of the total size of the subset (known BAC cases plus estimated BAC cases), should result in a smaller relative error. For example, if the subset for which estimates are desired (e.g., fatally-injured drivers) consists of fifty percent known BAC cases and fifty percent unknown BAC cases, and the accuracy of the estimates are within three percent, then the combined estimates can be expected to be within one and one-half percent. Clearly, as the proportion of cases with known BAC increases, so does the accuracy of the final estimates of alcohol involvement.

The retention of the estimated posterior probabilities for use as weights is a novel approach. Generally, discriminant analysis utilizes the posterior probabilities to determine to which of the mutually exclusive groups each case is most likely to belong. There is little or nothing in the literature that addresses the variability of these posterior probabilities. While research into this area is certain to prove fruitful, there is a desire to disseminate the details of the overall modeling approach, and to make maximum utilization of these new estimates.

REFERENCES

(1) Fell, J.C., "Alcohol Involvement in Traffic Accidents: Recent Estimates from the National Center for Statistics and Analysis", NHTSA Technical Report, DOT-HS-806-269, May 1982.

(2) Cerrelli, E.C., "Alcohol in Fatal Accidents, National Estimates - U.S.A", Mathematical Analysis Division, NHTSA Technical Note DOT-HS-806-371, January 1983.

(3) Maxwell, D., "Classification and Estimation of Alcohol Involvement in Fatalities", Office of Alcohol Countermeasures, NHTSA Technical Note DOT-HS-806-372, January 1983.

(4) Johnson, R.A., and Wichern, D.W., Applied Multivariate Statistical Analysis, Prentice Hall, 1982.

(5) BMDP Statistical Software Manual, University of California Press, 1983.

APPENDIX

Table A-1
Final Discriminant Functions
Drivers of Passenger Cars, Under 21 Years of Age, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	10.73908	10.83629	11.80195
Drinking	.31774	7.29068	8.21109
Hour	.54394	.68609	.72875
Lic Stat	7.86100	8.26625	7.82023
Restraint	-2.56759	-2.50540	-3.31611
Severity	3.34846	3.58588	4.39581
SSS	1.46887	.99457	.70750
Constant	-106.26575	-114.25049	-119.30400

Table A-2
Final Discriminant Functions
Drivers of Passenger Cars, Under 21 Years of Age, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	10.12845	10.32778	10.45204
Dr Record	-1.72137	-1.55225	-1.58752
Hour	.38203	.51343	.55945
Severity	5.56488	6.24930	6.94729
SSS	.86699	.61613	.08855
Wkday/end	1.64246	2.19678	2.11047
MLDA	-.18731	-1.00676	-.76015
Constant	-96.63130	-103.83919	-106.31039

Table A-3
Final Discriminant Functions
Drivers of Passenger Cars, 21 Years of Age and Older, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	.19033	.16911	.17937
Drinking	2.32158	10.21068	11.50245
Dr Record	.96757	1.03303	1.09779
Hour	.55378	.66990	.70592
Severity	2.84507	3.24652	4.10704
Lic Stat	12.03953	11.81247	11.22057
Restraint	1.80014	1.31890	1.00636
SSS	2.05068	1.66487	1.34295
Constant	-16.14600	-21.33804	-21.83783

Table A-4
Final Discriminant Functions
Drivers of Passenger Cars, 21 Years of Age and Older, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	.14738	.12773	.12937
Dr Record	.42374	.62688	.66358
Hour	.46662	.59546	.62981
Severity	8.21520	8.25901	9.45727
Restraint	1.61068	1.91121	1.09274
SSS	1.76663	1.54427	.96617
Constant	-10.59947	-13.20008	-13.30957

Table A-5
Final Discriminant Functions
Drivers of Light Trucks and Vans, Under 21 Years of Age, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	11.02654	10.99303	11.42585
Day	1.53468	1.81225	1.77367
Drinking	1.56986	13.55412	14.24122
Hour	.25490	.42648	.45718
Restraint	-.62109	.21926	-2.08300
Severity	6.92521	7.43468	9.02115
SSS	-.24725	-.83051	-1.22462
MLDA	3.71382	3.70739	4.48099
Constant	-108.28178	-117.85539	-126.63025

Table A-6
Final Discriminant Functions
Drivers of Light Trucks and Vans, Under 21 Years of Age, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	9.15010	9.50464	9.51943
Hour	.55351	.71092	.74825
Severity	2.79452	2.71704	4.06324
SSS	.74436	.40901	-.38840
WKday/end	.11702	.37782	1.46226
Constant	-87.25464	-97.11098	-98.66747

Table A-7
Final Discriminant Functions
Drivers of Light Trucks and Vans, 21 Years of Age and Older, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	.20490	.18380	.19556
Drinking	1.77164	11.76080	13.19036
Hour	.53628	.65996	.70017
Lic Stat	10.23694	10.09150	9.49824
Severity	1.81117	2.21052	3.30850
SSS	2.18695	1.84645	1.35583
Constant	-14.32828	-20.63632	-21.13071

Table A-8
Final Discriminant Functions
Drivers of Light Trucks and Vans, 21 Years of Age and Older, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Hour	.36806	.46625	.51519
Severity	7.32409	7.49388	8.67708
SSS	2.68525	2.30826	1.66630
Wkday/end	1.30159	1.65557	1.84195
Constant	-6.95483	-9.69887	-9.74499

Table A-9
Final Discriminant Functions
Drivers of Motorcycles, Under 21 Years of Age, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Drinking	.06188	5.37844	5.90553
Hour	.60967	.62800	.80799
SSS	2.54653	2.29127	1.68338
Constant	-5.53814	-10.29995	-10.83416

Table A-10
Final Discriminant Functions
Drivers of Motorcycles, Under 21 Years of Age, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	6.54194	6.95099	6.88546
Helmet	-.86379	-1.37273	-2.10976
SSS	2.54706	2.07284	1.29092
Wkday/end	-.20564	1.35991	.67202
MLDA	.67161	-.93586	-1.64856
Constant	-59.30678	-68.21915	-64.86424

Table A-11
Final Discriminant Functions
Drivers of Motorcycles, 21 Years of Age and Older, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	.35293	.35104	.36593
Drinking	1.61630	6.41254	7.36898
Hour	.66661	.78053	.85187
Severity	12.95771	12.97147	13.56446
SSS	2.57069	2.15067	1.64105
Wkday/end	1.44933	1.70901	1.95849
Constant	-17.42056	-21.31224	-22.87764

Table A-12
Final Discriminant Functions
Drivers of Motorcycles, 21 Years of Age and Older, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Dr Record	.38132	.48198	.47691
Hour	.49288	.60746	.70862
Helmet	2.63108	2.42216	2.04470
SSS	2.30841	1.93521	1.19611
Wkday/end	1.23259	1.72343	1.83723
Constant	-5.81421	-8.31187	-8.20286

Table A-13
Final Discriminant Functions
Drivers of Medium and Heavy Trucks, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Drinking	.85733	10.09247	13.39544
Hour	.29020	.33483	.37962
Lic Stat	13.57156	13.19830	11.96371
Severity	3.48854	3.72731	4.87295
SSS	2.94311	2.83564	2.37122
Constant	-10.17829	-15.82434	-17.51682

Table A-14
Final Discriminant Functions
Drivers of Medium and Heavy Trucks, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Hour	.24107	.28736	.30395
SSS	1.36765	1.74670	.44792
Constant	-1.90794	-5.36587	-4.67106

Table A-15
Final Discriminant Functions
Drivers of Utility Vehicles, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
DrinKing	.81872	9.80287	10.57981
Hour	.48429	.57661	.58037
Lic Stat	10.07356	10.69228	9.37048
Severity	3.46946	3.92615	4.59040
Constant	-9.91816	-17.51096	-15.96442

Table A-16
Final Discriminant Functions
Drivers of Utility Vehicles, UnKnown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Hour	.48595	.58063	.67241
Severity	8.99121	7.47868	9.57067
SSS	1.41613	.71901	.71875
Constant	-8.33707	-9.57623	-10.99894

Table A-17
Final Discriminant Functions
Drivers of Miscellaneous Vehicles, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Drinking	.68273	14.88675	13.39121
Hour	.39353	.42597	.55790
Severity	4.37229	4.13872	5.83454
Wkday/end	1.09624	2.12001	2.86956
Constant	-4.54257	-13.98225	-14.37662

Table A-18
Final Discriminant Functions
Drivers of Miscellaneous Vehicles, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Hour	.38214	.49263	.57966
Constant	-2.29898	-5.48687	-5.57224

Table A-19
Final Discriminant Functions
Nonoccupants, Known PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	.14895	.14268	.13967
Day	1.09121	1.08352	1.18158
Drinking	-.23927	4.44152	6.75861
Hour	.71817	.85867	.87732
Sex	7.36981	6.97977	6.73339
Roadway	11.27456	11.68572	12.54485
Constant	-20.87111	-25.47418	-26.59036

Table A-20
Final Discriminant Functions
Nonoccupants, Unknown PRAI
1982 - 1983

	Group		
Variables	BAC=0.00	0.01-0.09	0.10+
Age	.12053	.10828	.10308
Hour	.63363	.74380	.78826
Sex	6.52685	5.90016	5.84551
Wkday/end	.47750	.77723	.91621
Roadway	8.54713	8.91903	9.75944
Constant	-15.70486	-18.13780	-18.02211