NPE-01

ADI Ref: 517

DOT-HS-802 594

HIGHWAY SAFETY PROGRAMS EFFECTIVENESS MODEL
Final Technical Report

Anthony N. Mucciardi, Ph. D.
Elsie C. Orr
Jian K. Chang, Ph. D.

Contract No. DOT-HS-6-01496
Contract Amount - $28,994

ADAPTRONICS, INC.
Westgate Research Park
7700 Old Springhouse Road
McLean , Virginia 22101

September 1977

FINAL REPORT

This document is available to the public through
The National Technical Information Service,
Springfield, Virginia 22161

Prepared for:

U.S. DEPARTMENT OF TRANSPORTATION
National Highway Traffic Safety Administration
Office of Contracts and Procurement
Washington, D.C. 20590

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle
Final Technical Report - Highway Safety Programs Effectiveness Model

5. Report Date
September 1977

6. Performing Organization Code

7. Author(s)
A. Mucciardi, E. Orr, J. Chang

8. Performing Organization Report No.
517

9. Performing Organization Name and Address
Adaptronics, Inc.
7700 Old Springhouse Road
McLean, Virginia 22101

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DOT-HS-6-01496

12. Sponsoring Agency Name and Address
National Highway Safety Administration
400 Seventh Street, S.W.
Washington, D. C. 20590

13. Type of Report and Period Covered
Final Technical Report
9/76 - 7/77

14. Sponsoring Agency Code

15. Supplementary Notes
N/A

16. Abstract

The purpose of this project was to construct a model capable of functionally relating highway safety (DOT/NHTSA) program outputs to (intermediate) risk factors and then to accidents, injuries and fatalities. The model inputs and outputs were obtained from a conceptual Causal Network which displayed the factors believed to influence the occurrence of an accident and their postulated interdependencies in leading to an accident. Also depicted in the network were the outputs of the highway safety activities as they were believed to interact with the intervening factors.

The models constructed were each (nonlinear) polynomial functions known as Adaptive Learning Networks (ALNs). The ALN methodology was applied to the factors set forth in a Causal Network constructed especially for this project. The relationships between the program outputs, the intervening factors, and the occurrence of accidents displayed in the network were tested along with various other variable combinations utilizing nationally representative data. In essence, the postulated network was checked and appropriately altered so as to trace quantitatively the effects of the outputs of highway safety programs in deterring accidents through the control of the intervening factors. This deterrent effect was estimated by asympotically reducing the outputs of the highway safety programs to zero and observing the impact of these reductions on the intervening factors, and in turn, the effect of these alterations in the intervening factors on accident occurrences.

The major results of this study were:

- Nonlinear, multivariate models possessing good accuracy have been synthesized

(continued)

17. Key Words
Causal Network; highway safety; risk factor; accidents; traffic conditions; Adaptive Learning Network

18. Distribution Statement
Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	70	

Reproduction of completed page authorized

16. Abstract (Continued)

for the intermediate risk factors using accident data collected in the State of Indiana.

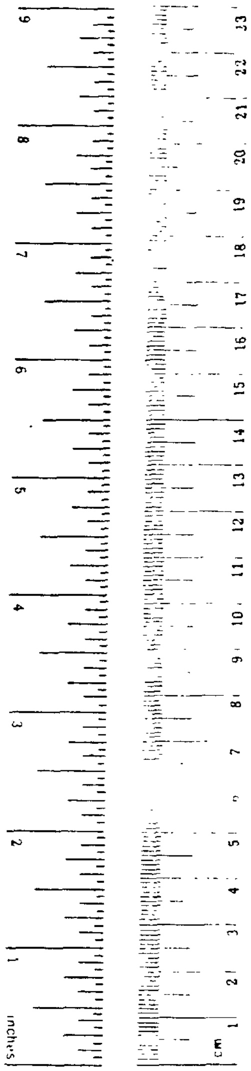
- The conjectured Causal Network was restructured by examination of which network variables were determined by the models to influence maximally a given risk factor.
- The effect of a particular exogenous variable -- driver age -- on intermediate risk factors was established quantitatively and it was shown how this information could be used to evaluate highway safety program outputs that might influence such variables.
- The influence of driver age was found to vary from small to considerable in predicting several highway risk factors.

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

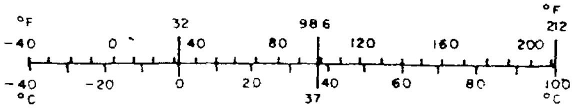
Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
	AREA			
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
	MASS (weight)			
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
	VOLUME			
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
	TEMPERATURE (exact)			
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
	AREA			
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
	MASS (weight)			
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
	VOLUME			
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
	TEMPERATURE (exact)			
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F



FOREWORD

This final report documents the materials, methods, results, conclusions and recommendations of the project entitled "Highway Safety Programs Effectiveness Model" sponsored by the Department of Transportation, National Highway Traffic Safety Administration, under Contract No. DOT-HS-6-01496. The research was conducted during the period September 1976 through February 1977.

Dr. Anthony N. Mucciardi was the Project Manager for Adaptronics, Inc. The authors thank the NHTSA Contract Technical Managers, Messrs. Dennis Pastorelle and George Booth, for their advice, encouragement, and guidance throughout this project.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

Page

APPENDICES

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION AND SUMMARY

1.1 PROJECT BACKGROUND

In early 1973, a systematic approach to assessing the developments and achievements of the U. S. highway safety programs was begun. Three successive phases of inquiry were established:

- Phase I studied how NHTSA state and community grants were spent by the states, in terms of equipment and services, and the catalytic effects of these funds produced from FY 1968 through FY 1973.
- Phase II yielded a broader examination of highway safety activities nationwide. This study measured national program outputs of highway safety efforts at all governmental levels from 1969 through 1974, using indicators of performance such as ratios and percentages.
- Phase III started with the findings of the earlier studies, and attempted to determine the effects of safety programs on the level of traffic accidents, injuries, and fatalities.

Preparation began for Phase III in the fall of 1975 with NHTSA literature searches to explore methodologies and techniques for approaching a detailed evaluation of national effectiveness. The ultimate objective of Phase III was to determine quantitatively the effects of highway safety programs on the occurrence of accidents, injuries, and fatalities.

A number of necessary components were recognized as being essential groundwork toward achieving the Phase III objective. These consisted of:

- Identifying those factors which related to the occurrence of accidents, injuries, and fatalities, and defining the framework in which they operated;
- Determining how these factors interrelated in influencing the occurrence of accidents, injuries, and fatalities; and

- Determining the structure in which the outputs of the highway safety programs impacted the occurrence of accidents, injuries, and fatalities through the alteration and control of these intervening factors.

Two efforts were initiated to examine and partially develop these components.

The first of these efforts was designed to approach all of the above components in an exploratory fashion -- the result being the construction of a Causal Network which ultimately displayed the factors believed to influence the occurrence of an accident and their postulated interdependencies in leading to an accident. Also depicted in the network were the outputs of the highway safety activities as they were believed to interact with the intervening factors. Such a network provided the means of relating program outputs to crash reduction, since safety efforts were intended to impact the factors associated with an accident and thereby reduce the occurrences of accidents. The expected benefit of a highway safety countermeasure program was estimated through knowledge of the functional relationship between the outputs of the proposed activity and the associated factors, and in turn the influence of those factors on crashes.

The development of the methodology and technology required to establish these functional relationships constituted the second of the two initial efforts and is the subject of this project and report. This effort was intended to model mathematically the structure developed in a Causal Network and to test that structure against nationally representative data. The technique explored in this initial modeling task is known as an Adaptive Learning technique. This approach to modeling is based on the premise that if a relationship exists between one or more independent variables and one dependent variable, that relationship must be encoded in any data collected on these variables. This premise is employed by Adaptive Learning in the sense that a given data base is analysed to determine if any functional relationships display themselves in the data. If such functional relations are found, those variables also correspond in the real world. Conversely, if no functional relations are found, it is concluded from the above premise that the variables are not predictably related in the real world.

These procedures have been completely automated by Adaptronics and were used in this study to explore the potential of the Adaptive Learning technique for modeling highway safety relationships. This approach was applied to the factors set forth in a Causal Network constructed especially for this project. The relationships between the program outputs, the intervening factors, and the occurrence of accidents displayed in the network were tested along with various other variable combinations utilizing nationally representative data. In essence, the postulated network was checked and appropriately altered so as to trace quantitatively the effects of the outputs of highway safety programs in deterring accidents through the control of the intervening factors. This deterrent effect was estimated by asympotically reducing the outputs of the highway safety programs to zero and observing the impact of these reductions on the intervening factors, and in turn, the effect of these alternations in the intervening factors on accident occurrences.

1.2 PROJECT STATEMENT AND OBJECTIVES

The purpose of this project, "Highway Safety Programs Effectiveness Model," was to construct a core model to identify and represent mathematically those interactions outlined in a conceptual Causal Network.

The specific project objectives were:

- Review for methodological validity, rigor, and feasibility, NHTSA's proposed evaluation approach of creating a mathematical model of the accident-occurrence structure.
- Apply Adaptronics analysis techniques and supporting software to the highway safety program impact assessment model design.
- Conceptualize and construct a mathematical model capable of functionally relating highway safety program outputs to the intermediate risk factors and then to accidents, injuries, and fatalities.

1.3 MODELING METHODOLOGY OVERVIEW

To understand the modeling technique employed and its application to highway safety, it is helpful to detail better that portion of the Causal Network which supports the modeling effort. A hypothetical Causal Network is displayed in Figure 1.1. (The network of the figure does not show the outputs of the highway safety programs or the "bottom line" of occurrences of accidents, injuries, and fatalities.) This network is depicted in a form believed to be conducive to realization of the model and not necessarily representative of the actual form of the Causal Network currently being researched and constructed. However, this hypothetical Causal Network will suffice for describing the model.

The network of the figure flows to the right, i.e., a line from factor A to factor B (B to the right of A) is interpreted as representing a suspected influence of factor A on factor B.

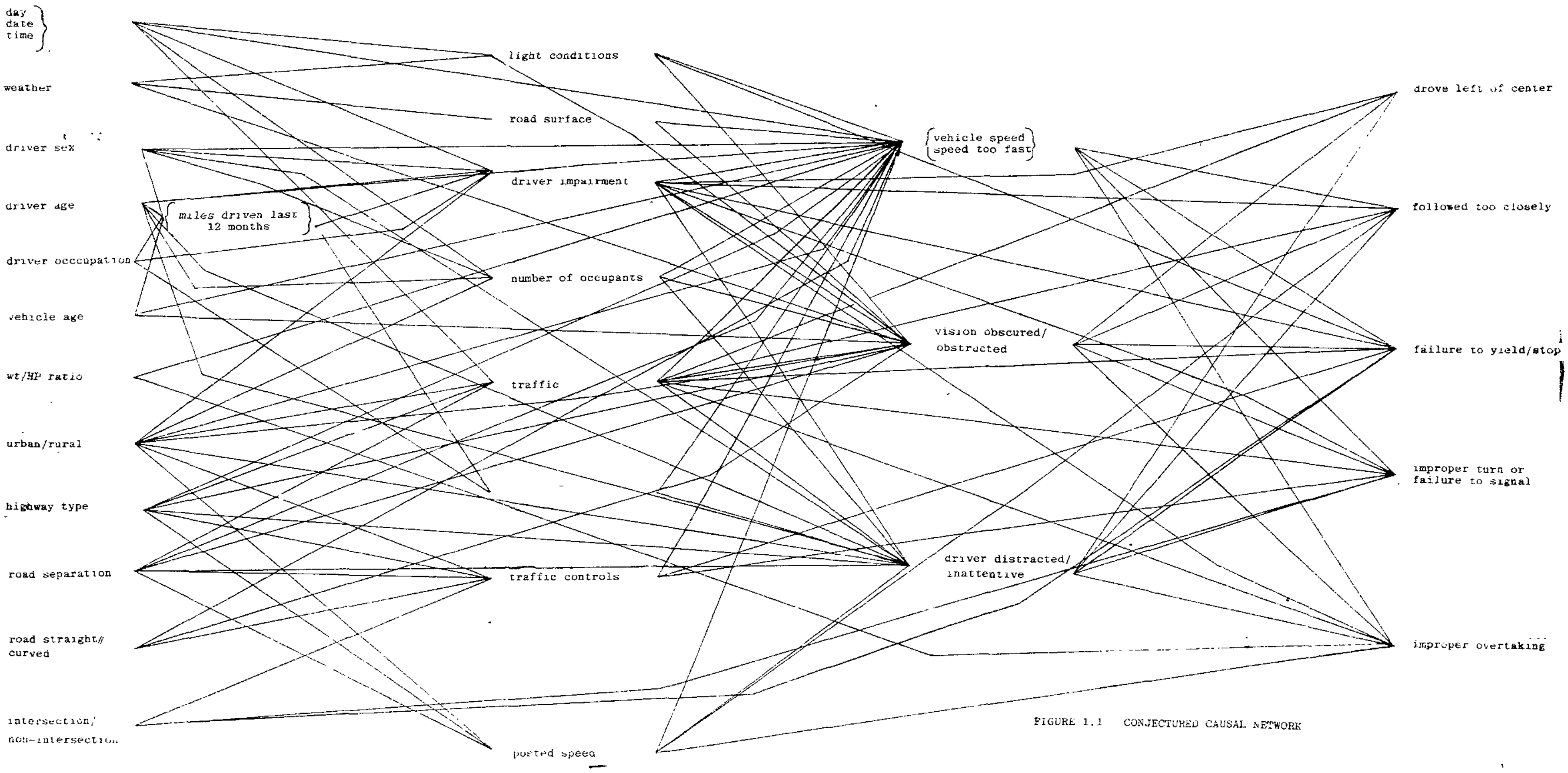


FIGURE 1.1 CONJECTURED CAUSAL NETWORK

All such factors (the A's) that flow into a single given factor (B) are suspected of either individually or jointly influencing the given B-factor.

As an example, select as the B-factor "driver impairment" (Figure 1.1). The A-factors are those believed to influence driver impairment as shown in the network by a line leading to this B-factor; namely, day-date-time, driver sex, driver age, driver occupation, urban vs. rural environment, and miles driven during the last 12 months. These six A-factors are called the independent variables for the dependent variable "driver impairment."

A model (i.e., an equation) could now be constructed to represent the relationship between the independent variables (A-factors) and the dependent variable (B-factor). This was accomplished as follows:

- The six independent variables were used as inputs for modeling driver impairment and their structure and coefficients were learned from recorded data for these variables, without reliance on assumptions by the analyst about mathematical structure. The input parameters that were most informative for the modeling purpose (i.e., predicting driver impairment) were automatically selected. The technique used to perform this task is called an Adaptive Learning Network (ALN) technique.
- The input variables did not need to be individually correlated with the modeled (dependent) variable "driver impairment." Often, nonlinear combinations of the inputs were correlated with the dependent variable, and when this occurred, these nonlinear combinations were fou-nd by the ALN method. Also, the input variables did not need to be statistically independent; various factors could be used as inputs even if they showed strong cross correlations.
- As the model (equation) of the relationship between "driver impairment" (the dependent variable) and the six factors (independent variables) evolved during synthesis, it became as rich in interactions between the input variables, in their nonlinearities, and in their multinomial structure as required for optimal fitting of the data.

- The model could possess as many degrees of freedom as necessary (even more than the number of data points used for its generation), but data overfitting was avoided. Note: the proof that overfitting had been controlled was to demonstrate on an independent evaluation set of data that the model accuracy rate was the same as that obtained on the data for which the model was synthesized; this proof was obtained routinely using known algorithms. The model was also realizable in extreme situations involving very large or very small amounts of data.
- Once the model was obtained, its use to obtain predictions required little computing effort.

This modeling approach was employed for each selected dependent variable (B-factor) displayed in the network (Figure 1.1). The combined use of these dependent-variable models comprised the overall model, and as such could be used to determine program impact as outlined in Section 1.1. Notice that the model identified the key risk factors (driver impairment, following too closely, etc.) as well as determined their quantitative importance. This knowledge could be used to decide which highway safety programs were needed to lessen the undesirable effects of these risk factors.

1.4 MAJOR RESULTS

The major objectives of this project have been accomplished. Specifically:

- Nonlinear, multivariate models possessing good accuracy have been synthesized for the intermediate risk factors (Figure 1.1) using accident data collected in the State of Indiana.
- The conjectured causal network (Figure 1.1) was restructured by examination of which network variables were determined by the models to influence maximally a given risk factor.
- The effect of a particular exogenous variable -- driver age -- on intermediate risk factors was established quantitatively and it was shown how this information could be used to evaluate highway safety program outputs that might influence such variables.
- The influence of driver age was found to vary from small to considerable in predicting several highway risk factors.

1.5 CONCLUSIONS AND RECOMMENDATIONS

It is concluded that the causal network approach of presenting the complex functional relationships between accident, risk factors, and endogenous and exogenous variables is mathematically sound and has utility in assessing highway safety program impact. Computer simulations performed by Adaptronics demonstrated that the adaptive learning network modeling methodology can be used effectively in quantitative modeling of causal networks.

One of the main difficulties encountered in this project was in coping with the definition and encoding procedures of the Indiana accident data base. As an example, the techniques for assessing "light conditions" and "road conditions" via visual examination created a considerable variation among different observers. It is recognized that these data were recorded under sometimes difficult circumstances and, occasionally, not even on the same day as the accident. However, it would definitely be of benefit to obtain objective measurements whenever possible. For instance, a light meter could be used to record light conditions if measured reasonably soon after the accident and a hand-held profilometer could be employed to measure the road surface condition.

It is additionally recommended that future data bases be collected with a better balance between the number of cases wherein a risk factor is cited and not-cited as accident-causative.

Finally, non-accident data should be collected. Even though there exist methods of synthesizing a pattern classifier when only accident-involved data are available, it is easier and more meaningful to design a classifier to discriminate between the accident-involved and non-accident populations when both data sets are available.

2. USE OF CAUSAL NETWORKS IN ASSESSING HIGHWAY SAFETY PROGRAM EFFECTIVENESS

2.1 BACKGROUND

The National Highway Traffic Safety Administration has been conducting the Highway Safety Program Impact Assessment to determine the impact of highway safety programs on the occurrence of traffic accidents, injuries, and fatalities. In the development of this assessment, a conceptual complex Causal Network approach is to be constructed. A contract for the "Construction of a Comprehensive Causal Network" is currently being supported by DOT/NHTSA, and the Center for the Environment & Man, Inc. is the contractor [10]. Their Causal Network will allow functional statements of the program output, risk factor, and accident occurrence environment to be made and it will provide the interactive capability of using actual accident data for an efficient and effective analysis.

2.2 CONCEPT OF CAUSAL NETWORKS

In the conceptual development of the assessment of highway safety program effectiveness, there is recognition that program performance levels are not capable of being related directly to accident levels in terms of avoiding or retarding growth trends. A complex network of intervening variables is at work and programs are being directed toward their alteration and control. These intervening variables are commonly referred to as "risk factors" or "factor variables". Figure 2.1 is a graphical representation of a conceptual Causal Network. It can be seen that highway safety program outputs P_1, P_2, ..., P_k give rise to "activities" (e.g., a program decision to lower speed limits may produce more visible police cars on the roads, advertising campaigns, etc.). These,

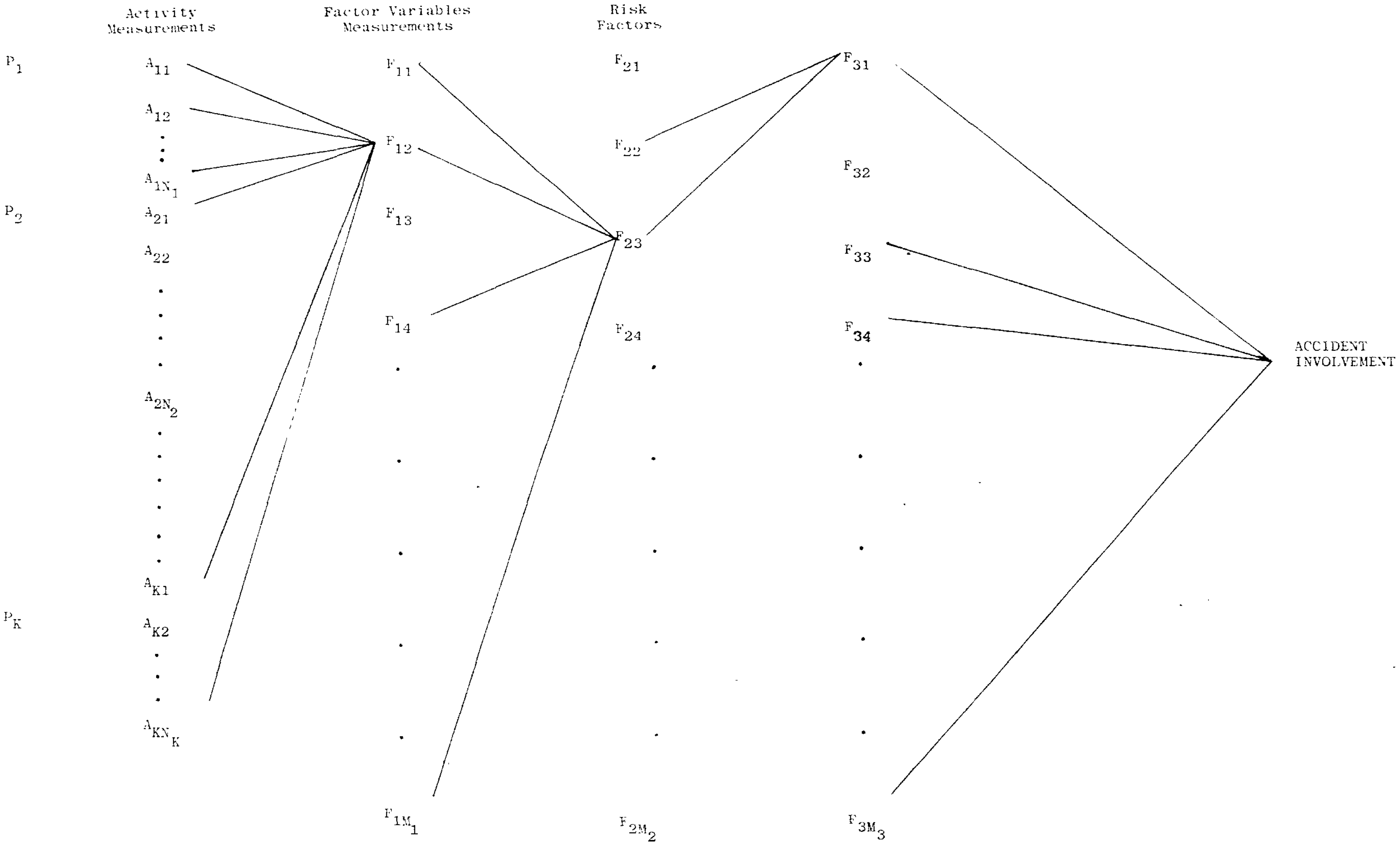


FIGURE 2.1: CONCEPTUAL CAUSAL NETWORK

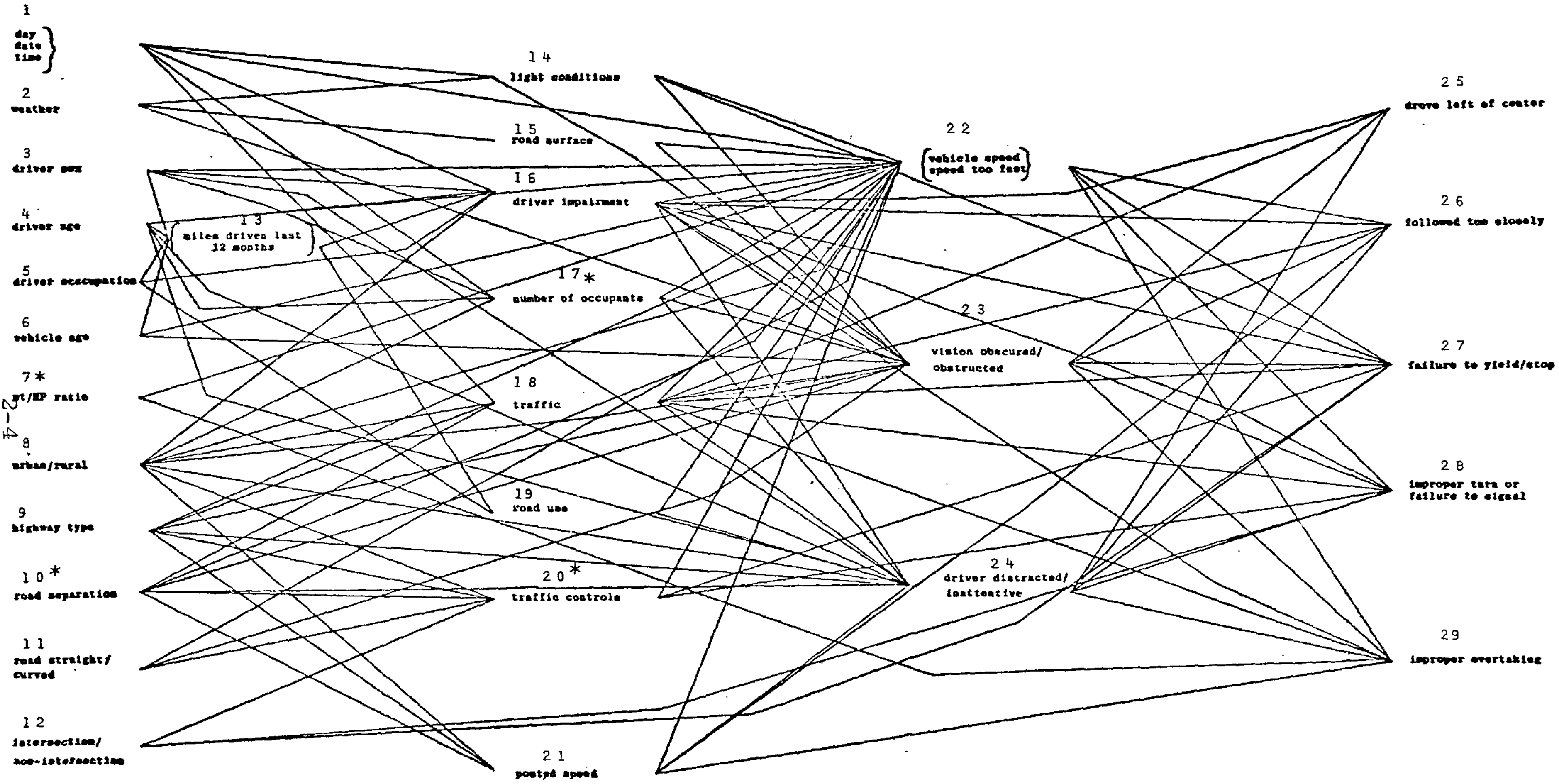
in turn, relate to the risk factors (e.g., "speed" is a risk factor that may lead to accident involvement). Certain risk factors are interrelated and lead ultimately to accidents. Thus, the link from a DOT-sponsored program output to effects on accidents is an indirect one.

As described above, in the relationships shown by a Causal Network one might find a particular risk factor affected by a multitude of program activities, with each activity making its individual impact at various levels given varying circumstances. Likewise, a single program output might affect more than one risk factor, again varying its impact given different conditions. To understand, diagram, and measure those complex relationships and hence to be in a position to make definite findings regarding program effectiveness, these conceptual Causal Networks provide guidance regarding the appropriate mathematical models to use.

A typical Causal Network, modeling part of the conceptual causal network, was constructed by the first Contract Technical Manager, Mr. D. Pastorelle, and others and it is given in Figure 2.2. For example, Risk Factor 14 (light conditions) is influenced by Variables 1 (day, date, time) and 2 (weather). Similarly, Risk Factor 21 (posted speed) is influenced by Variables 8 (urban/rural), 9 (highway type), and 10 (road separation). These are conjectured functional relationships between two highway risk factors and some of the exogenous variables (day, date, time, driver age, driver occupation, etc.).

2.3 USE OF CAUSAL NETWORKS

The main use of a conceptual Causal Network is to aid in assessment of highway safety program effectiveness. The end result is not the construction of a given type of comprehensive Causal



* Not Available

FIGURE 2.2: TYPICAL CAUSAL NETWORK

Network, but rather use and simulation of the Causal Network to evaluate highway safety program effectiveness. Hence, any description of Causal Networks should state clearly the guidelines and procedures regarding how it is to be used to assess highway safety program effectiveness.

To use fully any Causal Network as a guide for modeling purposes, accident involvement levels (the last layer) have to be defined. One approach is to use damage costs as indications of the levels of accident involvement. Another possibility is to define the levels of accident involvement as the seriousness or severity of the accident by some evaluation criterion.

Due to the short duration of this project and the lack of an accident level definition in the data base used in this project, the last layer in the Causal Network -- accident involvement level -- was left as further work. The Adaptronics ALN models were synthesized for all the other layers of the Causal Network (Figure 2.2).

3. HIGHWAY ACCIDENT DATA BASE

3.1 INTRODUCTION

To show the utility of the ALN Modeling technique in this application, a highway accident data base was required. The highway accident data base was supplied by NHTSA. It consisted of a subset of the accident data collected under the "Tri-Level Study of the Causes of Traffic Accidents" by the Institute for Research in Public Safety at Indiana University [8]. A detailed description of this highway accident data base, denoted ITADB (Indiana Tri-Level Accident Data Base), is presented in Appendix A.

3.2 CHARACTERISTICS OF THE HIGHWAY ACCIDENT DATA BASE

A total of 98 variables of the ITADB was recorded for each of 720 accidents (i.e., observations). (A description of these 98 variables is given in Table A-1 in Appendix A.) Only 29 of the 98 variables appear in the Causal Network of Figure 2.2. However, it was found that often more than one of the 98 ITADB variables fell within the definition of a given variable in the 29-variable Causal Network, so some of the ITADB variables were combined. The relationship between the 29 variables used in the Causal Network and the 98 ITADB variables is presented in Table A-2 of Appendix A. Variables 7 (Wt/Hp ratio), 10 (road separation), 17 (number of occupants), and 20 (traffic controls) of the Causal Network were not recorded in the ITADB.

The 98 ITADB variables were divided into the following five types of variable:

Type 1 - Informational Variables
Traffic Units, Day of Week, etc.

Type 2 - Environmental Variables
Weather Condition, Condition of Road Surface, etc.

Type 3 - Exogenous Variables
Age, Sex, Marital Status, etc.

Type 4 - Numerical Variables
Speed Limit, Frequency of Driving a Particular Road, etc.

Type 5 - Risk Factor Variables
Recognition Error, Inattention, Position of Car on Road, etc.

3.3 LIMITATIONS OF THE DATA BASE

After Examination of the ITADB, a number of problems was revealed:

- There were missing or unknown variables in some of the records (observations) - In some of the records, values were missing. These values were assigned in the following way. The frequency distribution for the variable under question was determined using that subset of the 720 observations for which values were available. The frequency distribution was then used to bias the generation of a (uniformly distributed) random number. This value was substituted for the missing value. A different random number, so generated, was used to substitute for each missing value of the given variable in the data set.
- Some variables had unbalanced distributions - Unbalanced distributions of a number of the ITADB variables were troublesome. For example, ITADB Variable P36 -- "cross-flowing traffic" -- was cited as a causative accident factor only 9 times out of the 720 accidents. Usually more than one of the ITADB "P" variables composed one of the "x" variables, so the value assigned to the x variable was determined as follows. If any of the P variables was cited as accident-causative, the corresponding x variable was also. For example, x_{16} was defined as Driver Impairment. The three P variables that relate to x_{16} were Impairment Due to Alcohol, Impairment Due to Drugs, and Impairment Due to Fatigue. The values of 1 and 2 were used to denote "not cited (N/C)" and "cited (C)", respectively. So, if alcohol, drugs, or fatigue singly, or in any combination, were cited as a causative factor (i.e., assigned the value 2), then x_{16} was coded as a 2 also; otherwise, it received the value 1 if the three P variables were all not cited.

Although this procedure meant that the x variables were better distributed between the N/C and C values than were the P variables, there were still some x variables that had mainly N/C values. So, if these variables were among the set that would serve as candidate inputs for a model of another variable, an attempt was made to find the largest subset of the 720 observations for which all of the input variables and the output variable would simultaneously have the most balanced distribution.

- <u>The method of coding the value of some variables was not very appropriate for quantitative modeling purposes</u> - The third problem with the ITADB was the manner in which the variables were coded. How does one numerically code the day of the week, the hour of the day, the weather conditions, etc.? This is a commonly recurring problem in a number of fields including highway safety. The approach used in this project was to assign numerical values in the most rational manner possible so that all the variables could be treated as taking on discrete values for modeling purposes. The procedures used are described in Appendix C. As an example, those variables that were either N/C or C as accident-causative were assigned binary variables, 1 (N/C) or 2(C). The hour and day variables were each split into two trigonometric variables as follows:

  Hour -- sin $(2\pi h/24)$ and cos $(2\pi h/24)$

  Day -- sin $(2\pi d/7)$ and cos $(2\pi d/7)$

  Thus, numerical discontinuities that would otherwise appear between the 24th and 0th hour and the 7th and 1st day were avoided.

In summary, it is emphasized that the ITADB was not designed originally with the purposes of this project in mind. Instead, it was the only data base available that could easily and quickly be transferred from one computer file to another and that, also, reasonably satisfied the needs of this project. Consequently, certain steps had to be taken in the use of the data base for modeling purposes that could raise questions of appropriateness, validity, etc. Adaptronics is sympathetic to these concerns and had debated them internally and with NHTSA personnel. The decision was made to proceed with use of the ITADB because the purpose of this project was to demonstrate the feasibility of mathematically modeling and analyzing Causal Networks. In this spirit, and because of the small time (4 months) and funds allotted to this project, it to believed that this was a sound decision. Further work will certainly need to be performed with data bases that are more closely matched to the needs of model syntheses. The results of this project can give considerable guidance for such future efforts.

4. CONSTRUCTION OF HIGHWAY SAFETY PROGRAM EFFECTIVENESS MODEL VIA ADAPTIVE LEARNING TECHNIQUES

4.1 ADAPTIVE LEARNING NETWORK (ALN) MODEL

In principle, models that predict risk factors can be either derived analytically or empirically.

An analytical model is one obtained by "reasoning from first principles." That is, the investigator attempts to interrelate all pertinent physical laws thought to influence injury. The problem with the analytical approach to modeling is that many physical processes are so very complex as to defy reasoning from first principles. Constructing a mathematical model necessarily requires a number of approximations about the relationship of one variable to another. If the guesses are wrong, the model proves to be inaccurate. Furthermore, the model may become quite cumbersome due to a large number of coupled equations, so that the computer processing time increases to unacceptable amounts.

Empirical predictive methods involve finding a predictive equation that best fits the observed experimental data. But, with conventional empirical modeling methods, one still has to know which interrelationships are important in order to write the general terms of the equation. And the resultant models, like analytical ones, are inflexible. If unanticipated changes occur in the process, the models become obsolete.

A different approach introduced by Adaptronics incorporates "self-learning" principles. To construct a self-learning model, the analyst first decides what variables may be important, but it is not necessary to consider the effects of the variables upon one another. What is needed instead is a collection of data that is reasonably representative of the variety of situations that can occur in the system being modeled.

The next step is to construct a mathematical network, known as an Adaptive Learning Network (ALN), which is a nonlinear hypersurface linking inputs to output. A generalized equation is constructed to link an output value to each possible pair of input variables. Special purpose computer programs are used to find the coefficients (the weights assigned to the variables) for each equation that makes it best fit the data. Those equations and variables that consistently produce the smallest prediction errors are determined. Additional equations are then constructed that examine interactions among four, eight, or more variables instead of only two. These additional equations result in added layers in the network and are retained if they can improve the prediction accuracy.

A model in the form of a network that has had its coefficients determined and has been reduced to the essential variables is called "adaptively trained." The synthesis of this model has proceeded directly from examination of an experimental data base without human intervention; hence the term "self-learning." To make certain that the model has indeed discovered for itself the pertinent physical laws, additional experimental data not used in the training, or synthesis, phase are introduced to test the ability of the model to generalize on its prior experience in dealing with new situations.

4.2 TYPES OF ALN MODELS

In this project, 15 nonlinear ALN models were synthesized to predict each of the tentative highway risk factors (given in the Causal Network of Figure 2.2). There were 15 such factors (not counting the first layer). The resulting ALN models were used in one of two ways depending on the nature of the dependent (i.e., modeled) variable.

If the dependent variable was of a "continuous" nature, such as vehicle speed, the ALN model was constructed to yield the output as a continuous variable. However, if the dependent variable could only assume two values as in N/C (=1) or C (=2), the ALN model was used as a classifier. In this case, the modeled hypersurface partitioned nonlinearly the input data space into two regions -- one associated with N/C outcomes and the other associated with C outcomes. So, for example, if a particular input vector was determined by the model to be on the N/C side of the separating decision boundary, a value of 1 was output. Most of the 15 models were of the classifier type due to the characteristics of the ITADB.

4.3 FORM OF ALN MODELS

The methodology associated with ALN synthesis is described more fully in References [3-8] by Barron and Mucciardi. In summary, two-input one-output "elements" are used to construct an adaptive learning network. The output of each element, y, is a quadratic function of its two inputs x_i and x_j:

$$y = w_o + w_1x_i + w_2x_j + w_3x_ix_j + w_4x_i^2 + w_5x_j^2$$

All combinations of inputs are considered two-at-a-time as above. For given identities of x_i and x_j, an optimization algorithm is used to find the coefficients, w, that yield the smallest error in fitting y to the values of x_i and x_j in a "fitting" subset of the data. Those combinations of variables yielding a low error rate (on an independent "selection" subset of the data) are then retained and the rest discarded. Thus, the candidate input list is pruned to the most informative subset. This produces the first layer in the ALN.

The outputs of Layer 1 become inputs to Layer 2 and the process is now repeated. Since each input to Layer 2 is a function of two x's, we are now considering functions of functions; thus the complexity of the model increases, but more slowly than its functional power. Only those combinations from Layer 1 are retained that produce the greatest improvement in accuracy. Now the outputs from Layer 2 become inputs to Layer 3, and so on.

The training procedure is terminated when it is established that the addition of further layers would produce an "overfitting" condition; that is, the model would become so adept at fitting the data used to train it that it would be unable to generalize to data not previously seen. Special algorithms are used to detect and avoid this condition.

An ALN Model thereby assumes the form of a multinomial -- a polynomial in many variables -- of the (automatically) selected inputs. The extent and type of non-linearities in model structure can be discovered and implemented during model synthesis. Thus, the ALN methodology is a powerful tool for use in data modeling instances where little or no knowledge exists regarding the functional relationship of dependent to independent variables.

4.4 FOUR APPROACHES TO MODEL SYNTHESIS

In consultation with the NHTSA Contract Technical Manager, four approaches to model synthesis were devised. The approaches differed only in which variables were used as the independent variable inputs when constructing a model for a particular dependent variable.

Approach	Variables Used as Model Inputs
I	Those that had a direct link to the dependent variable in the conjectured Causal Network.
II	Only those that appeared in the immediately preceding layer.
III	Those that appeared in any of the previous layers.
IV	Same as III, plus those that appeared in the same layer as the dependent variable.

All four approaches could not be evaluated due to time and cost considerations. Approach IV was selected because it was the most inclusive.

The 15 risk factor models were therefore constructed in the following way. First, the dependent variable was identified. Second, the candidate independent (i.e., input) variables were, via Approach IV, all those in the same layer and any preceding layers of the Causal Network (Figure 2.2). Third, the ALN modeling algorithm was used to determine automatically: (a) the subset of candidate inputs most relevant for modeling accurately the dependent variable, (b) the structure of the model, and (c) the weighting coefficients for the various terms within the model. Fourth, a fraction of the data that was not used to synthesize the model was then employed to establish model accuracy on data not previously seen.

One of the very desirable benefits of the adaptive learning algorithm in this project was its capacity to discover -- from the data -- the model structure. This meant that the conjectured Causal Network could be used as a guide to initiate the modeling efforts, but that another structure was found through use of the algorithm. The final Causal Network -- "wired" automatically from accident data -- could then be compared to the original structure to search for causative links not previously considered or to reinforce already conjectured links.

5. RESTRUCTURING OF CAUSAL NETWORKS VIA ALN MODELS

5.1 MODELING RESULTS

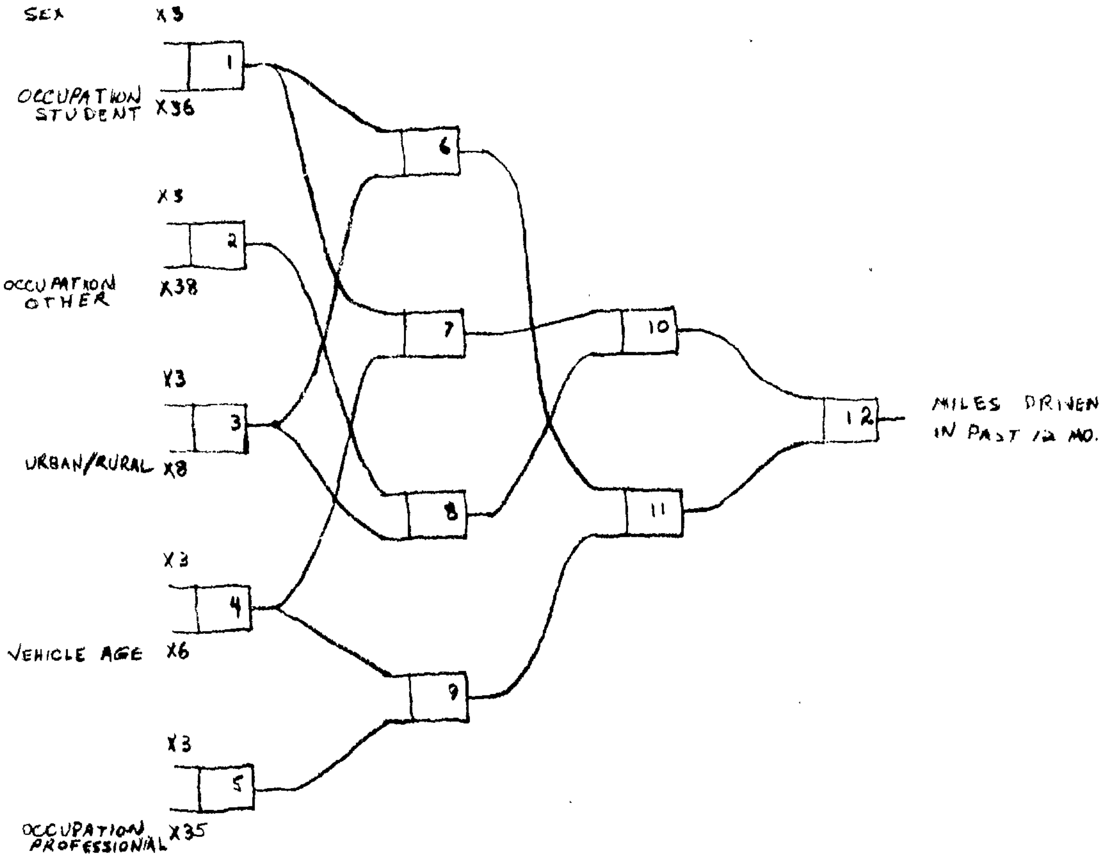
ALN models were constructued for each of the variables in the conjectured Causal Network for Layers 2 through 5. These included Variable 13 (miles driven during last 12 months) through Variable 29 (improper overtaking), inclusive.

The 15 resultant models are shown in Figures 5.1 through 5.15. In each figure the inputs that were selected are given as well as how they interact. The latter result is obtained by tracing a particular input variable's path through the net. The weighting coefficients for each element are given at the bottom of each figure.

As described in the previous section, all the variables to be modeled with the exception of 13, 18, 19 and 21 were binary valued. Hence, the ALN models were trained as classifiers for these 11 variables. Each of the 11 binary variables was coded as 1 for "not-cited" and 2 as "cited" as an accident-causative factor. The ALN's output was interpreted as N/C if it was less than 1.5 and, C otherwise. Thus, in 11 of 15 cases, the ALN models were pattern classifiers.

5.2 COMPARISON OF CONJECTURED AND RESTRUCTURED CAUSAL NETWORKS

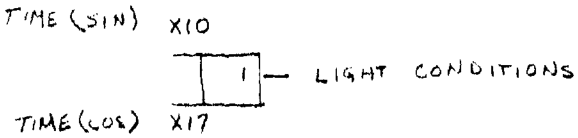
Using the ALN models, each node in the Causal Network was reconstructed and compared to the original conjectured structure. Appendix B shows the reconstruction of the Causal Network using the ALN models along with the original structure.



```
                                NETWORK WEIGHTING COEFFICIENTS
ELEM        W0             W1             W2             W3             W4             W5

  1     .23755E+03    -.54235E+02    -.52234E+02
  2     .32349E+03    -.40679E+02    -.10122E+03
  3     .23781E+03    -.41158E+02    -.10731E+02
  4     .22780E+03    -.42659E+02    -.30776E+01
  5     .17795E+03    -.45527E+02     .27691E+02
  6     .36276E+03    -.19781E+01    -.22246E+01     .18879E-01
  7     .49922E+03    -.30867E+01    -.31834E+01     .25851E-01
  8     .20013E+03    -.83293E+00    -.11578E+01     .10922E-01
  9     .43625E+03    -.26877E+01    -.26686E+01     .22579E-01
 10     .23475E+03    -.24613E+01     .18379E-01     .14870E-01     .71675E-02    -.57273E-02
 11     .76261E+03    -.71464E+01    -.26290E+01     .26301E-01     .12937E-01    -.24351E-02
 12     .17513E+03     .24634E+01    -.37804E+01    -.11990E-01     .21014E-02     .16899E-01
```

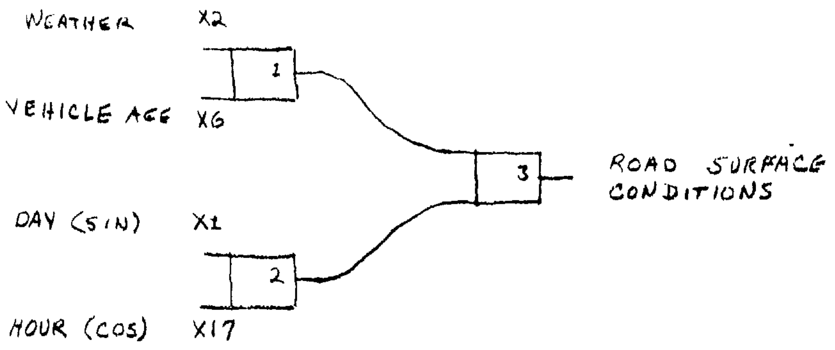
FIGURE 5.1: MILES DRIVEN IN PAST 12 MONTHS - MODEL STRUCTURE



```
                                 NETWORK WEIGHTING COEFFICIENTS
ELEM        W0            W1            W2             W3            W4            W5

  1     .13070E+01    .11754E+00    .39986E+00    -.27656E+00    .24100E+00    .31622E+00
```

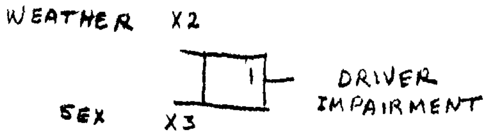
FIGURE 5.2: LIGHT CONDITIONS - MODEL STRUCTURE



```
                                    NETWORK WEIGHTING COEFFICIENTS
LEM          W0              W1              W2               W3              W4              W5

 1     .25686E+01   -.23753E+01   -.18545E-01    .71529E-02    .10934E+01   -.48934E-04
 2     .13714E+01    .13461E+00   -.47791E-01    .94904E-01    .11689E+00    .98017E-01
 3    -.22999E+01    .19879E+01    .17184E+01   -.80606E+00
```

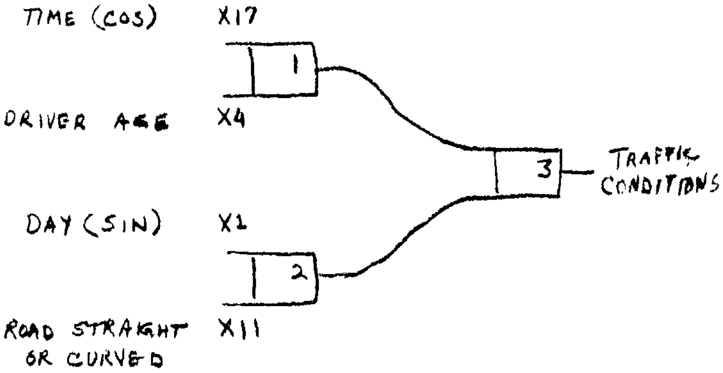
FIGURE 5.3: ROAD SURFACE CONDITIONS - MODEL STRUCTURE



```
                              NETWORK WEIGHTING OOEFFICIENTS
ELEM         W0           W1           W2           W3           W4           W5

  1     .18899E+00   .85357E+00   .55875E-01
```

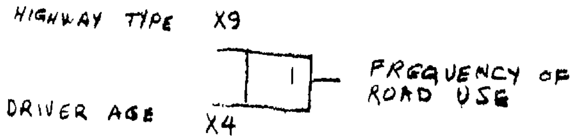
FIGURE 5.4: DRIVER IMPAIRMENT - MODEL STRUCTURE



```
                              NETWORK WEIGHTING COEFFICIENTS
ELEM        W0            W1            W2            W3            W4            W5

  1    .35838E+01   .51690E+00   -.44833E-01   -.54995E-02   .32093E-01   .42970E-03
  2   -.30000E+01   .16129E-01    .30000E+01    .60387E-01   .62552E-01  -.25000E+01
  3   -.13931E+02   .54073E+01    .54694E+01   -.17300E+01
```

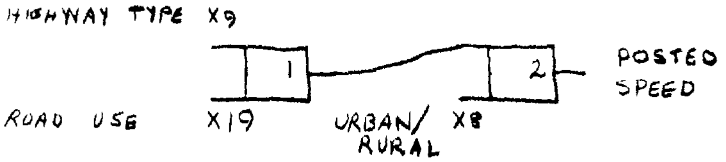
FIGURE 5.5: ROAD STRAIGHT OR CURVED - MODEL STRUCTURE



```
                                   NETWORK WEIGHTING COEFFICIENTS
ELEM        W0              W1              W2              W3              W4              W5

  1      .30735E+01    -.30404E+00    -.16834E-01    -.13784E-01    .44707E+00     .38177E-03
```

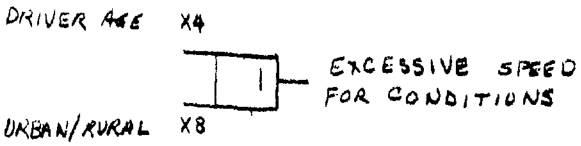
FIGURE 5.6: FREQUENCY OF ROAD USE - MODEL STRUCTURE



```
                                  NETWORK WEIGHTING COEFFICIENTS
ELEM          W0              W1              W2              W3              W4              W5

  1     .18158E+01      .31674E+01     -.38007E+00     -.30339E-01     -.22314E+00      .33390E-01
  2     .44772E+01      .31502E+00     -.25546E+01      .28437E+00
```

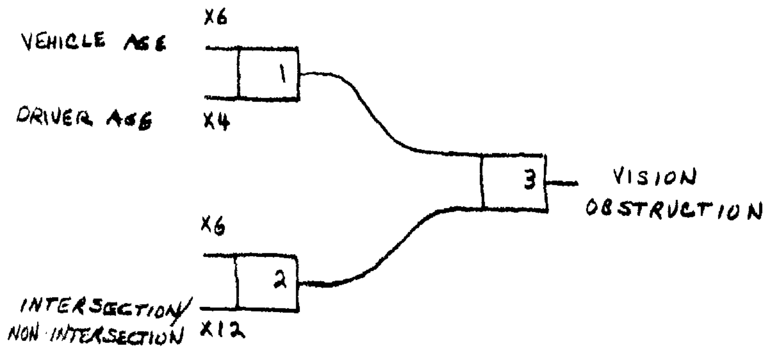
FIGURE 5.7: POSTED SPEED - MODEL STRUCTURE



```
                                    NETWORK WEIGHTING COEFFICIENTS
 ELEM          W0               W1             W2              W3              W4             W5

   1     .22375E+01     -.50278E-02    -.39703E+00
```

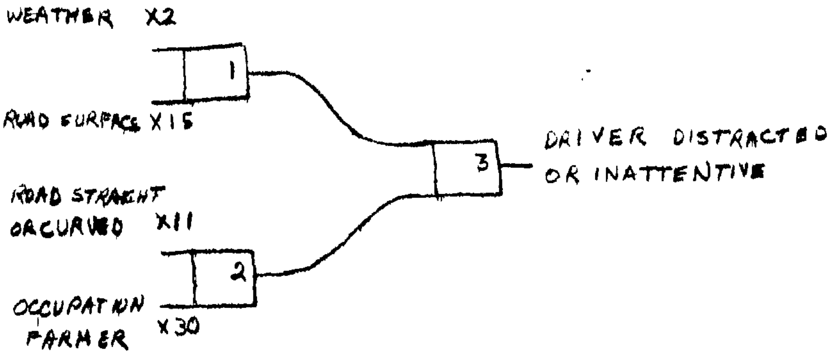
FIGURE 5.8: EXCESSIVE SPEED FOR CONDITIONS - MODEL STRUCTURE



```
                              NETWORK WEIGHTING COEFFICIENTS
ELEM        W0              W1              W2              W3              W4              W5

  1     .19221E+01     -.10086E-01     -.14057E-01      .43924E-03     -.20134E-02      .80505E-04
  2     .15000E+01      .41083E-01     0.              -.25063E-01     -.22854E-02     0.
  3     .27635E+01     -.21798E+01     -.17538E+01      .20815E+01
```

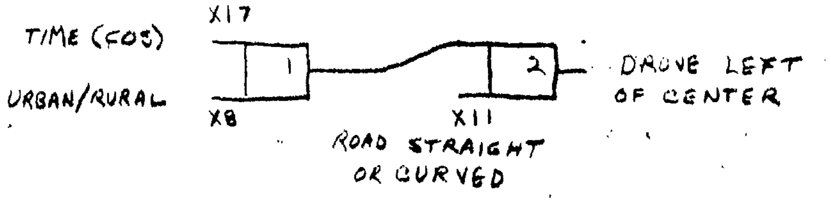
FIGURE 5.9: VISION OBSTRUCTION - MODEL STRUCTURE



```
                                   NETWORK WEIGHTING COEFFICIENTS
ELEM         W0              W1              W2              W3              W4              W5

  1     .15854E+01     .23305E+00     -.29893E+00
  2     .21715E+01    -.12507E+00     -.52320E+00
  3    -.13110E+01     .95742E+00      .91796E+00
```

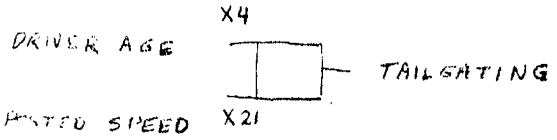
FIGURE 5.10: DRIVER DISTRACTED OR INATTENTIVE - MODEL STRUCTURE



```
                                    NETWORK WEIGHTING COEFFICIENTS
ELEM          W0                W1                W2                W3                W4                W5

  1     .12500E+01      -.37673E+00       .50000E+00       .22066E+00      -.22130E+00      -.25000E+00
  2    -.10211E+01       .17373E+01       .13768E+01      -.88092E+00
```

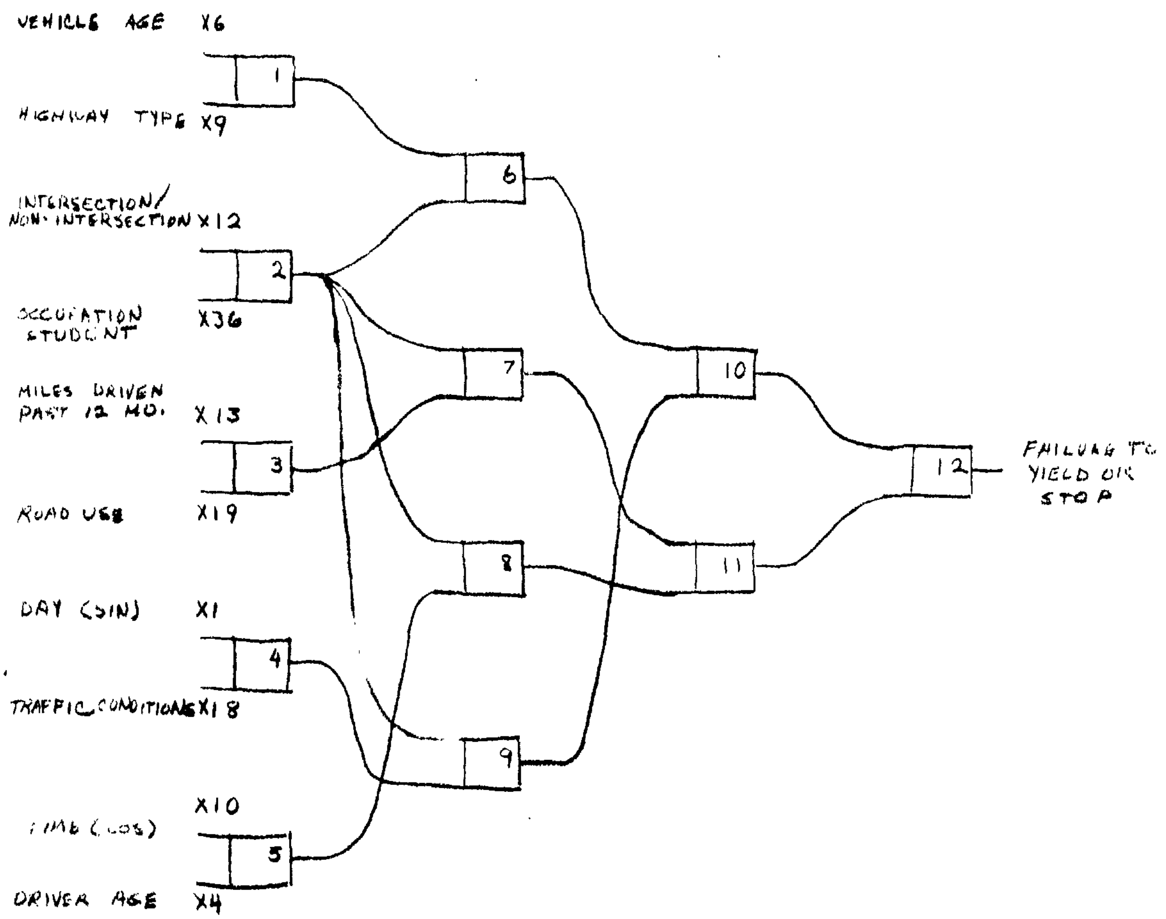
FIGURE 5.11: DROVE LEFT OF CENTER - MODEL STRUCTURE



```
                                  NETWORK WEIGHTING COEFFICIENTS
ELEM          WC               W1               W2               W3               W4               W5

  1     .22320E+01     -.20571E-01     -.39890E-01     -.68728E-03     .91375E-04     -.49893E-07
```

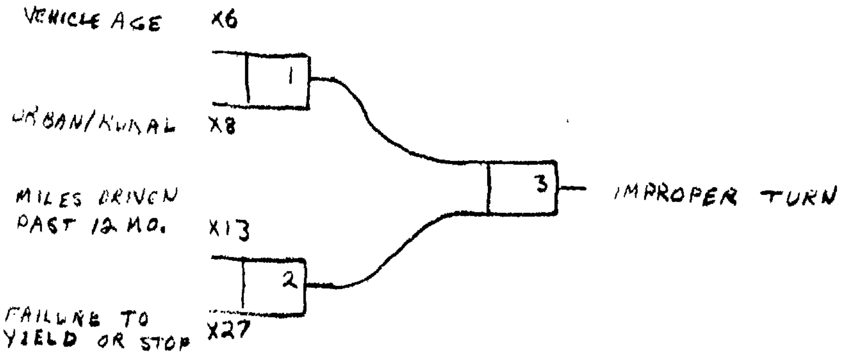
FIGURE 5.12: TAILGATING - MODEL STRUCTURE



NETWORK WEIGHTING COEFFICIENTS

ELEM	W0	W1	W2	W3	W4	W5
1	.11250E+01	.15355E-01	.75000E+00	.94972E-02	-.22758E-02	-.37500E+00
2	.13750E+01	.15000E+01	-.85156E+00	.26388E-01	-.59375E+00	.29492E+00
3	.14951E+01	.27941E-03	-.22586E-01	.15348E-03	-.22076E-05	.36996E-02
4	.18348E+01	-.19192E+00	-.26114E+00	.77697E-01	-.23363E-01	.42515E-01
5	.16568E+01	.18720E+00	-.96785E-02	-.56377E-02	.11704E+00	.67437E-04
6	-.44613E+01	.24470E+01	.28437E+01	-.99018E+00		
7	-.11136E+01	.67229E+00	.88873E-01	.52572E+00		
8	-.68121E+01	.41595E+01	.39154E+01	-.18151E+01		
9	.64171E+01	-.34359E+01	-.49132E+01	.32500E+01		
10	.13216E+01	-.37626E+00	-.69253E+00	.78363E+00		
11	-.20811E+01	.20301E+01	.18564E+01	-.98844E+00		
12	.22845E+00	.46950E+00	.11405E+00	.17372E+00		

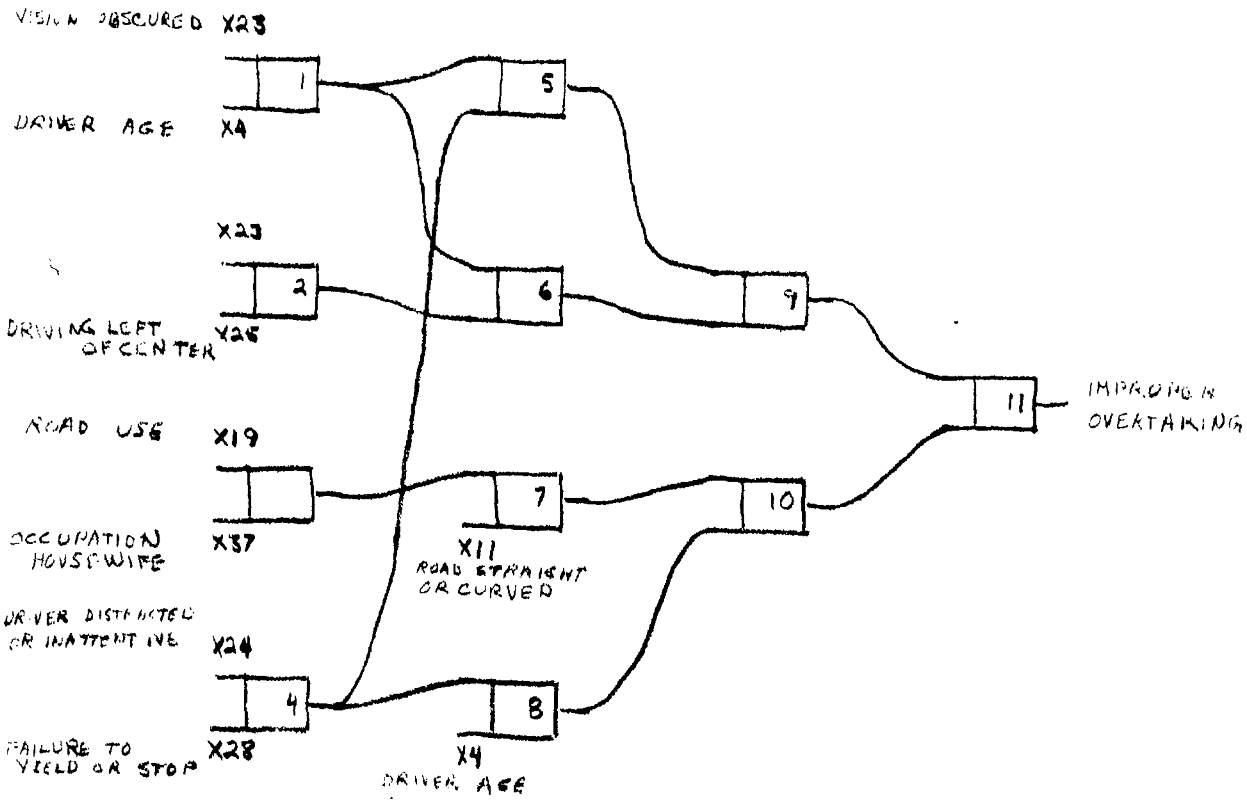
FIGURE 5.13: FAILURE TO YIELD OR STOP - MODEL STRUCTURE



```
                                  NETWORK WEIGHTING COEFFICIENTS
ELEM         W0              W1              W2              W3              W4              W5

  1      .50000E+00      .11869E+00      .50000E+00     -.48878E-01     -.30880E-02      0.
  2      .20000E+01      .29614E-02     -.75000E+00     -.70318E-03     -.51079E-05      0.
  3      .17793E+01     -.85327E+00     -.10756E+01      .12472E+01
```

FIGURE 5.14: IMPROPER TURN - MODEL STRUCTURE


VISION OBSCURED X23
DRIVER AGE X4
X23
DRIVING LEFT OF CENTER X25
ROAD USE X19
OCCUPATION HOUSEWIFE X37
DRIVER DISTRACTED OR INATTENTIVE X24
FAILURE TO YIELD OR STOP X28
1
2
4
5
6
7
8
9
10
11
X11 ROAD STRAIGHT OR CURVED
X4 DRIVER AGE
IMPROPER OVERTAKING

NETWORK WEIGHTING COEFFICIENTS

ELEM	W0	W1	W2	W3	W4	W5
1	.11055E+01	.23107E+01	-.72197E-01	.29344E-01	-.11001E+01	.42579E-03
2	.19252E+01	-.12500E+01	.10000E+01	.18524E-01	.37500E+00	-.50000E+00
3	.15000E+01	.73154E-01	.25000E+00	.23650E+00	-.39754E-01	-.50000E+00
4	.25000E+00	.13750E+01	.50000E+00	.74401E+00	-.81250E+00	-.50000E+00
5	-.51685E+01	.37626E+01	.35336E+01	-.19051E+01		
6	.48281E+01	-.33340E+01	-.31902E+01	.28421E+01		
7	.15844E+01	.49046E-01	-.15053E+01	.82611E+00		
8	.27134E+01	-.67110E+00	-.73253E-01	.44257E-01		
9	-.48262E+01	.37386E+01	.37551E+01	-.21626E+01		
10	.11757E+01	-.82105E+00	-.89005E+00	.12864E+01		
11	.27760E+01	-.17264E+01	-.18536E+01	.17866E+01		

FIGURE 5.15: IMPROPER OVERTAKING - MODEL STRUCTURE

The comparison between the original and restructured network is given in Table 5.1 for the 15 variables. The 15 dependent variables are shown as columns, and an open circle is used to denote a link that was conjectured to exist between a given pair of independent and dependent variables. An "x" is used to denote those links that were found by the respective ALN models. A circle with an "x" within it denotes agreement between the two procedures. For example, the first of the 15 dependent variables was x_{13}, "miles driven during last 12 months," shown in column one. Four factors were conjectured to be predictive of x_{13} -- (1) driver sex, (2) driver age, (3) driver occupation, and (4) vehicle age. The ALN model found that factors (1), (3), and (4) were indeed predictive of x_{13} but not so for factor (2). In addition, another factor, urban/rural, that was not conjectured to link to x_{13} was found to be relevant. It can be seen that there were cases in which the agreement was high (e.g., x_{13}), quite different (e.g., x_{27}), and considerably simpler due to the ALN restructuring process (e.g.,x_{16}).

The computer classification results derived from the ALN models are summarized in Table 5.2. The data set was divided into three subsets: Fitting, Selection, and Evaluation. The Fitting set was used to train the adaptive learning network, the Selection set for selecting the best subset of independent variables, and the Evaluation set to test the performance of the ALN model.

The best results were obtained for dependent Variable 16 (driver impairment), which was 90 percent accurate for the fitting, 93 percent accurate for selection, and 95 percent accurate for evaluation. The worse evaluation results were for dependent Variable 26 (followed too closely), which was only 42 percent accurate in classification.

TABLE 5.1

COMPARISON BETWEEN CONJECTURED AND ALN-DETERMINED CAUSAL NETWORK

Independent	13 Miles Driven	14 Light Cond.	15 Road Surf.	16 Driver Impair.	18 Traffic Cond.	19 Road Use	21 Posted Speed	22 Vehicle Speed	23 Vision Obs./Obst.	24 Driver Dis. Inatt.	25 Left of Center	26 Fol. Too Closely	27 Fail to Yield/	28 Imp. Turn/ Fail.to Signal	29 Improper Overtaking
7. Day			X	0	⊗			0					X		
2. Weather		0	⊗	X					0	X					
3. Dr. Sex	⊗			⊗				0		0					
4. Dr. Age	0			0	X	X		⊗	X	0		X	X		X
5. Dr. Occup.	⊗			0		0				X			X		X
6. Veh. Age	⊗		X					0	⊗				X	X	
8. Urb./Rur.	X			0	0		⊗	⊗	0	0	X			X	
9. Hway Type					0	X	⊗	0	0	0			X		
17. Time		⊗	X	0	⊗			0			X		X		
11. Str./Curve					X			0	0	X	X				X
12. Inters./Non Int									X				⊗	0	
13. Mi. Driven				0		0							X	X	
14. Lt. Cond..								0	0				0		
15. Rd. Surface								0	0	X					
16. Dr. Impair.								0	0	0	0	0	0	0	0
18. Traffic								0	0	0	0	0	⊗	0	0
19. Rd. Use							X	0		0			X		
21. Post. Speed								0		0		⊗	0		0
22. Vehicle Speed												0	0	0	0
23. Vision Obs.											0	0	0	0	⊗
24. Dr. Dist.Inatt.											0	0		0	⊗
25. Left of Ctr.															X
26. Foll. Close															
27. Yield/Stop														X	
28. Turn/Sig.															X
29. Overtaking															

0 Variable Given in the Conjectured Causal Network

X Variable Selected by ALN Model

⊗ Variable Originally Conjectured and Selected by ALN Model

TABLE 5.2

ALN CLASSIFICATION RESULTS

Variable Number and Variable Identification		Classification Results In Percentage		
		Fitting	Selection	Evaluation
14	Light Condition	81	86	79
15	Road Surface	89	93	91
16	Driver Impairment	90	93	95
22	Vehicle Speed	70	75	64
23	Vision Obscured	62	54	46
24	Driver Distracted	55	58	52
25	Drove Left of Center	66	66	65
26	Followed Too Closely	63	55	42
27	Failure to Yield/Stop	59	63	56
28	Improper Turn	57	50	51
29	Improper Overtaking	76	69	71

By examining closely the data set, it can be seen why the performance of the ALN model was poor on Variable 26 (tailgating). Out of the 720 records, 711 of these records had zero value (not cited as a factor for the accident), and 9 records had other values. Hence, Variable 26 contained virtually no information in analyzing the causal network because this variable was not cited at all 99 percent (711/720) of the time.

The remainder of the ALN classification results varied from 50 percent to over 90 percent. The conclusion reached from computer analysis of the causal network was that the ALN methodology could indeed be used to assess the highway safety program effectiveness and to analyze the accident data base quantitatively.

The links that were found in the restructured Causal Network can be examined by highway safety planners to assess the effects of past and future actions.

5.3 EXAMPLE OF RESTRUCTURED CAUSAL NETWORK

An example of a restructured Causal Network is shown in Figure 5.16 (which is the same as Figure B.8) for variable x_{22}, "vehicle speed." It can be seen that 14 factors were conjectured originally to influence vehicle speed. Only two of these -- driver age and urban/rural -- were found to be needed. Therefore, this risk factor in the Causal Network could be predicted using only 2 of the 14 conjectured links, thereby reducing the data collection demands. The remainder of the 14 networks are given in Appendix B; the correspondence between Figure 2.2 and Figures B.1 through B.15 is summarized in Table 5.1.

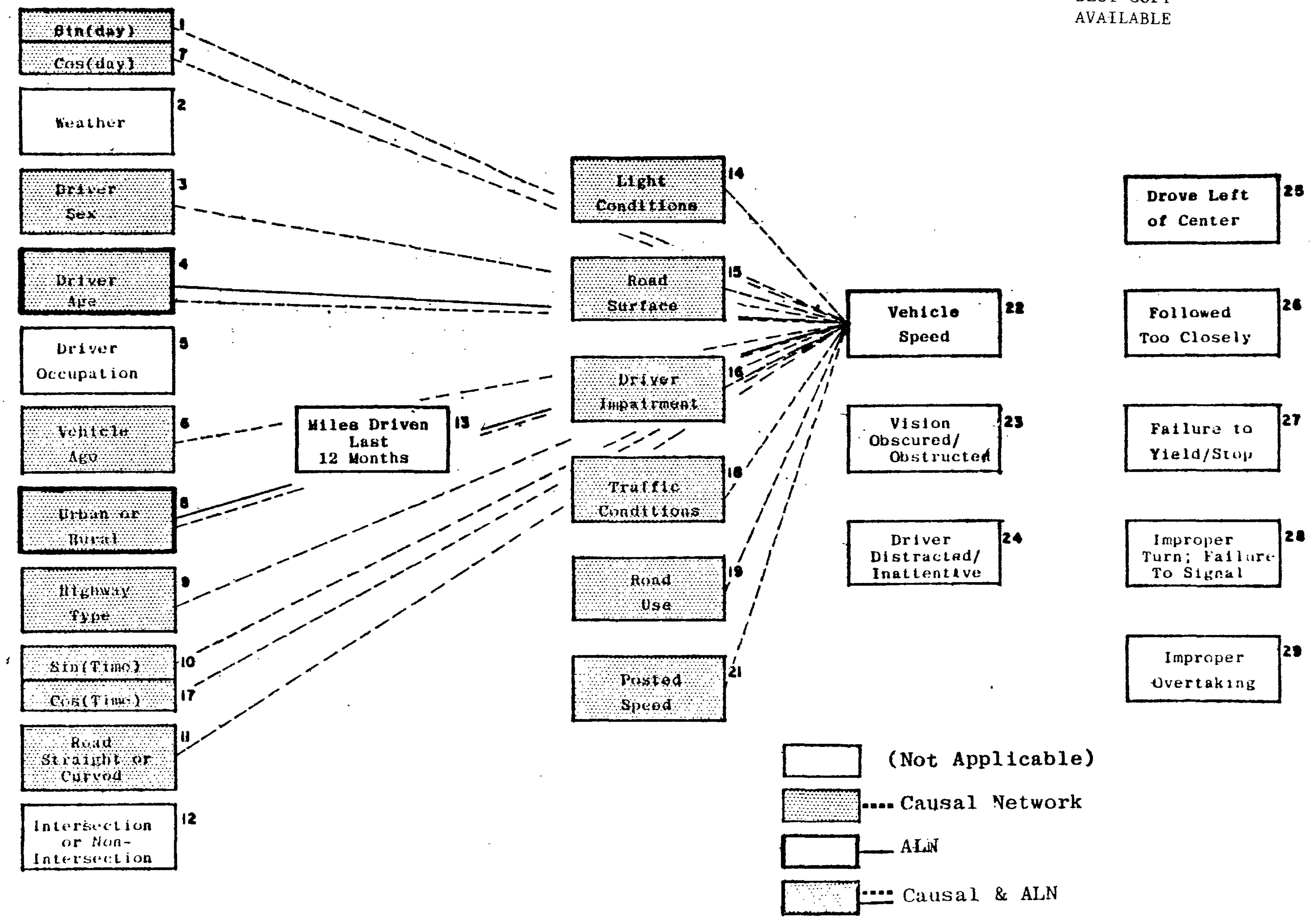


FIGURE 5.16: RESTRUCTURED CAUSAL NETWORK: VARIABLE 22 - VEHICLE SPEED

6. EFFECTS OF DRIVER AGE ON ACCIDENT-CAUSATIVE RISK FACTORS

6.1 RISK FACTORS INFLUENCED BY DRIVER AGE

Driver age is considered to be an important factor in highway accidents. For this reason it was decided to study this variable after the 15 models had been created to establish the quantitative relationship between driver age and the risk factors found to be influenced by it. This exercise also served to demonstrate the main objective of the project, which was to establish the utility of the ALN approach for making quantitative use of Causal Networks linking highway safety program outputs to accident involvement.

It was found, by ALN synthesis, that seven risk factors were influenced directly and two risk factors were influenced indirectly by driver age:

Direct Influence

- Traffic Conditions
- Road Use
- Vehicle Speed
- Vision Obstructed
- Tailgating
- Failure to Yield/Stop
- Improper Overtaking

Indirect Influence

- Posted Speed
- Improper Turn/Failure to Signal

Hence, using the appropriate ALN model, the effect of driver age on that risk factor could be studied quantitatively.

6.2 QUANTITATIVE EFFECT OF DRIVER AGE

For example, the ALN model for x_{22}, Vehicle Speed, evolved into a fairly simple structure of only a one-element network of the two inputs, driver age (DA) and urban/rural (U/R):

$$\text{Vehicle Speed} = w_0 + w_1 DA + w_2 U/R + w_3(DA)(U/R) + w_4(DA)^2 + w_5(U/R)^2$$

The coefficients w_3 to w_5 were found to be equal to zero, resulting in Vehicle Speed, VS, as a linear function of DA and U/R:

$$VS = 2.24 - 0.005(DA) - 0.397(U/R)$$

Since DA varied from 16 to 82 and U/R was binary, taking on values 1 or 2, the value of their respective coefficients did not reflect their relative importance on VS. To find this, each coefficient needed to be multiplied by the standard deviation of its associated variable, thus:

$$\frac{\Delta VS}{\Delta DA} = (-0.005)\ \sigma_{DA} = (-0.005)(14.99) = -0.072$$

and,

$$\frac{\Delta VS}{\Delta U/R} = (-0.397)\ \sigma U/R = (-0.397)(0.483) = -0.192$$

Therefore, the rate of change of VS with respect to DA (i.e., the first derivative) was -0.072 and with respect to U/R was -0.192, on a normalized basis.

In the latter regard, two items were of interest. First, both partial derivatives were negative, meaning that Vehicle Speed was found generally to decrease as driver age increased and/or as the driving was done in an urban setting. (The latter result followed from the coding of U/R as 1 for rural and 2 for urban; so as U/R increased, VS decreased, and U/R increased by shifting from a rural to urban road.)

Second, the ratio of U/R's effect to DA's effect on VS was 0.192/0.072, which is equal to 2.65. Hence, Vehicle Speed was considerably more influenced by the Urban/Rural risk factor than by the Driver Age exogenous variable.

The contour plot of Figure 6.1 shows the effect of DA and U/R in graphical form. The line represents the locus of points for which VS = 1.5, that is, the boundary between VS not being cited as an accident-causative risk factor (VS<1.5) and being cited (VS>1.5). It can be seen that DA has very little effect in causing VS to become accident-causative, VS>1.5, for a given value of U/R. However, when U/R is rural (U/R=1), VS is more often cited as an accident-causative factor (VS>1.5).

Three other contour plots are shown in Figures 6.2 - 6.4 for which DA effects the respective risk factor in a more complex, nonlinear manner.

In Figure 6.2, vision obstructed is the dependent variable and vehicle age, driver age and intersection/non-intersection were the independent variables. The ALN equation in this case was nonlinear in the three independent variables. Since there were three independent variables and one of them, Variable 12 (intersection/ non-intersection) was binary, the decision boundaries for vision obstructed were plotted separately for intersections and non-intersection. As expected, the probability that vision obstructed was cited as a factor for the accident at the intersection was higher than that at the non-intersection.

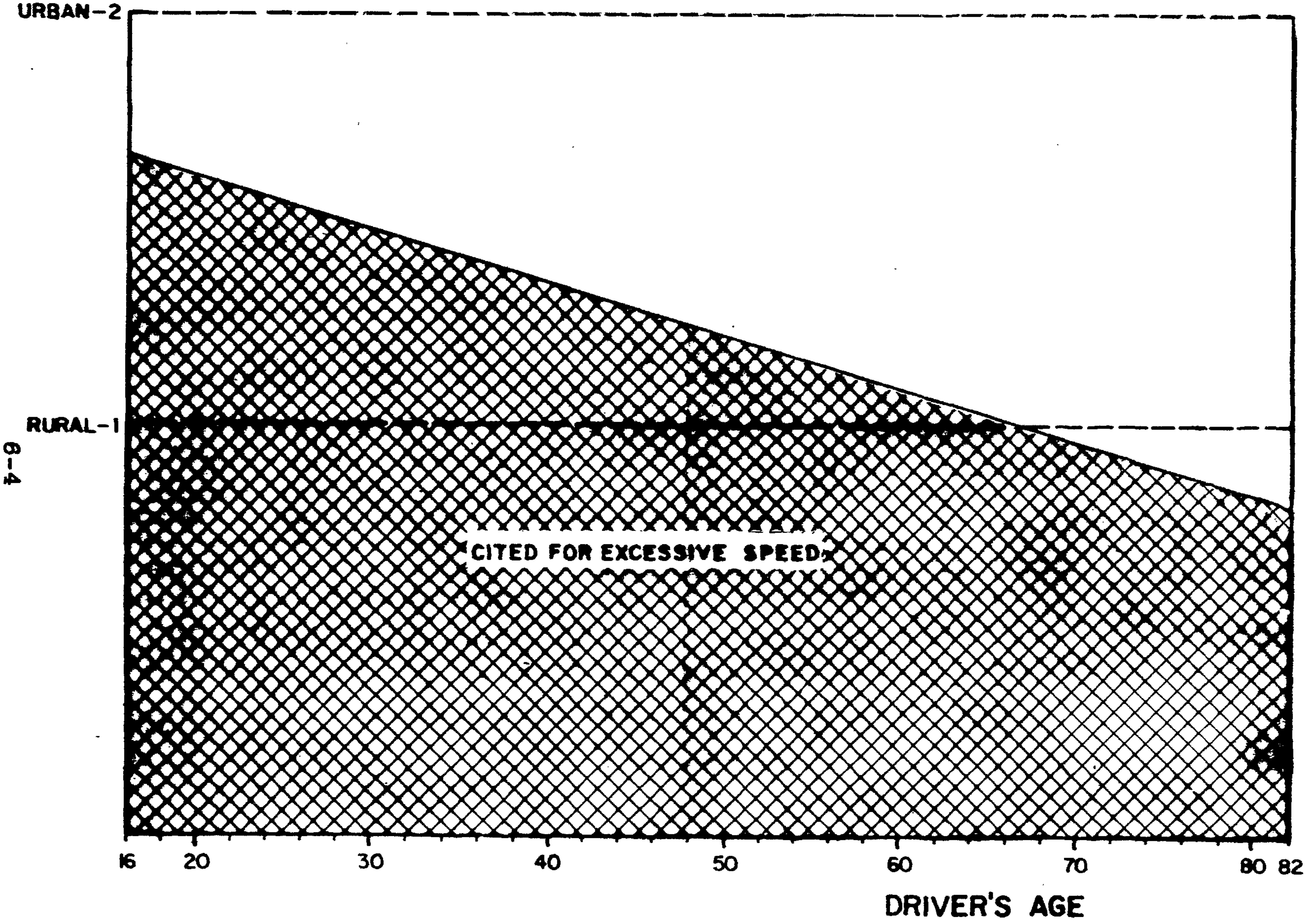


FIGURE 6.1: DECISION REGIONS FOR DEPENDENT VARIABLE 22 (EXCESSIVE SPEED -- FIG. 5-8)

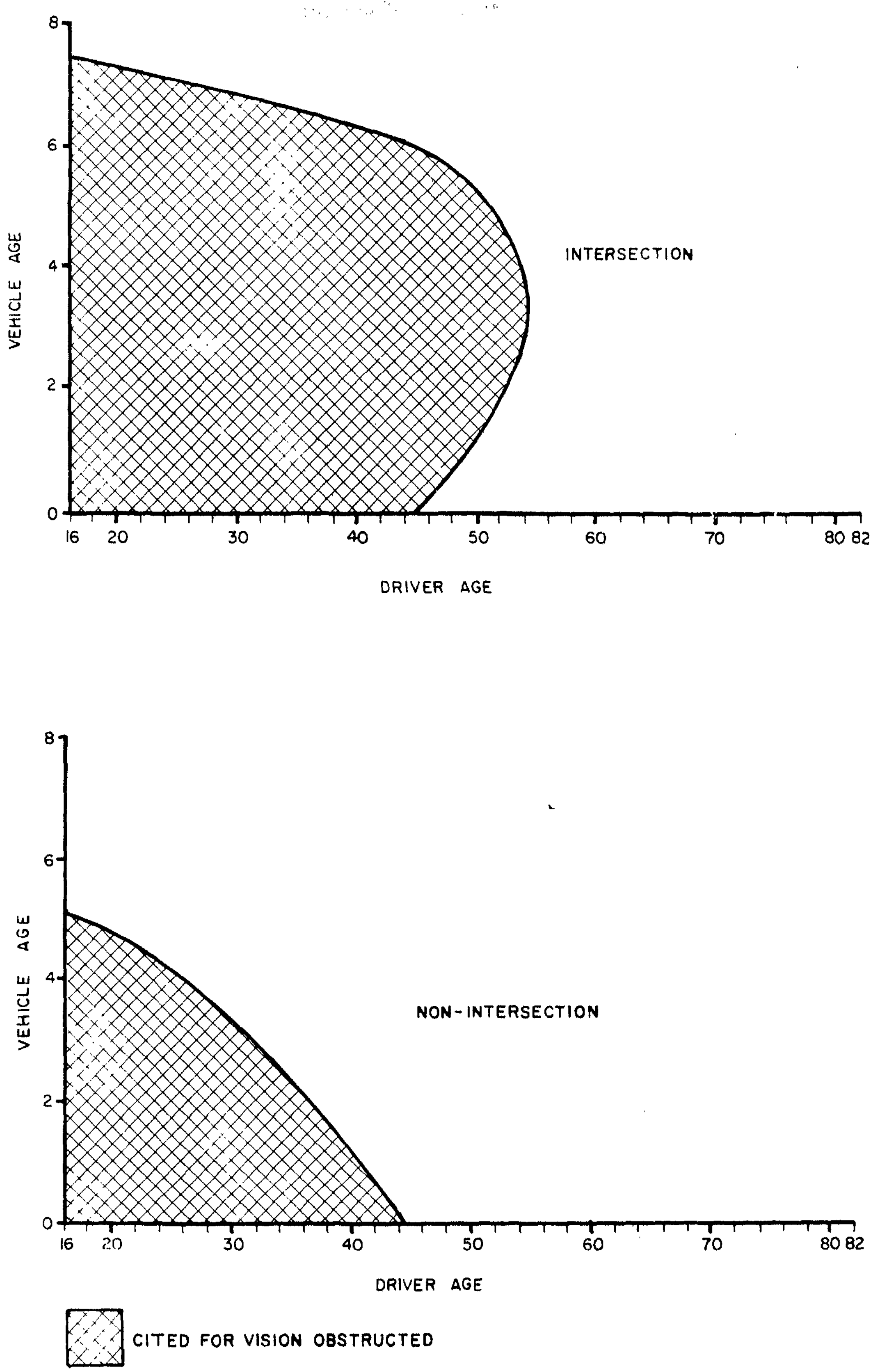

8
6
4
2
0
VEHICLE AGE
INTERSECTION
16
20
30
40
50
60
70
80 82
DRIVER AGE
8
6
4
2
0
VEHICLE AGE
NON-INTERSECTION
16
20
30
40
50
60
70
80 82
DRIVER AGE
▨ CITED FOR VISION OBSTRUCTED

Figure 6.3 is the contour plot for dependent Variable 26 (tailgating); the independent variables were posted speed and driver age. The results show that for drivers below 30 years of age and posted speeds less than 55 miles per hour, tailgating was likely to be cited as a factor contributing to the accident.

A complex and interesting contour plot resulted from the traffic condition model shown in Figure 6.4. The independent variables were driver age, day of week, time of day and road straight/ curved. Traffic condition was coded as 1 for heavy, 2 for moderate, 3 for light and 4 for none. (The contour plots for traffic conditions are for the mean value of 2.5.) Around the noon hour, traffic conditions were moderate-heavy -- regardless of day of the week and driver age. Similarly, during the midnight hours (hours 21 to 24 and 1 to 4), traffic conditions were none-light regardless of day of the week and driver age. The importance of this contour plot is that it gives the analyst a visual picture of the complex causal relationship between the dependent variable and the input independent variables.

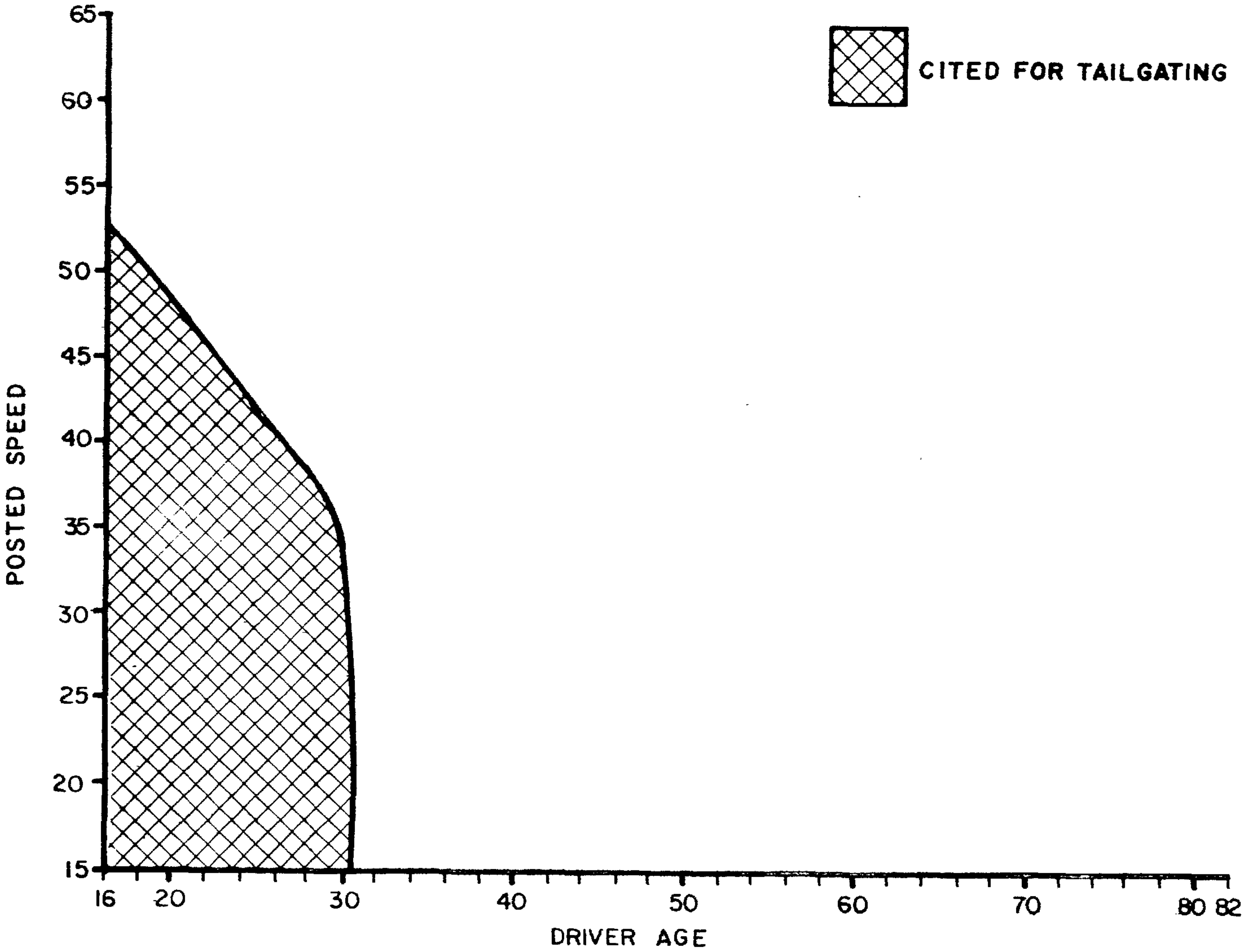


FIGURE 6.3: DECISION REGIONS FOR VARIABLE 26 (TAILGATING -- FIG. 5.12)

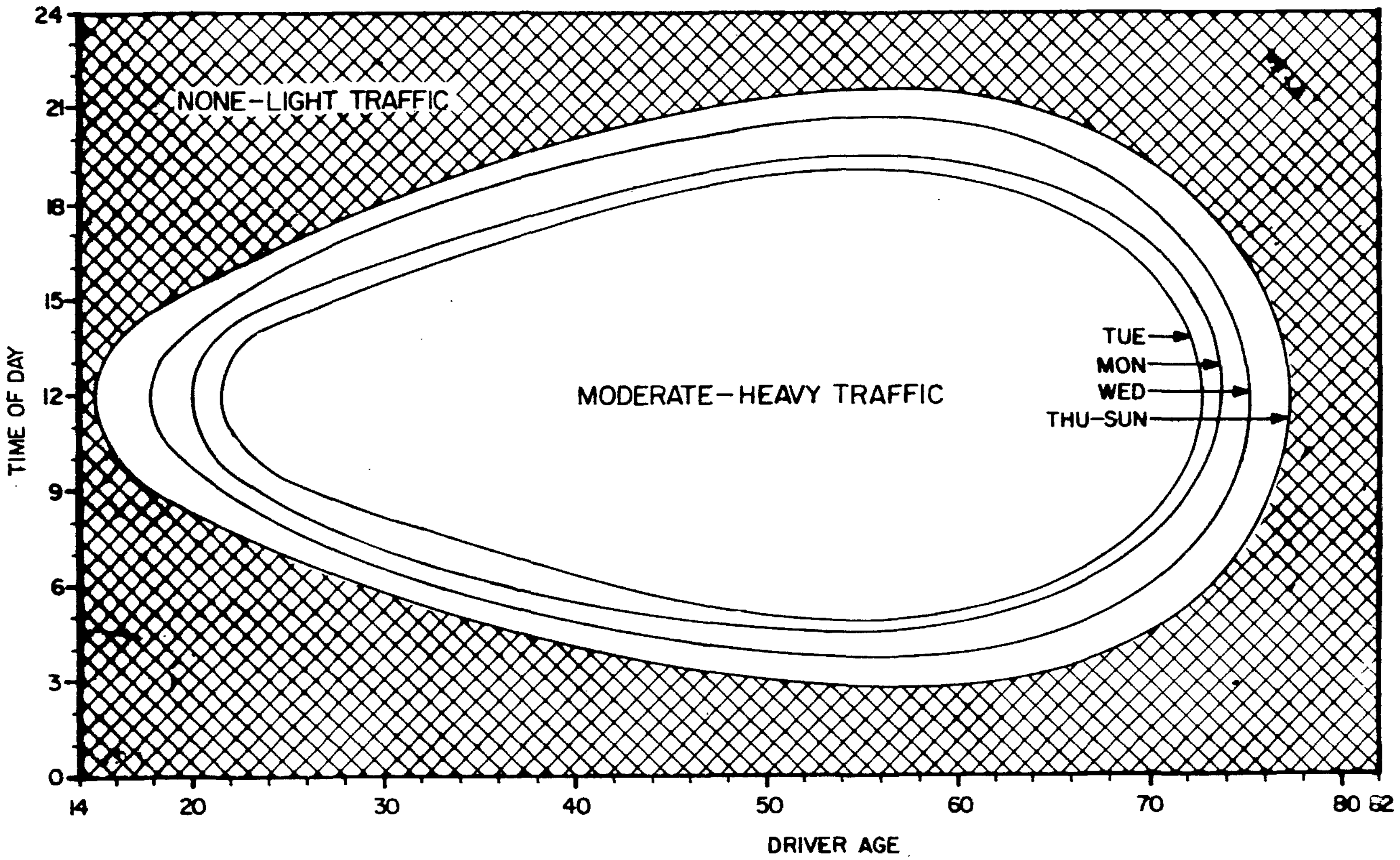


FIGURE 6.4: IMPORTANCE OF TRAFFIC CONDITIONS (FIG. 5.5)

7. REFERENCES

1. Overall, John E. and Klett, C. J. Applied Multivariate Analysis, McGraw Hill, 1972.

2. Gonzales, R. C. and Howington, L. C. "Machine Recognition of Abnormal Behavior in Nuclear Reactors", Proceedings The Third International Conference on Pattern Recognition, Coronado, California., References: Nov. 8-11, 1976, pp. 8-16.

3. Barron, R. L., "Theory and Application of Cybernetic Systems: An Overview," Proc. IEEE 1974 National Aerospace Electronics Conference (NAECON '74), Dayton, Ohio, May 13-15, 1974, pp. 107-118.

4. Mucciardi, A. N. and E. E. Gose, "An Automatic Clustering Algorithm and Its Properties in High-Dimensional Spaces," IEEE Trans. Computers, Vol. SMC-2, No. 2, April 1972, pp. 247-254.

5. Mucciardi, A. N., "Elements of Learning Control Systems With Applications to Industrial Processes," Proc. 1972 IEEE Conference on Decision and Control, New Orleans, La. December 13-15, 1972, pp. 320-325.

6. Mucciardi, A. N., "New Developments in Water Shed Pollution Forecasting Using Adaptive Trainable Networks, "Proc. 1974 IEEE Conference on Decision and Control, Dallas, Texas, October 2-4, 1974.

7. Mucciardi, A. N., "Adaptive Nonlinear Modeling for Ultrasonic Signal Processing," Proc. Interdisciplinary Workshop for Quantitative Flaw Definition, June 1974, D. O. Thompson, (ed.) AFML-TR-74-238, pp. 194-212.

8. "Tri-Level Study of the Causes of Traffic Accidents: Interim Report II." Volume 1: Causal Factor Tabulation and Trends. Volume 2: Radar and Anti-Lock Bracking Payoff Assessment. Institute for Research in Public Safety, Indiana University, December 31, 1974.

9. Chang, J. K. and Mucciardi, A. N. "Highway Safety Programs Effectiveness Model: Task I and II Short-Term Review Findings" Interim Report, Adaptronics, Inc. Nov. 26, 1976.

10. Joksch, H.C. "Construction of A Comprehensive Causal Network: Objectives and Scope of Phase I" The Center for The Environment and Man, Inc., Contract No. DOT-HS-6-01506.

APPENDIX A

CHARACTERISTICS OF THE ITADB HIGHWAY ACCIDENT DATA BASE

A.1 ANALYSIS OF THE INDIANA TRI-LEVEL ACCIDENT DATA BASE

In the Indiana tri-level accident data base subset, a total of 98 variables were recorded for each of 720 accidents. Table A.1 gives the descriptions of these 98 variables. However, only 29 variables (exogenous variables, risk factors, etc.) were shown in the Causal Network (Figure 2.2). The relationships between the 29 Causal Network variables and the 98 ITADB variables are given in Table A.2.

Variable 7 (wt/HP ratio), 10 (road separation), 17 (number of occupants), and 20 (traffic controls) of the Causal Network. were not recorded in the ITADB.

The 98 variables in the ITADB could be divided into the following five types of variable :

Type 1 - Informational Variables
Traffic Units, Day of Week, etc.

Type 2 - Environmental Variables
Weather Condition, Condition of Road Surface, etc.

Type 3 - Exogenous Variables
Age, Sex, Marital Status, etc.

Type 4 - Numerical Variables
Speed Limit, Frequency of Driving a Particular Road, etc.

Type 5 - Risk Factor Variables
Recognition Error, Inattention, Position of Car on Road, etc.

TABLE A.1

INDIANA TRI-LEVEL ACCIDENT DATA BASE VARIABLES DESCRIPTION

Variable Number	Description
P01	Phase No. (2,3,4,5)
P03	Number of Traffic Units (1,2,3,4)
P06	Traffic Unit Number
P08	Day of Week of Accident
P09	Hour of Day of Accident
P10	Condition of Road Surface
P11	Weather Conditions
P12	Urbanization at Accident Location
P13	Highway Classification
P14	Accident Location Classification
P15	Character of Road-Horizontal
P16	Light Conditions
P17	Type of Road Surface
P18	Speed Limit at Accident Location
P19	Sex of Vehicle Driver
P20	Age of Vehicle Driver
P21	Occupation of Vehicle Driver
P22	100's of Miles Driven in Last 12 Months
P23	Age of Vehicle
P24	Drugs Taken Within 48 Hours of Accident
P25	Alcohol Consumed Within 24 Hours of Accident
P26	Traffic Conditions at Time of Accident

Table A.1: (Continued)

Variable Number	Description
P27	Frequent Driving Road
P28	Recognition Errors
P29	- Driver Failed to Observe, Stop for Stop Sign
P30	- Recognition Delays - Reason Identified
P31	* Inattention
P32	- Traffic Stopped, Slowing
P33	- Position of Car on Road
P34	- Road Features - e.g., curve, lane
P35	- Road Signs, Signals
P36	- Cross-Flowing Traffic
P37	- Inattention - Other
P38	* Internal Distraction
P39	- Event in Car - e.g., Sudden Noise
P40	- Radio, Tape Adjustment
P41	- Window Adjustment
P42	- Conversation with Passenger
P43	- Internal Distraction - Other
P44	* External Distraction
P45	- Other Traffic
P46	- Driver - Selected Outside Activity
P47	- Activity of Interest Outside Vehicle
P48	- Sudden Event Outside Vehicle
P49	- External Distraction - Other
P50	* Improper Lookout
P51	- Pulling Out from Parking Space
P52	- Entering Traffic from Street, Alley
P53	- Prior to Changing Lanes, Passing
P54	- Improper Lookout - Other
P55	* Perception Delays - Other, Unknown
P56	- Traffic Stopped, Slowing
P57	- Position of Car on Road
P58	- Road Features - e.g., Curve, Lane
P59	- Road Signs, Signals
P60	- Cross-Flowing Traffic
P61	- Perception Delays - Other
P62	Comprehension, Reaction Delays
P63	- Delayed Comprehension
P64	- Delayed Reaction

Table A.1: (Continued)

Variable Number	Description
P65	Improper Maneuver
P66	- Turned From Wrong Lane
P67	- Drove in Wrong Lane for Direction
P68	- Drove in Wrong Direction of Travel
P69	- Passed at Improper Location
P70	- Improper Maneuver - Other
P71	Improper Driving Technique
P72	- Cresting Hills - Driving in Center Road
P73	- Breaking Too Late, Inappropriately
P74	- Stopping Too Far Out in Intersection
P75	- Driving Too Close to Center Line, Edge
P76	- Slowed Too Rapidly
P77	- Improper Driving Technique - Other
P78	Excessive Speed
P79	- For Road Design - Regardless of Traffic
P80	- In Light of Traffic, Pedestrians
P81	- In Light of Weather Conditions
P82	- Combination of Design, Traffic, Weather
P83	- Excessive Speed - Other
P84	Tailgating
P85	Inadequate Signal
P86	- Failure to Signal for Turn
P87	- Failure to Use Horn to Warn
P88	- Inadequate Signal - Other
P89	Alcohol Impairment
P90	Other Drug Impairment
P91	Fatigue

TABLE A.1: (Continued)

Variable Number	Description
P92	View Obstructions
P93	- Hillcrests, Dips, etc.
P94	- Roadside Embankments, Escarpments
P95	- Roadside Structures and Growth
P96	- Stopped Traffic
P97	- Parked Traffic
P98	- View Obstructions - Other

TABLE A.2

RELATIONSHIP BETWEEN CAUSAL NETWORK VARIABLES AND INDIANA DATA BASE VARIABLES

Causal Network Variable Number	Related Variables from Indiana Data Base
1	P08, P09
2	P11
3	P19
4	P20
5	P21
6	P23
7	-
8	P12
9	P13
10	-
11	P15
12	P14
13	P22
14	P16
15	P10
16	P89, P90, P91
17	-
18	P26
19	P27
20	-
21	P18
22	P78, P79, P80, P81, P82, P83
23	P92, P93, P94, P95, P96, P97, P98
24	P30, P31, P38, P44, P50, P55
25	P33, P57, P67, P68, P72, P75
26	P84
27	P29, P36, P51, P52, P60
28	P66, P85
29	P53, P69

The above five variable types were not mutually exclusive. For instance, the age variable was an exogenous variable (Type 3) as well as a numerical variable (Type 4). A partitioning of the 98 variables of the ITADB into the five variable types is shown in Table A.3.

Upon examination of the ITADB, a number of problems was revealed: (i) missing or unknown variables, (ii) unbalanced distributions of variables, and (iii) method of coding variables.

If the variable was missing or unknown, one of the two following methods could have been used to assign the missing value:

(I) Sample Average Method - The missing variable could have been estimated by the sample average from those records or observations similar to the missing one.

(II) Monte Carlo Method - The missing variable could have been replaced by the outcome of a random experiment whose probability distribution was the frequency of occurrence of this variable in the data base.

The Monte Carlo Method was used in this investigation.

In the simulation of the Causal Network by the ALN modeling approach, the values of variables Types 1, 3, and 4 were used directly. The Type 2 (environmental) variables and the Type 5 (risk factor) variables were modified prior to the simulations as follows:

(I) No Transformation - The value of the variable as coded in the ITADB was used.

TABLE A.3

TYPES OF VARIABLE IN INDIANA DATA BASE

Type	Variables	Description
1	P01 to P09	Informational Variables
2	P10 to P17	Environmental Variables
3	P19, P21	Exogenous Variables
4	P18, P22, P23, P27	Numerical Variables
5	P28 to P98	Risk Factor Variables

For example, variable P11 (weather condition) in the ITADB was coded as "1" for clear, "2" for rain, "3" for snow, "4" for fog, and "8" for other. The same coding was used in the highway accident data in this study when no transformation was used.

(II) Counting Method - This method set the value of the variable equal to one plus the number of cited ITADB variables that were related to this variable (Table A.2).

The counting method was chosen to code variables 22 to 25 and 27 to 29 in this study because of the small data base in ITADB and unbalanced distributed variables. Variable 27 (failure to yield/stop) was related to variable P36 (cross-flowing traffic), etc. P36 is the variable related to the ITADB and indicated by the prefix P. P36 was not cited 711 times and cited only 9 times as a factor for the accident in the ITADB. Hence, this variable P36 was highly unbalanced.

The complete summary of the ITADB is given in Table A.4. Column 1 is the variable number related to the Causal Network and Column 3 is the corresponding variable in ITADB. The number of possibly different values a particular variable could achieve is given in Column 5. The frequency of missing values is listed in Column 6.

A.2 TYPE OF CODING FOR THE HIGHWAY DATA BASE USED IN THIS STUDY

Variable 1 - Day and Time: Day and Time were replaced by the following four variables to avoid discontinuities between the seventh and first days and between 2400 and 0001 hours, respectively:

Variable Number	Description	Related Variables From IDB	Description	Number of Different Values	Frequency of Missing Values	Values	Frequency of Values
1	sin ($\frac{2\pi}{7}$day)	P08	Day of the Week	7	21	1 Mon	96
						2 Tue	120
						3 Wed	129
						4 Thur	111
						5 Fri	106
						6 Sat	73
						7 Sun	64
2	Weather	P11	Weather Conditions	5	37	1 clear	520
						2 rain	135
						3 snow	18
						4 fog	1
						8 other	9
3	Driver Sex	P19	Driver Sex	2	55	1 male	130
						2 female	235
4	Driver Age	P20	Driver Age	53	60	See attached table	
5	Driver Occupation	P21	Occupation	9	64	1 farmer	2
						2 laborer	81
						3 semi-skilled	62
						4 skilled	102
						5 white-collar	18
						6 professional	112
						7 student	178
						8 housewife	43
						9 other	28
6	Vehicle Age	P23		22	69	0	18
						1	74
						2	75
						3	72
						4	56
						5	76
						6	56
						7	52
						8	46
						9	16
						10	28
						11	19
						12	11
						13	6
						14	5
						15	1
						16	2
						17	1
						19	1
						20	3
						21	1
						23	2
7	cos ($\frac{2\pi}{7}$day)	P08	Day	7	21	--Same as Variable 1--	
8	Urban/Rural	P12	Urban/Rural	2	24	0 rural	229
						1 urban	467
9	Highway Type	P13	Highway Type	2	21	0 county,city	569
						1 state	130
10	sin ($\frac{2\pi}{24}$time)	P09	Time of Day	24	21	1 am	6
						2	12
						3	3
						4	1
						5	3
						6	7
						7	1
						8	10
						9	25
						10	16
						11	18
						12 noon	52
						13 1 pm	66
						14	17
						15	64
						16	104
						17	94
						18	64
						19	9
						20	32
						21	31
						22	11
						23 11 pm	15
						24 midnight	5

Table A.4 - (Continued)

Variable Number	Description	Related Variables From IDB	Description	Number of Different Values	Frequency of Missing Values	Values	Frequency of Values
11	Road Straight or Curved	P15	Road Straight or Curved	2	44	1 straight	556
						2 curved	109
12	Intersection or Non Intersection	P14	Location Classification	6	22	1 intersection roadway	342
						2 culvert int.	1
						3 non road	92
						4 RR Crossing	2
						5 bridge over-pass	1
						8 other	260
13	Miles driven - last 12 months	P22	Miles driven-last 12 mo.	59	134	--See Attached Table --	
14	Light Conditions	P16	Light Condition	3	37	1 day	556
						2 dark	110
						3 dawn or dusk	17
15	Road Surface	P10	Road Surface	4	35	1 dry	487
						2 wet	166
						3 snow,ice	28
						8 other	4
16	Driver Impairment	P89	Due to alcohol	4	0	0 N/C	696
						10 Cas-Pos.	11
						20 Cas-Prob.	11
						30 Cas-Certain	2
		P90	Due to Drugs	4	0	0	703
						10	9
						20	6
						30	2
		P91	Due to Fatigue	4		0	703
						10	10
						20	6
						30	1
17	cos $(\frac{2\pi}{24}\text{time})$	P09	Time of Day	24	21	--Same as Variable 10--	
18	Traffic	P26	Traffic Condition	5	28	1 heavy	62
						2 moderate	77
						3 light	89
						4 none	97
						8 could not	367
19	Road Use	P27	Frequency Driving Road	7	72	1 daily	311
						2 2/week	108
						3 1/week	60
						4 2/mo.	25
						5 1/mo.	29
						6 seldom	85
						7 first time	30
21	Posted Speed	P18	Speed Limit	10	154	1 20 mph	35
						2 25 mph	5
						3 30 mph	344
						4 35 mph	47
						5 40 mph	21
						6 45 mph	70
						7 50 mph	17
						8 55 mph	13
						10 65 mph	12
						11 other	2
22	Vehicle Speed-Speed too fast	P78	Excessive Speed	5	0	0	636
						2	4
						10	14
						20	33
						30	33
		P79	For Road Design, not traffic	5	0	0	670
						2	4
						10	7
						20	14
						30	25
		P80	In Light Traffic, Pedestrians	3	0	0	715
						10	2
						30	3
		P81	In Light of Weather Cond.	4	0	0	705
						10	3
						20	11
						30	1

Table A.4 - (Continued)

Variable Number	Description	Related Variables From IDB	Description	Number of Different Values	Frequency of Missing Values	Values	Frequency of Values
22 Continued		P82	Comb. Design, Traffic and Weather	4	0	0 10 20 30	719 1 6 3
		P83	Other	4	0	0 10 20 30	716 1 2 1
23	Vision Obscured	P92	View Obstruction	6	0	0 1 2 10 20 30	611 1 2 21 56 26
		P93	Hillcrests, dips, etc.	3	0	0 20 30	712 4 1
		P94	Roadside Embankments,etc.	4	0	0 10 20 30	703 1 6 7
		P95	Roadside Structures & Growth	5	0	0 2 10 20 30	678 2 5 25 6
		P96	Stopped Traffic	4	0	0 10 20 30	703 5 7 5
		P97	Parked Traffic	5	0	0 1 10 20 30	693 1 3 17 6
		P98	Other View Obstructions	4	0	0 10 20 30	711 4 2 3
24	Driver Distracted, Inattentive	P30	Recognition Delay-Reason Indent.	4	0	0 10 20 30	471 31 60 155
		P31	Inattention	4	0	0 10 20 30	635 19 23 13
		P38	Internal Distraction	4	0	0 10 20 30	674 8 14 24
		P44	External Distraction	4	0	0 10 20 30	697 5 1 11
		P50	Improper Lookout	4	0	0 10 20 30	615 2 26 77
		P55	Perception Delays - other unknown	5	0	0 1 10 20 30	675 1 12 14 18
25	Drove Left of Center	P33	Position of Car on Road	4	0	0 10 20 30	708 [illegible] [illegible] [illegible]
		P57	Position of Car on Road	2	0	0 10	[illegible] 1
		P67	Drove in Wrong Lane for Direction	2	0	0 30	718 2

Table A.4 - (Continued)

Variable Number	Description	Related Variables From IDB	Description	Number of Different Values	Frequency of Missing Values	Values	Frequency of Values
25 (Continued)		P66	Turned From Wrong Direction of Travel	4	0	0 10 20 30	717 1 1 1
		P72	Cresting Hills-Driving in Center of Road	4	0	0 10 20 30	712 1 4 3
		P75	Driving to Close Center Line, Edge	3	0	0 10 20	716 1 3
26	Followed too Closely	P84	Tailgating	4	0	0 10 20 30	711 4 4 1
27	Failure to Yield/Stop	P29	Driver Fail obs. or Stop for Stop Sign	4	0	0 10 20 30	689 3 2 26
		P36	Cross-Flowing Traffic	4	0	0 10 20 30	710 2 3 5
		P51	Pulling Out from Parking Space	3	0	0 20 30	714 1 5
		P52	Entering Traffic from Street, Ally	4	0	0 10 20 30	649 1 19 51
		P60	Cross-Flowing Traffic	4	0	0 10 20 30	712 2 3 3
28	Improper Turn, Failure to Signal	P66	Turned From Wrong Lane	3	0	0 10 30	710 1 9
		P85	Inadequate Signal	5	0	0 1 10 20 30	687 1 20 10 2
29	Improper Overtaking	P53	Prior to Changing Lanes,Passing	3	0	0 20 30	710 1 9
		P69	Passed at Improper Location	3	0	0 20 30	713 2 5

Variable 4 - Driver Age Frequency

Age	Frequency	Age	Frequency	Age	Frequency
16	21	34	4	52	10
17	35	35	7	53	7
18	38	36	9	54	5
19	49	37	6	55	5
20	53	38	8	56	2
21	54	39	6	57	1
22	41	40	7	59	4
23	23	41	10	63	1
24	29	42	6	64	4
25	27	43	4	65	6
26	21	44	6	66	3
27	27	45	9	67	1
28	16	46	6	68	2
29	11	47	4	69	3
30	7	48	10	71	3
31	11	49	4	74	2
32	10	50	5	82	1
33	12	51	4		

Variable 13 - Miles Driven Last 12 Months

Miles	Frequency	Miles	Frequency	Miles	Frequency
100	2	7,500	3	22,500	1
200	2	8,000	21	23,000	4
500	4	9,000	10	24,000	3
600	1	9,500	2	25,000	25
1,000	11	10,000	102	27,500	1
1,200	4	10,500	2	29,000	1
1,500	2	11,000	9	30,000	17
1,800	1	12,000	55	35,000	7
2,000	12	12,500	2	38,000	1
2,500	3	13,000	12	40,000	7
3,000	12	13,500	1	45,000	1
3,500	3	14,000	9	50,000	8
4,000	12	15,000	68	65,000	1
4,500	1	16,000	6	70,000	1
5,000	34	17,000	3	75,000	1
5,500	1	17,500	7	80,000	1
6,000	18	18,000	3	82,000	1
6,500	1	19,000	1	100,000	4
7,000	13	20,000	44	127,000	1
7,200	1	22,000	2		

$$\sin\left(\frac{2\pi}{7}\text{ Day}\right) = \text{Variable 1}$$

$$\cos\left(\frac{2\pi}{7}\text{ Day}\right) = \text{Variable 7}$$

$$\sin\left(\frac{2\pi}{24}\text{ Time}\right) = \text{Variable 10}$$

$$\cos\left(\frac{2\pi}{24}\text{ Time}\right) = \text{Variable 17}$$

The values of Day were 1 (Monday), 2 (Tuesday), 3 (Wednesday, ...,7 (Sunday). The values of Time were 1 (One A.M., 2, 3, ..., 12 (Noon), ..., 24 (Midnight). (Note that variables 7, 10, and 17 in the conjectured Causal Network had no corresponding variables in ITADB. Hence these variable number positions were used as shown in the above equations.)

Variable 2 - Weather: Coded as binary -- 1 for clear and 2 for not clear.

Variable 5 - Driver Occupation: Replaced by 9 binary variables, numbered 30 to 38. Variable 30 took on a value of 1 for non-farmer and 2 for farmer. Variable 31 took on a value of 1 for non-laborer and 2 for laborer, etc.

Variable 8 - Urban/Rural: Coded as 1 for rural and 2 for urban.

Variable 9 - Highway Type: Coded as 1 for county/city and 2 for state.

Variable 12 - Intersection or Non-Intersection: Coded as 1 for intersection or non-road, and 2 for culvert intersection, railroad crossing, bridge overpass, or other.

Variable 13 - Miles Driven Last 12 Months: Coded in units of 100 miles driven in the past 12 months. Any value greater than 500 units was coded as 600; only ten records out of 720 had more than 50,000 miles driven in the last 12 months.

Variable 14 - Light Conditions: Coded as 1 for day and 2 for dawn, dusk, or dark.

Varibale 15 - Road Surface: Coded as 1 for dry road surface and 2 for wet, snow, ice, or other adverse conditions.

Variable 18 - Traffic: Coded as 1 for heavy, 2 for moderate, 3 for light, and 4 for none. When the value of this variable was equal to 8, meaning it could not be determined, this variable was changed to be interpreted as "unknown" (i.e., its value was generated by the Monte Carlo Method).

Variable 16 - Driver Impairment: Coded as 1 plus the number of cited ITADB "P" variables related to this variable.

Variables 22 to 25 and 27 to 29: Coded as 1 plus the number of cited "P" variables related to these variable numbers.

Variable 26 - Followed Too Closely: Coded as 1 for not cited, and 2 for cited.

The coding of the accident data is summarized in Table A.5. This accident data base was used to simulate the Causal Network in analyzing highway safety program effectiveness using the ALN modeling approach. Representative computer results were discussed in Section 5.

TABLE A-5
ACCIDENT DATA BASE FOR ALN MODELING

Variable Number	Description	Values	Frequency of Values
1	$\sin(\frac{2\pi}{7}\text{day})$	1 Mon	96
		2 Tue	120
		3 Wed	129
		4 Thurs	111
		5 Fri	106
		6 Sat	73
		7 Sun	64
2	Weather	1 Clear	520
		2 Other	153
3	Driver Sex	--Same as before --	
4	Driver Age	--Same as before --	
5	Driver Occupation Replaced by Variables 30 to 38		
6	Vehicle Age	--Same as before --	
7	$\cos(\frac{2\pi}{7}\text{day})$	Same as Variable 1	
8	Urban/Rural	1 rural	229
		2 urban	467
9	Highway Type	1 County,City	569
		2 State	130
10	$\sin(\frac{2\pi}{24}\text{time})$	--Same as before --	
11	Road Straight or Curved	--Same as before --	
12	Intersection or Non-Intersection	1 Intersection	434
		2 Non-intersection	264
13	Miles driven-last 12 months	Same as before except 60,000	10
14	Light Conditions	1 day	556
		2 other	127

TABLE A.5 (continued)

Variable Number	Description	Values	Frequency of Values
15	Road Surface	1 dry	487
		2 other	198
16	Driver Impairment		
	Due to Alcohol	1 N/C	696
		2 Cited	24
	Due to Drugs	1 N/C	703
		2 Cited	17
	Due to Fatigue	1 N/C	703
		2 Cited	17
17	$\cos(\frac{2\pi}{24}\text{time})$	--Same as before--	
18	Traffic	1 heavy	62
		2 moderate	77
		3 light	89
		4 none	97
19	Road Use	--Same as before--	
21	Posted Speed	--Same as before--	
22	Vehicle Speed - Speed too Fast	1 N/C	636
	Excessive Speed	2 cited	84
	For Road Design, not traffic	1 N/C	670
		2 cited	50
	In Light Traffic, pedestrians	1 N/C	715
		2 cited	5
	In Light of Weather Condition	1 N/C	705
		2 cited	15
	Comb. Design, Traffic and Weather	1 N/C	710
		2 cited	10
	Other	1 N/C	716
		2 cited	4

TABLE A.5 (continued)

Variable Number	Description	Values	Frequency of Values
23	Vision Obscured		
	View Obstruction	1 N/C 2 cited	614 106
	Hillcrests, dips, etc.	1 N/C 2 cited	712 8
	Roadside Embankments,etc.	1 N/C 2 cited	703 17
	Roadwide Structures and Growth	1 N/C 2 cited	678 38
	Stopped Traffic	1 N/C 2 cited	703 17
	Parked Traffic	1 N/C 2 cited	693 27
	Other View Obstructions	1 N/C 2 cited	711 9
24	Driver Distracted, Inattentive		
	Recognition Delay-Reason Indent.	1 N/C 2 cited	474 246
	Inattention	1 N/C 2 cited	635 85
	Internal Distraction	1 N/C 2 cited	674 46
	External Distraction	1 N/C 2 cited	697 23
	Improper Lookout	1 N/C 2 cited	615 105
	Perception Delays - other unknown	1 N/C 2 cited	675 45
25	Drove Left of Center		
	Position of Car on Road	1 N/C 2 cited	708 12
	Position of Car on Road	1 N/C 2 cited	719 1

TABLE A.5 (continued)

Variable Number	Description	Values	Frequency of Values
25 continued	Drove in Wrong Lane for Direction	1 N/C 2 cited	718 2
	Turned From Wrong Direction of Travel	1 2	717 3
	Cresting Hills-Driving in Center of Road	1 2	712 8
	Driving too Close Center Line, Edge	1 2	716 4
26	Followed too Closely		
	Tailgating	1 2	711 9
27	Failure to Yield/Stop		
	Driver Fail obs. or Stop for Stop Sign	1 2	689 31
	Cross-Flowing Traffic	1 2	710 10
	Pulling Out from Parking Space	1 2	714 6
	Entering Traffic from Street, Ally	1 2	649 71
	Cross-Flowing Traffic	1 2	712 8
28	Improper Turn, Failure to Signal		
	Turned from Wrong Lane	1 2	710 10
	Inadequate Signal	1 2	687 33
29	Improper Overtaking		
	Prior to Changing Lanes, Passing	1 2	710 10
	Passed at Improper Location	1 2	713 7

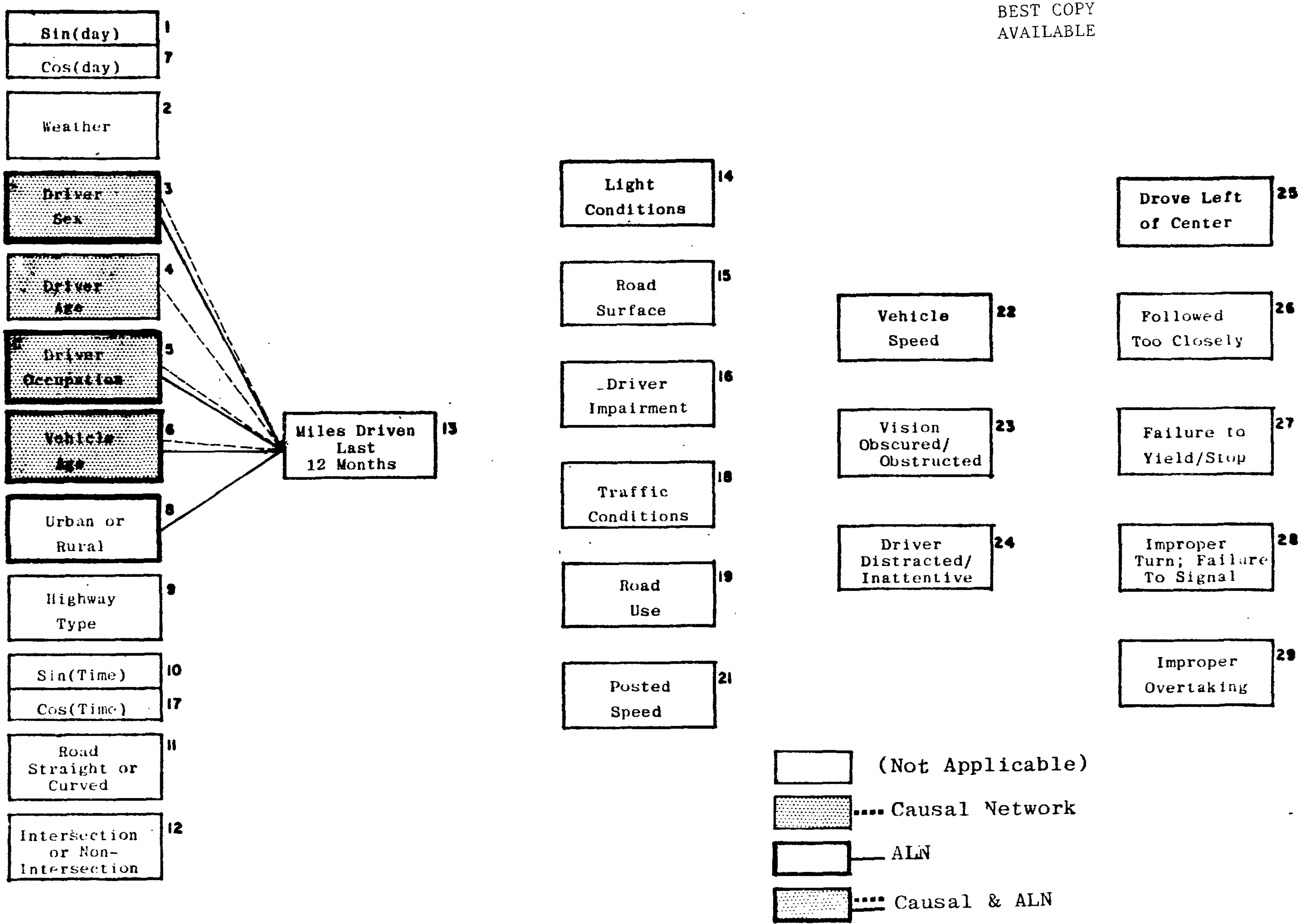


FIGURE B.1: RESTRUCTURED CAUSAL NETWORK: VARIABLE 13 - MILES DRIVEN LAST 12 MONTHS

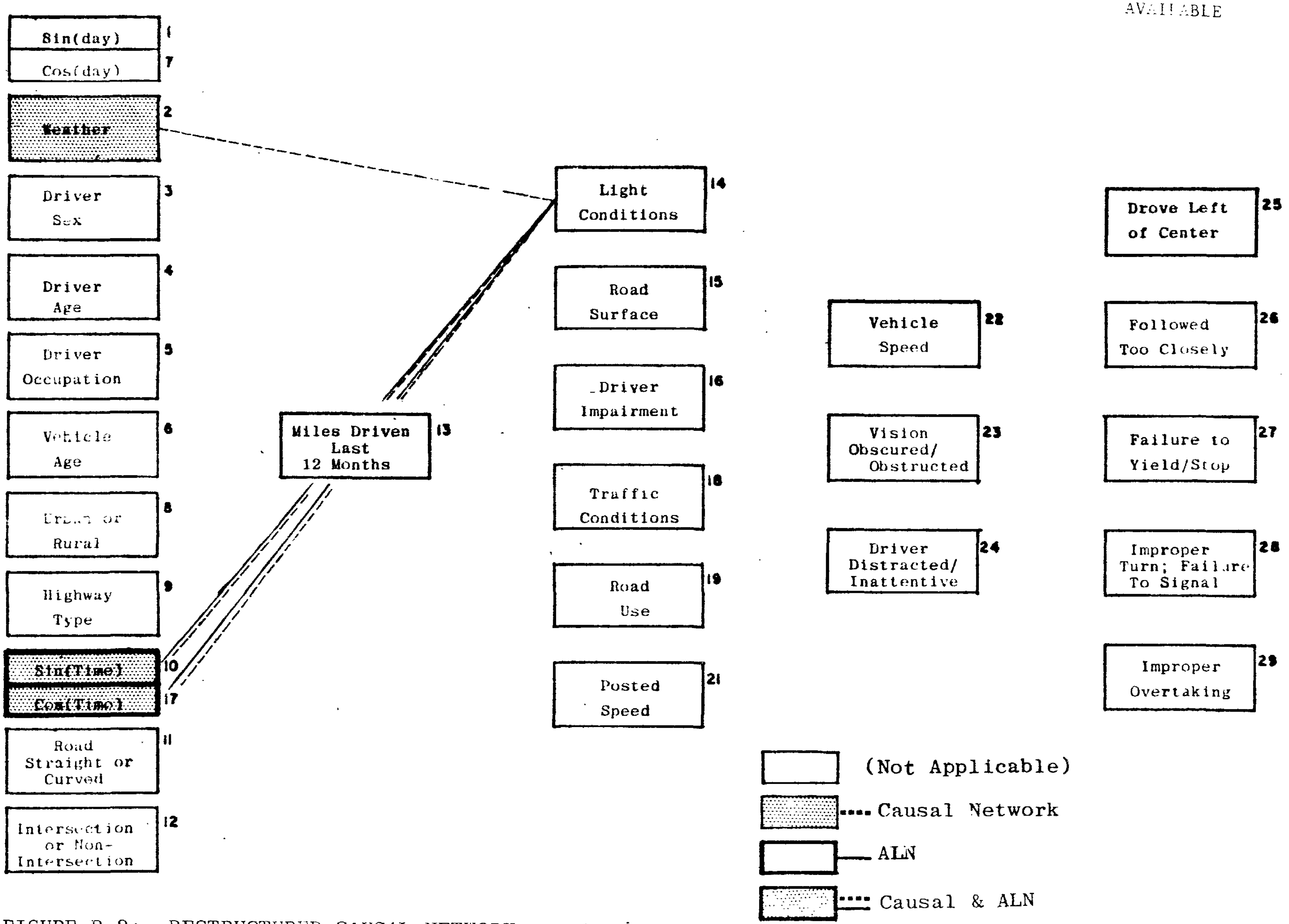


FIGURE B.2: RESTRUCTURED CAUSAL NETWORK: VARIABLE 14 - LIGHT CONDITIONS

BEST COPY
AVAILABLE

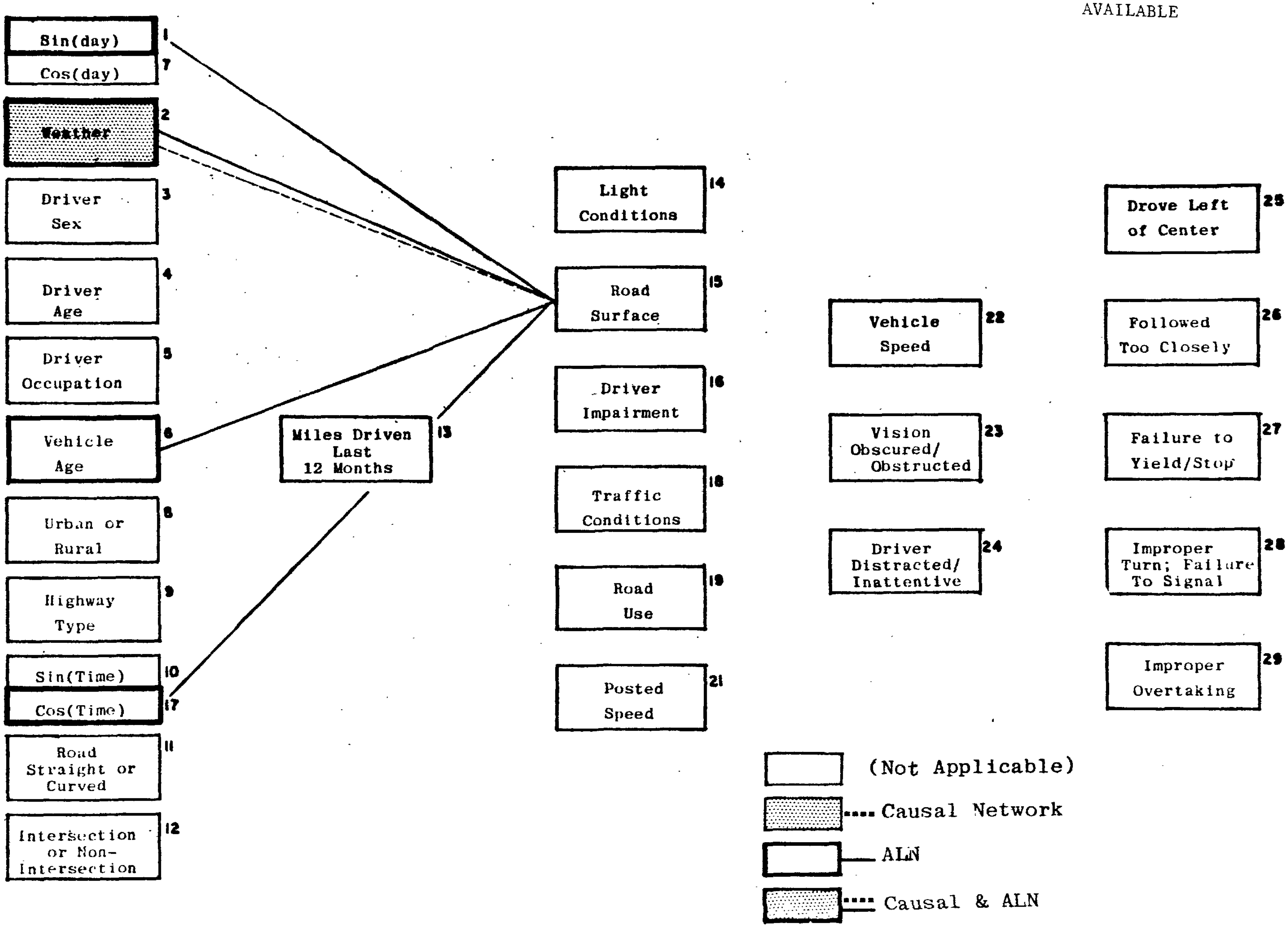


FIGURE B.3: RESTRUCTURED CAUSAL NETWORK: VARIABLE 15 - ROAD SURFACE

BEST COPY
AVAILABLE

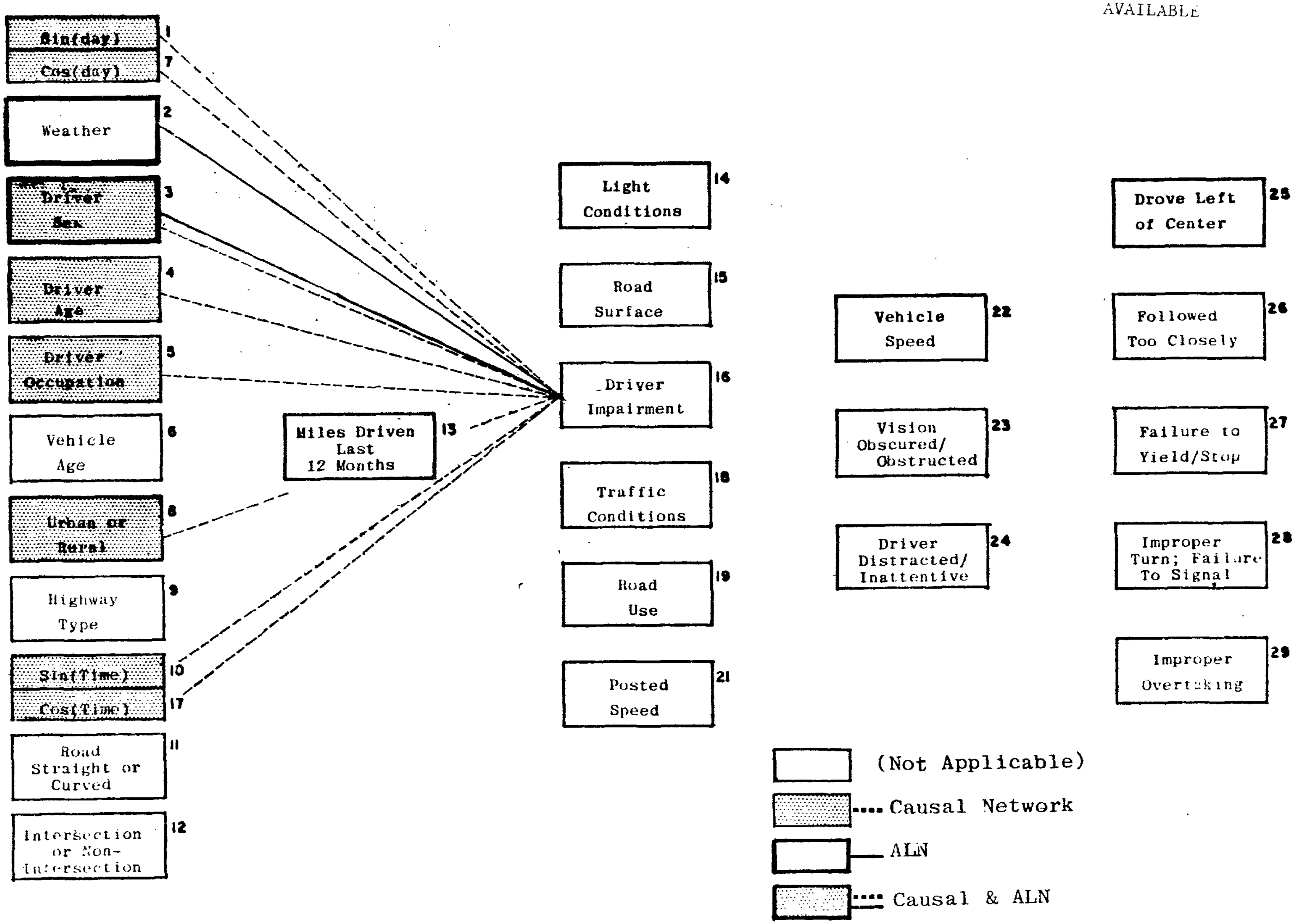


FIGURE B.4: RESTRUCTURED CAUSAL NETWORK: VARIABLE 16 - DRIVER IMPAIRMENT

BEST COPY AVAILABLE

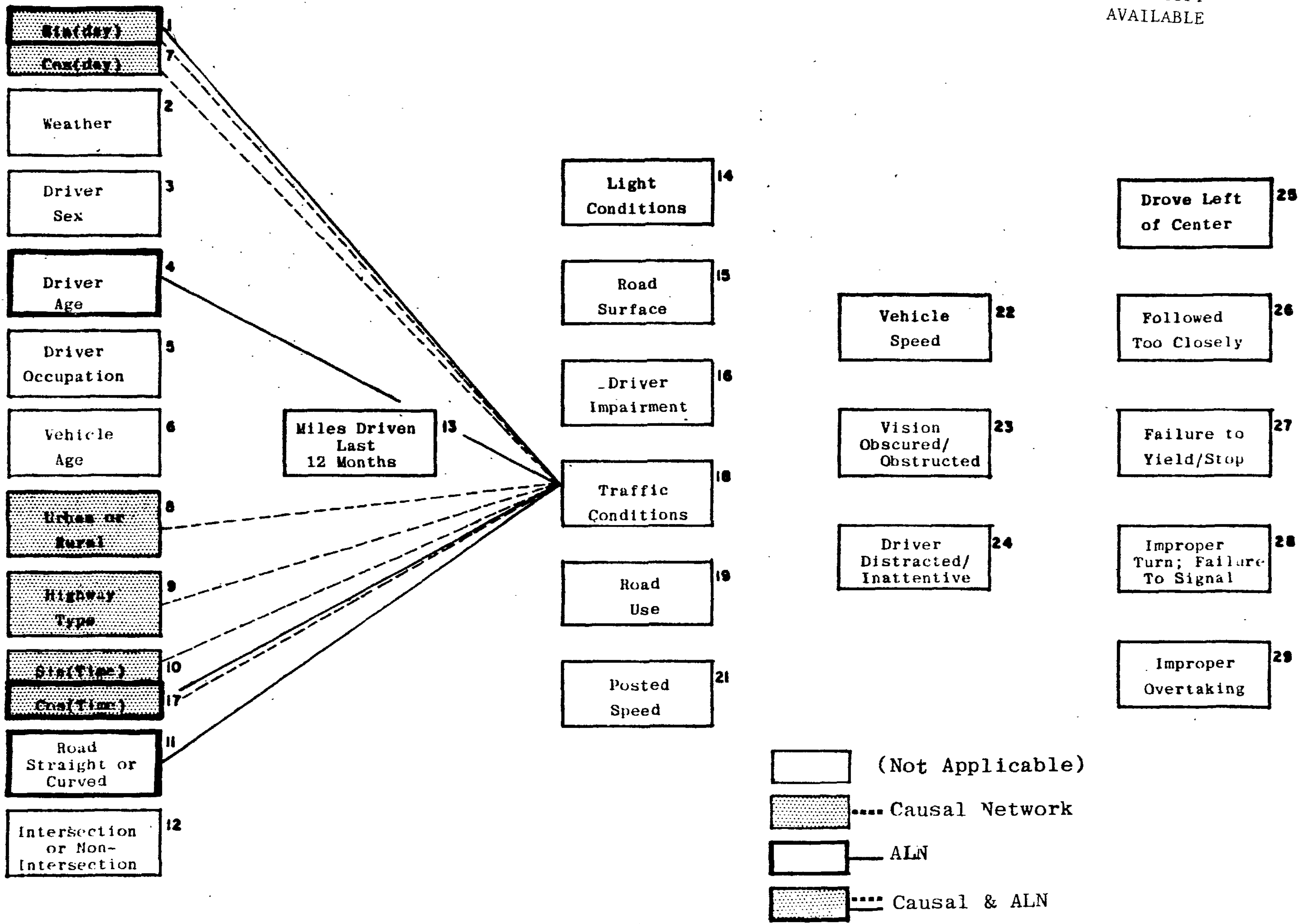


FIGURE B.5: RESTRUCTURED CAUSAL NETWORK: VARIABLE 18 - TRAFFIC CONDITIONS

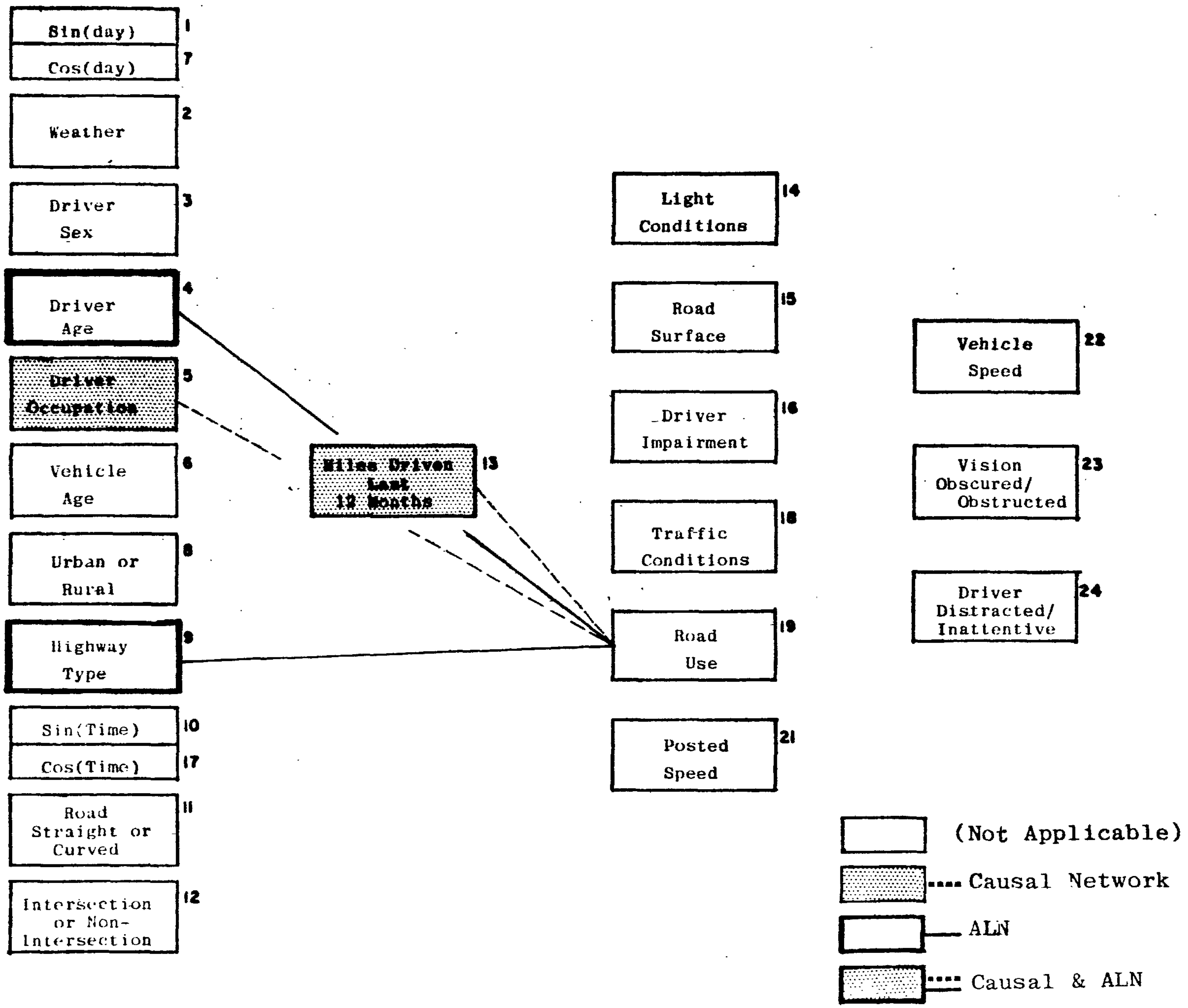


FIGURE B.6: RESTRUCTURED CAUSAL NETWORK: VARIABLE 19 - ROAD USE

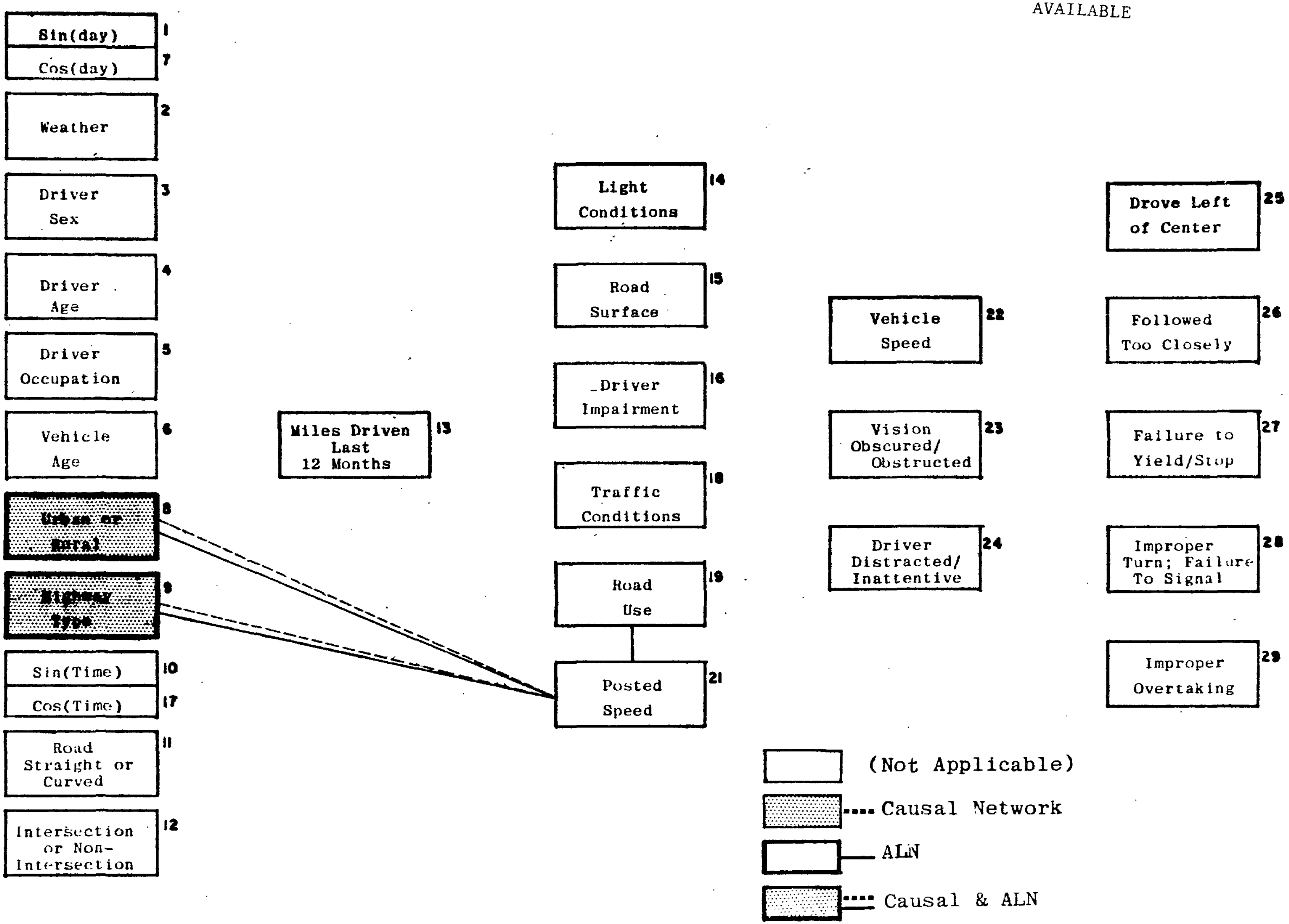


FIGURE B.7: RESTRUCTURED CAUSAL NETWORK: VARIABLE 21 - POSTED SPEED

BEST COPY
AVAILABLE

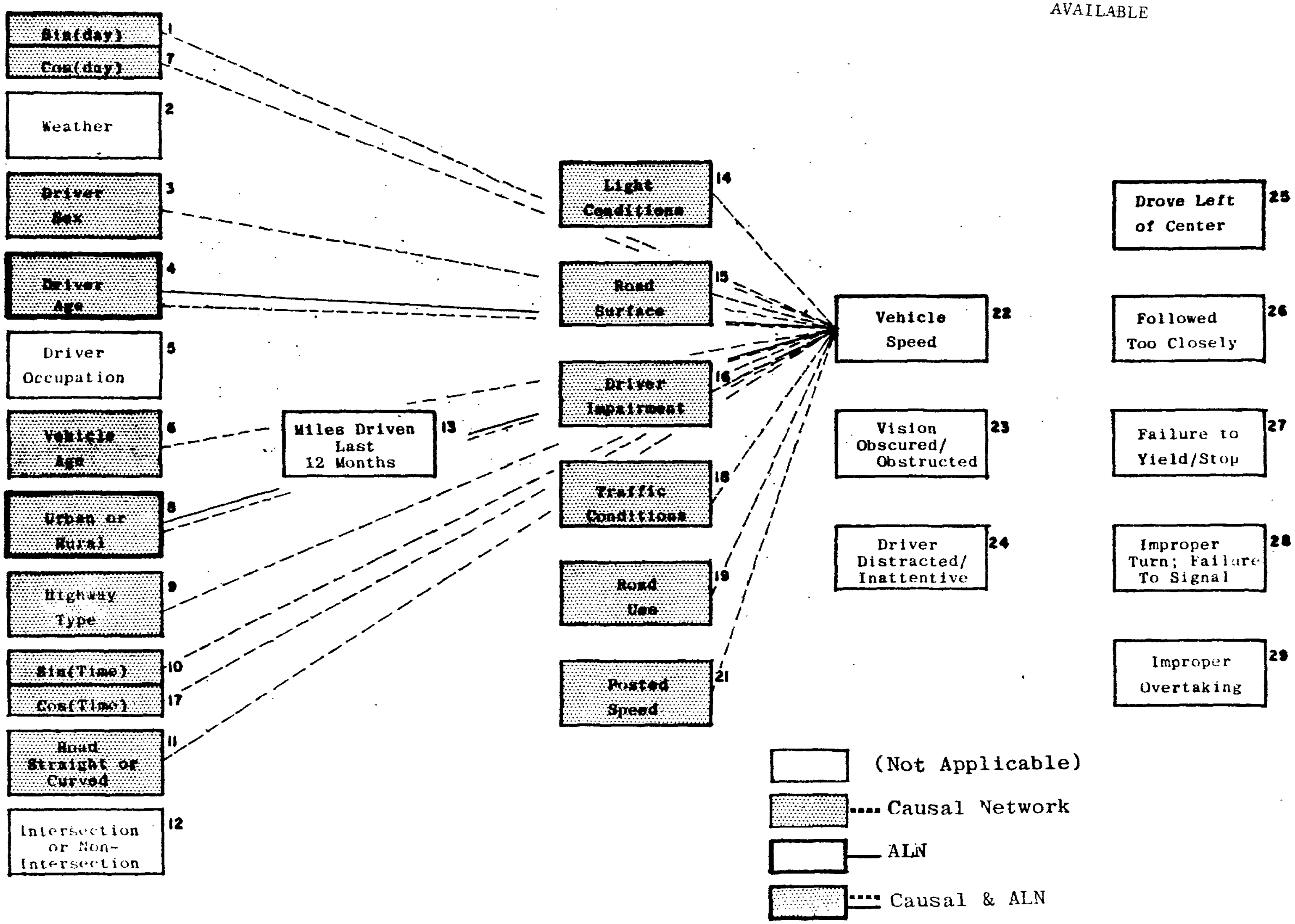


FIGURE B.8: RESTRUCTURED CAUSAL NETWORK: VARIABLE 22 - VEHICLE SPEED

BEST COPY AVAILABLE

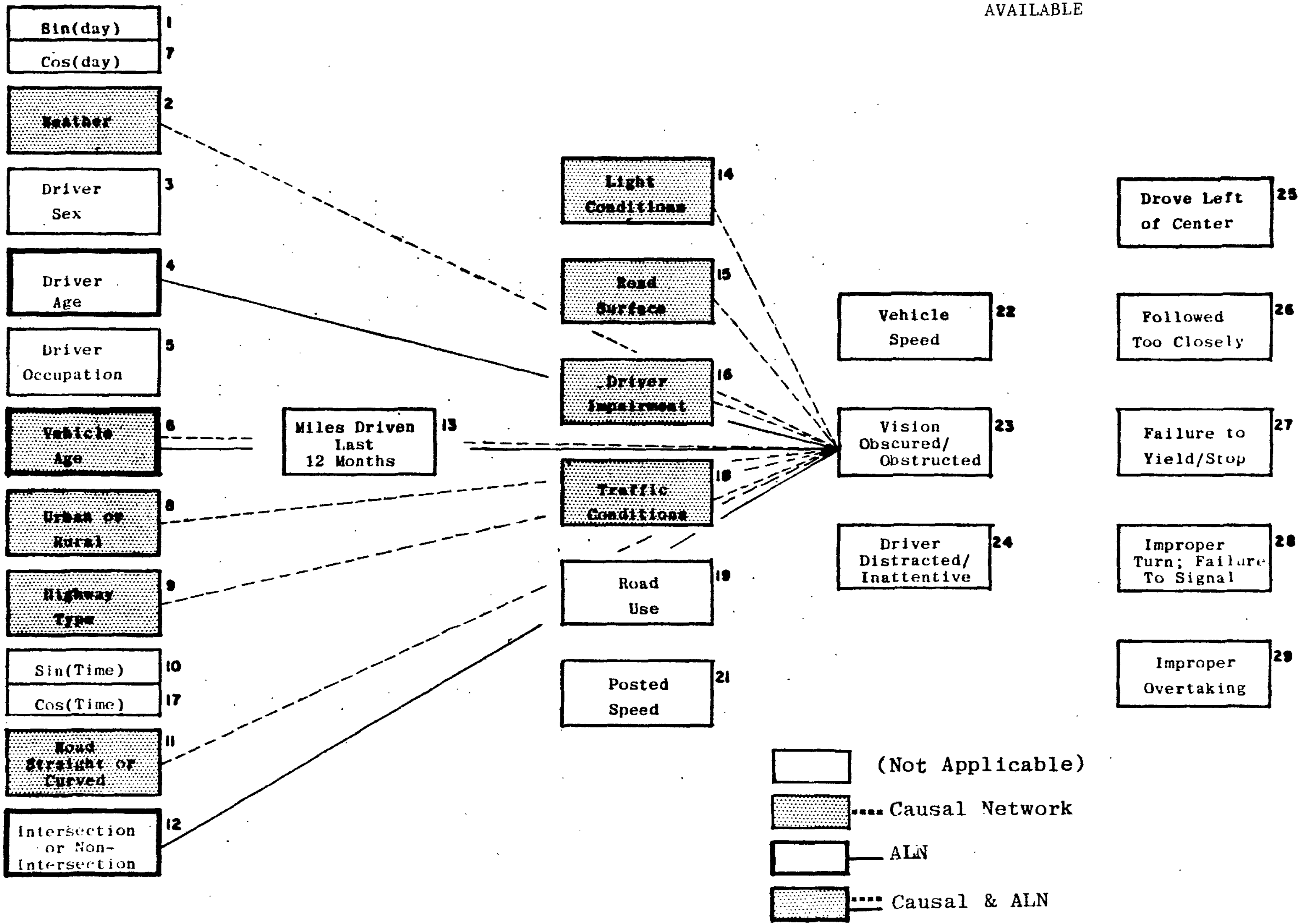


FIGURE B.9: RESTRUCTURED CAUSAL NETWORK: VARIABLE 23 - VISION OBSCURED/OBSTRUCTED

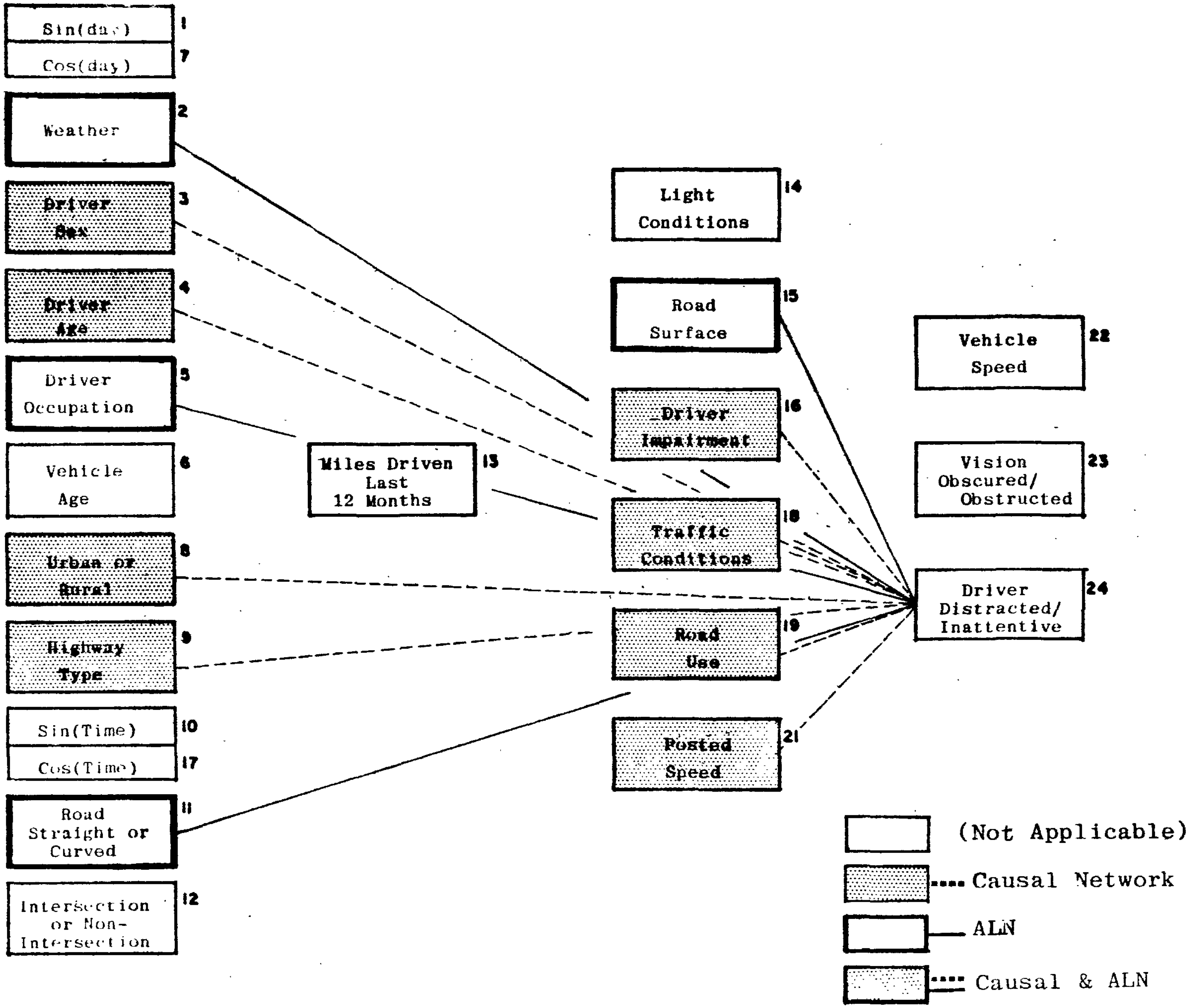


FIGURE B.10: RESTRUCTURED CAUSAL NETWORK: VARIABLE 24 - DRIVER DISTRACTED/INATTENTIVE

BEST COPY AVAILABLE

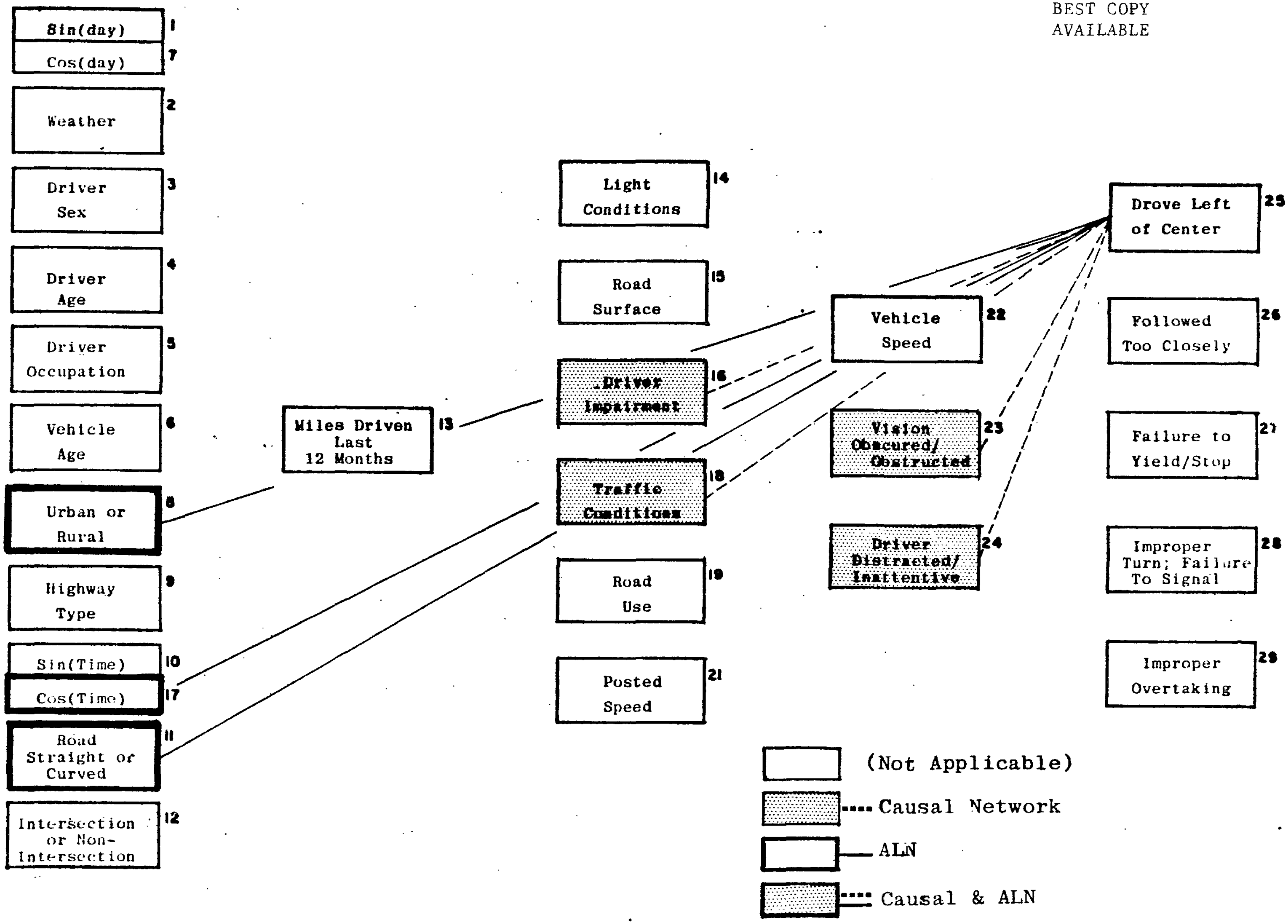


FIGURE B.11: RESTRUCTURED CAUSAL NETWORK: VARIABLE 25 - DROVE LEFT OF CENTER

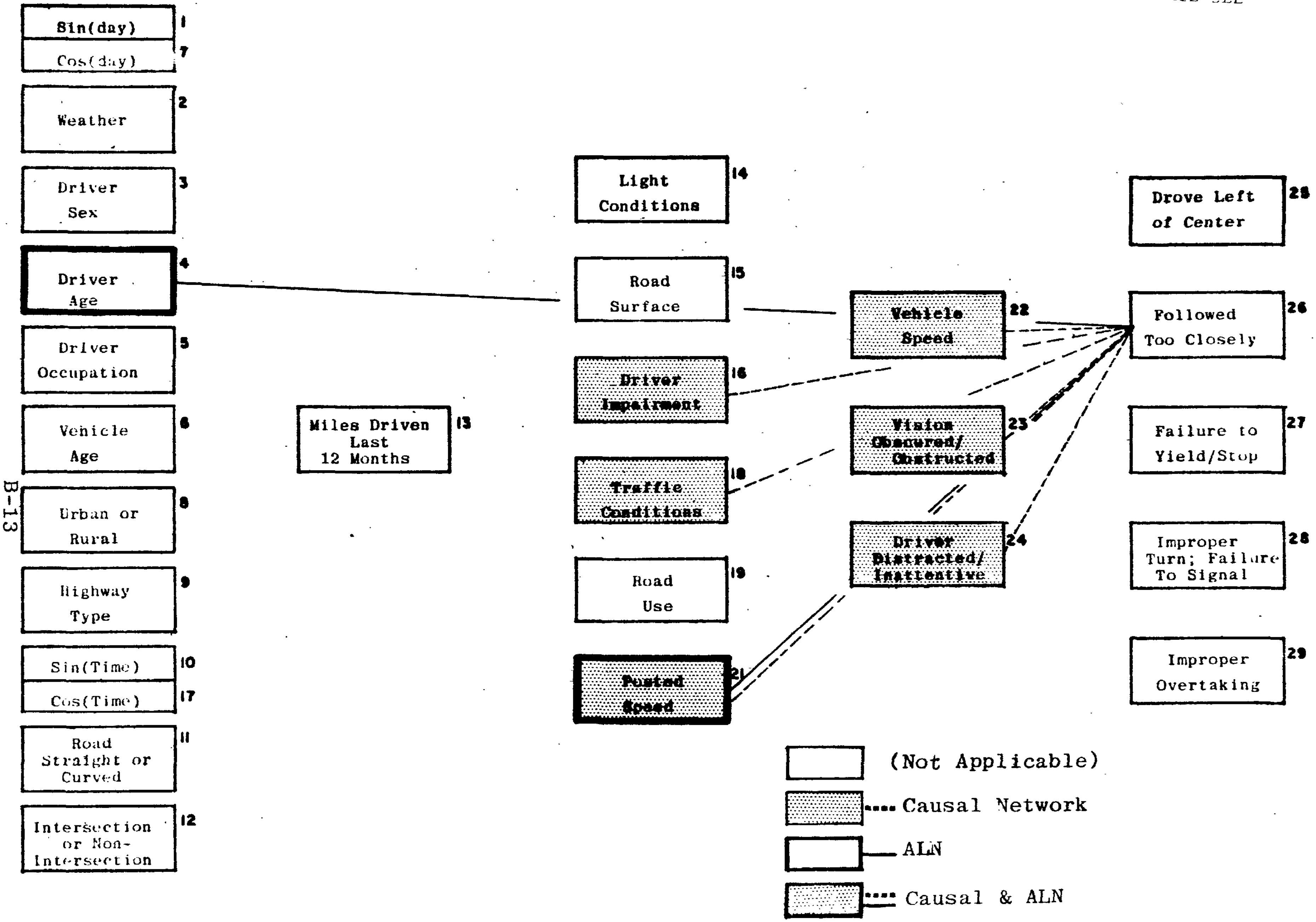


FIGURE B.12: RESTRUCTURED CAUSAL NETWORK: VARIABLE 26 - FOLLOWED TOO CLOSELY

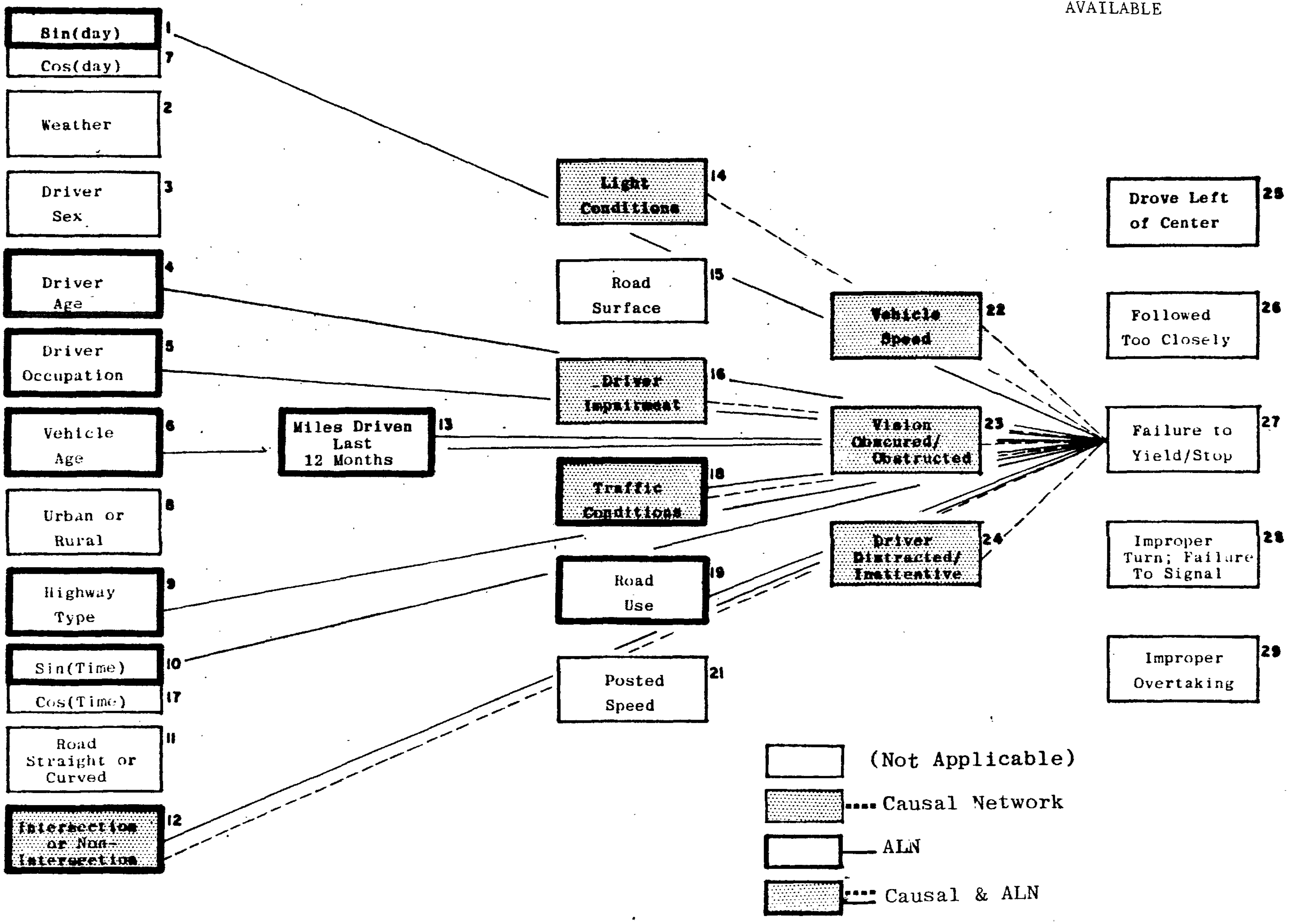


FIGURE B.13: RESTRUCTURED CAUSAL NETWORK: VARIABLE 27 - FAILURE TO YIELD/STOP

BEST COPY AVAILABLE

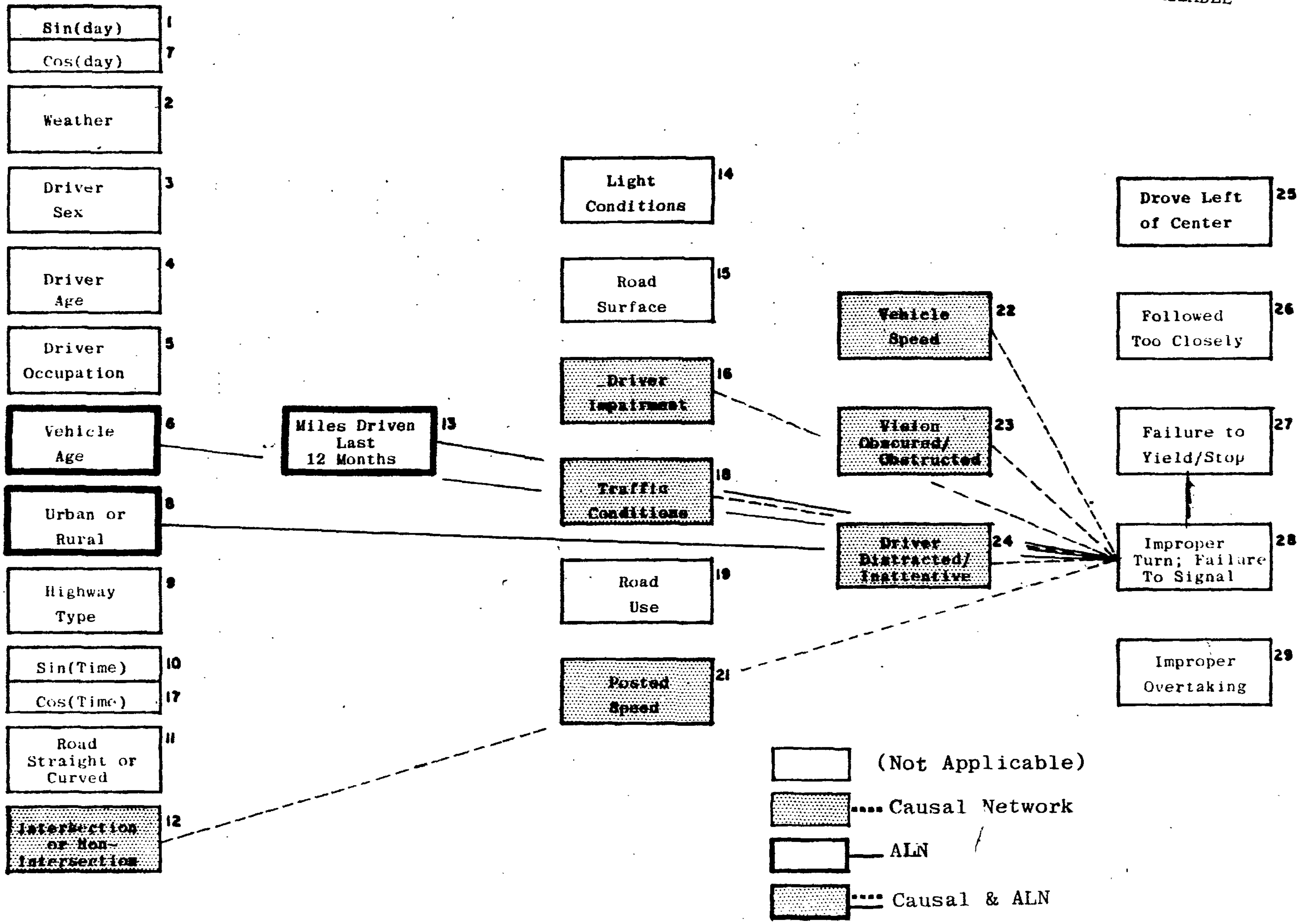


FIGURE B.14: RESTRUCTURED CAUSAL NETWORK: VARIABLE 28 - IMPROPER TURN; FAILURE TO SIGNAL

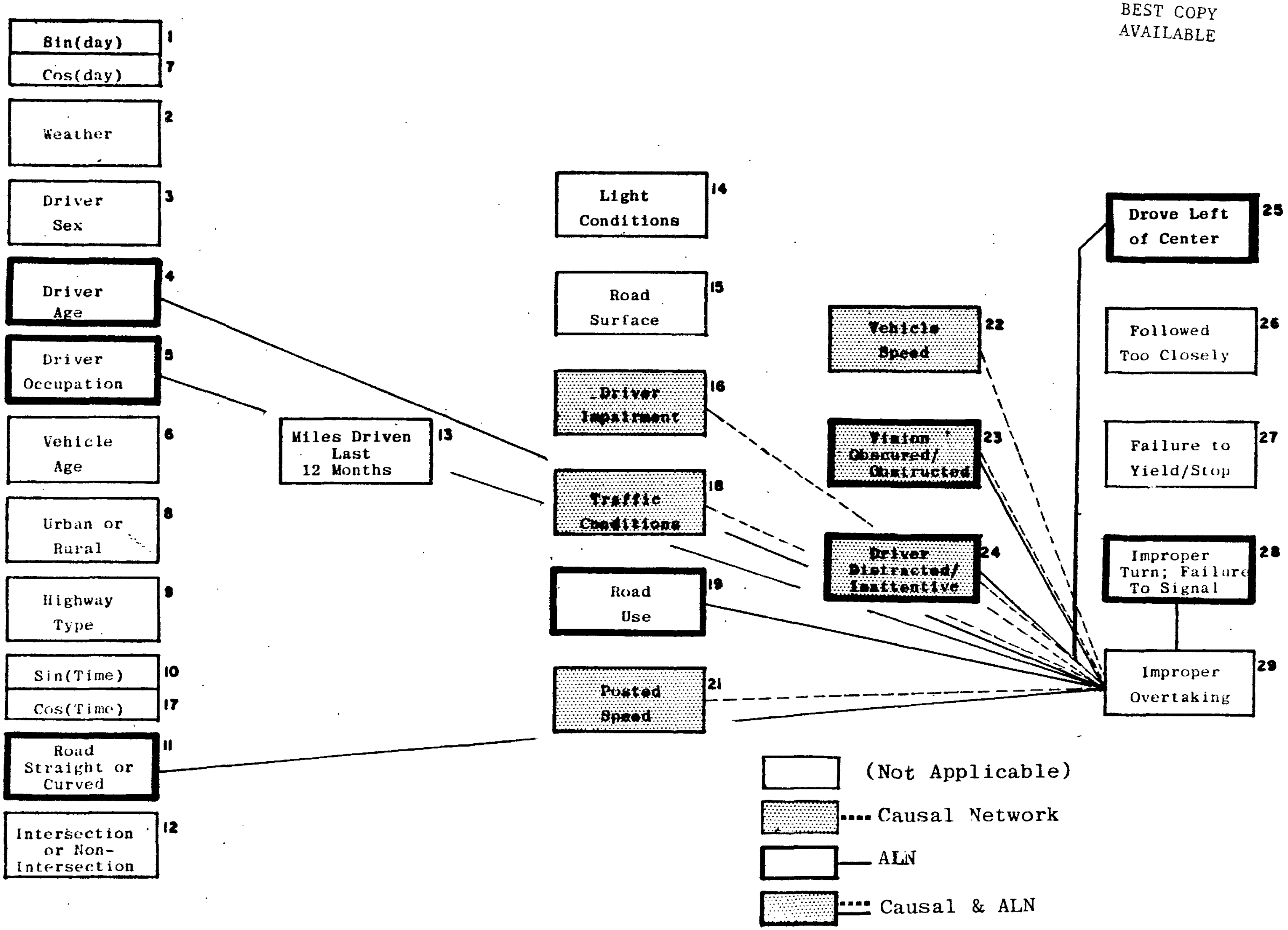


FIGURE B.15: RESTRUCTURED CAUSAL NETWORK: VARIABLE 29 - IMPROPER OVERTAKING

APPENDIX C

SYNTHESIS OF A PATTERN CLASSIFIER VIA MAHALANOBIS DISTANCE FUNCTION

One distinct characteristic of the highway accident data base was that data were primarily available from only one class, i.e., accident-involved driver population. There were no data available for the non-accident driver population. This section describes details for synthesizing a pattern classifier to discriminate between non-accident and accident-involved populations in such a situation.

The synthesis of a pattern classifier system to discriminate between the accident-involved driver population and non-accident driver population with only the accident-involved data available could be accomplished in the following way. Let X denote the input variable vector of N-dimensions which will be used to predict whether the observation is from an accident-involved population or not. X is a column vector. The transpose of X is the row vector X^t:

$$X^t = [x_1, x_2, \ldots, x_N]$$

where x_1 can be the driver sex, x_2 the driver occupation, x_3 the driver age, x_4 driver impairment, etc. Note that upper case letters denote a vector, whereas the components of the vector are given in lower case letters. Let $\{X_1, X_2, \ldots, X_{NT}\}$ be the set of observations, or training samples, that is available for synthesizing the pattern classifier. These training samples are all from the accident-involved driver population only.

Let U be the sample mean vector and C the covariance matrix for the accident-involved driver population:

$$U = \frac{1}{N_T} \sum_{i=1}^{N_T} X_i$$

$$C = \frac{1}{N_T-1} \sum_{i=1}^{N_T} (X_i-U)(X_i-U)^t$$

For any unknown input sample X, the Mahalanobis distance can be computed:

$$d_M(Y) = (Y-U)^t \, C^{-1}(Y-U)$$

We can construct the following pattern classifier:

Mahalanobis Distance Pattern Classifier

A sample X is said to be from the accident-involved driver population if:

$$d_M(X) \leq t_o$$

where t_o is a non-negative threshold.

The boundary of this region, $d_M(Y) = t_o$, defines a hyper-ellipsoid in the N-dimensional vector space. The value for the threshold t_o can be determined as follows. The sample mean vector U and the covariance matrix C are estimated using the training set $T = \{X_1, X_2, \ldots, X_{NT}\}$. By carefully selecting the samples for the training set T, the "optimum" threshold value can be obtained that will optimize the performance of the classification system. Another application of this Mahalanobis distance pattern classifier was described in Reference [2] by Gonzalez.

A.2 ESTIMATION OF CONDITIONAL MEMBERSHIP PROBABILITIES VIA ALN

The Adaptive Learning Ne:work (ALN) modeling technique is a non-probablistic approach. It is sometimes desirable to compute the conditional membership probability $P(k|X)$ for class k -- given that the input (variable) vector is X -- from the output of the ALN model. The following describes three methods for estimation of the conditional membership probability.

Let $\hat{y}$ be the output from the ALN model (the dependent variable). The input vector X is a column vector and $X^t = [x_1, x_2, \ldots, x_N]$ is a row vector. The input variables $x_1, x_2, \ldots, x_N$ are the independent variables. The objective is to compute the conditional membership probability $P(\hat{y}\varepsilon k|X)$ where "$\hat{y} \varepsilon k$" means that the output dependent variable is from class k when the input variable vector is X. In this project, three methods of estimating the conditional membership probabilities were defined; these are called the Distance Method, the Normal Distribution Method, and the Histogram Method.

Distance Method

The Distance Method estimates the conditional membership probabilities using the distance function between the predicted value of the dependent variable, $\hat{y}$, and the true values y_1 for class 1 and y_2 for class 2. The conditional membership probabilities are given by:

$$P\{\hat{y}\varepsilon 1|X\} = \frac{d_2}{d_1+d_2}$$

$$P\{\hat{v}\varepsilon 2|X\} = \frac{d_1}{d_1+d_2}$$

where $d_i = |\hat{y} - y_i|$ $(i = 2)$ is the Euclidean distance between $\hat{y}$ and y_i.

Normal Distribution Method

The Normal Distribution Method assumes that the conditional membership probability density functions of the dependent variable $\hat{y}$, given that it is class i, are normally distributed, $N(\mu_i, \sigma_i^2)$; i = 1, 2; with mean μ_i and variance σ_i^2. The conditional membership probabilities are given by:

$$P\{\hat{y}\varepsilon i|X\} = (2\pi)^{-\frac{1}{2}}\sigma_i^{-1} \exp \{-\frac{1}{2\sigma_i^2}(\hat{y}-\mu_i)^2\}$$

where, i = 1, 2.

To compute the above conditional membership probabilities, one needs to estimate the mean μ_i and variance σ_i^2 for each class i. There are two ways to accomplish this purpose:

(1) Assume that both the mean and variance are unknown and use the sample data to estimate their values. Let $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ be the estimators respectively, then:

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{y}_{ij}$$

$$\hat{\sigma}_i^2 = \frac{1}{N_i-1} \sum_{j=1}^{N_i} (\hat{y}_{ij} - \hat{\mu}_i)^2$$

where $\{\hat{y}_{ij}, j = 1, 2, \ldots, N_i\}$ are the training samples for class i; i = 1, 2; and N_i is the number of training samples for class i.

(2) Assume that the variance is unknown and the mean is $\hat{\mu}_i = y_i$. The unknown variance σ_i^2 is estimated by its sample variance:

$$\sigma_i^2 = \frac{1}{N_i-1} \sum_{j=1}^{N_i} (\hat{y}_{ij} - y_i)^2$$

where $\hat{y}_{ij}$ is defined as before.

Histogram Method

The Histogram Method estimates the conditional probabilities by their relative frequencies of occurrence based on the training samples. The histogram of the dependent variable $\hat{y}$ is computed as follows. First, the range of the dependent variable is partitioned into a total of N_T intervals. Let $\hat{y}_{max}$ and $\hat{y}_{min}$ be the maximum and minimum values for the dependent variable. Next, define interval I_j by:

$$I_j = \{\hat{y}: (j-1)\ell + \hat{y}_{min} \leq \hat{y} < j\ell + \hat{y}_{min}\}; \; j = 1, 2, \ldots, N_T$$

where,

$$\ell = (\hat{y}_{max} - \hat{y}_{min})/N_T.$$

Then the relative frequency of occurrence, p_{ij}, is defined to be the number of training samples from class i which fall into the interval I_j:

$$\bar{p}_{ij} = \#\{y \varepsilon I_j | X \varepsilon i\}/N_i;$$

$$i = 1, 2; \; j = 1, 2, \ldots, N_T.$$

Hence, the conditional probabilities are given by

$$P\{\hat{y}\varepsilon i|X\} = P\{\hat{y}\varepsilon I_j|X\} = p_{ij};$$
$$i = 1, 2.$$

An Example

Any one of the above methods can be used to compute the conditional membership probabilities $P\{\hat{y}\varepsilon i|X\}$; $i = 1, 2$. The following is a specific example in which these conditional probabilities are computed.

Let the dependent variable $\hat{y}$ be the highway accident variable x_{25} ("drove left of center"). The independent variables are x_i, $i = 1, 2, 3, \ldots, x_{24}$, where:

x_1 = Day, Date, Time

x_2 = Weather

x_3 = Driver Sex

.

.

.

x_{24} = Driver Distracted/Inattentive

The dependent variable y is equal to 1 when the cause ("drove left of center) is not cited as the reason for the accident, and y is equal to 2 when "drove left of center" is cited as the reason for the accident. The independent variables $x_1, x_2, \ldots, x_{24}$ are used by the ALN model to estimate the dependent variable $\hat{y}$. We would like to compute conditional membership probabilities

$$P\{\hat{y}\varepsilon i|X\}$$

where class 1 means "drove left of center" is not cited as the reason for the accident and class 2 means it is cited as the reason for the accident.

Using the Distance Method to compute these conditional probabilities:

$$P\{\hat{y}\varepsilon 1|X\} = \frac{|\hat{y}-y_2|}{|\hat{y}-y_1|+|\hat{y}-y_2|}$$

$$P\{\hat{y}\varepsilon 2|X\} = \frac{|\hat{y}-y_1|}{|\hat{y}-y_1|+|\hat{y}-y_2|}$$

where $\hat{y}$ is the output of ALN model when the input independent variables are x_1, x_2, ..., x_{24}. For example, if $\hat{y} = y_1$, then

$$P\{\hat{y}\varepsilon 1|X\} = \frac{|y_1-y_2|}{0+|y_1-y_2|} = 1,$$

$$P\{\hat{y}\varepsilon 2|X\} = 0.$$

Similarly, if $\hat{y} = \frac{1}{2}(y_1 + y_2)$, we have

$$P\{\hat{y}\varepsilon 1|X\} = P\{\hat{y}\varepsilon 2|X\} = \frac{1}{2}.$$

Thus, we have seen that one can compute the conditional probabilities using the ALN. The following is another hypothetical example.

Prediction of Accident Probability

Let the dependent variable $\hat{y}$ be the highway accident indicator such that $\hat{y}$ is equal to y_1 when no accident has occurred and equal to y_2 when it is an accident. Let the independent variables be x_1, x_2, ..., x_{29}, which include all exogenous variables, risk factor variables, etc. given in the Causal Network. The ALN modeling will optimally select the subset of independent variables which is best in predicting or estimating the dependent variable y, highway accident.

Let X be this subset of independent variables, where $X = \{x_1, x_2, \ldots, x_k\}$. Any one of the above three methods can be used to compute the accident probability.

The accident probability is given by:

$$P\{\hat{y}\varepsilon 2 | X\}$$

where class 2 is for accident, i.e. when $\hat{y} = y_2$.

Note that in this hypothetical example it has been assumed that the data base contained samples from both the accident-involved population and the non-accident-involved population. If the samples from the non-accident-involved population were not available, the above method could not be used to compute the conditional membership probability. However, the Mahalanobis distance pattern classifier described earlier in this appendix can be synthesized to discriminate between the accident-involved population and the non-accident-involved population.