A METHODOLOGY TO DEVELOP A DECISION MODEL USING A LARGE CATEGORICAL DATABASE WITH APPLICATION TO IDENTIFYING CRITICAL VARIABLES DURING A TRANSPORT-RELATED HAZARDOUS MATERIALS RELEASE

by

Renee M. Clark

B.S., University of Pittsburgh, 1991

M.S. in I.E., University of Pittsburgh, 1995

M.S. in M.E., Case Western Reserve University, 1999

Submitted to the Graduate Faculty of

the School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

SCHOOL OF ENGINEERING

This dissertation was presented

by

Renee M. Clark

It was defended on

December 14, 2005

and approved by

Richard D. Day, PhD, Assistant Professor, Department of Biostatistics

Jayant Rajgopal, PhD, Associate Professor, Industrial Engineering

Larry J. Shuman, PhD, Professor and Associate Dean, Industrial Engineering

Harvey Wolfe, PhD, Professor, Industrial Engineering

Mary E. Besterfield-Sacre, PhD, Associate Professor, Industrial Engineering
Dissertation Director

ABSTRACT

A METHODOLOGY TO DEVELOP A DECISION MODEL USING A LARGE CATEGORICAL DATABASE WITH APPLICATION TO IDENTIFYING CRITICAL VARIABLES DURING A TRANSPORT-RELATED HAZARDOUS MATERIALS RELEASE

Renee M. Clark, Ph.D.

University of Pittsburgh, 2006

An important problem in the use of large categorical databases is extracting information to make decisions, including identification of critical variables. Due to the complexity of a dataset containing many records, variables, and categories, a methodology for simplification and measurement of associations is needed to build the decision model. To this end, the proposed methodology uses existing methods for categorical exploratory analysis. Specifically, latent class analysis and loglinear modeling, which together constitute a three-step, non-simultaneous approach, were used to simplify the variables and measure their associations, respectively. This methodology has not been used to extract data-driven decision models from large categorical databases.

A case in point is a large categorical database at the DoT for hazardous materials releases during transportation. This dataset is important due to the risk from an unintentional release. However, due to the lack of a data-congruent decision model of a hazmat release, current decision making, including critical variable identification, is limited at the Office of Hazardous Materials within the DoT. This gap in modeling of a release is paralleled by a similar gap in the hazmat transportation literature. The literature has an operations research and quantitative risk assessment focus, in which the models consist of simple risk equations or more complex,

theoretical equations. Thus, based on critical opportunities at the DoT and gaps in the literature, the proposed methodology was demonstrated using the hazmat release database. The methodology can be applied to other categorical databases for extracting decision models, such as those at the National Center for Health Statistics.

A key goal of the decision model, a Bayesian network, was identification of the most influential variables relative to two consequences or measures of risk in a hazmat release, dollar loss and release quantity. The most influential variables for dollar loss were found to be variables related to container failure, specifically the causing object and item-area of failure on the container. Similarly, for release quantity, the container failure variables were also most influential, specifically the contributing action and failure mode. In addition, potential changes in these variables for reducing consequences were identified.

TABLE OF CONTENTS

LIST OF TABLES

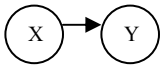
LIST OF FIGURES

NOMENCLATURE

BTS	Bureau of Transportation Statistics. Agency within the DoT that undertook the Safety Data Initiative.
DoT	United States Department of Transportation.
GeNIe	Software for decision modeling, including the development of Bayesian networks. Developed by the Decision Systems Lab at the University of Pittsburgh and used within this research.
HMIRS	Hazardous Materials Incident Reporting System. Large database maintained by the Office of Hazardous Materials (OHM) within the DoT to record releases of hazardous materials during commercial transport. Informally referred to as the Release database.
Lambda/ Loglinear Parameter(λ)	Parameter indicating the strength of the effect or association between two or more variables. Variables X and Y are <u>not</u> directly associated if $\max\|\lambda_{ij}\| < 0.20$, where i and j represent any category combination of X and Y.
Latent Gold	Software for performing latent class analysis. Developed by Statistical Innovations, Inc. and used within this research.

Marginal association	Association between two variables determined by summing over, or ignoring, all other variables in the model.
Partial association	Association between two variables after adjusting or correcting for the effects of other variables. If there is a partial association between two variables, they are not conditionally independent given other variables.
Modified *LISREL* Approach	Latent structure modeling in which latent class analysis and loglinear modeling are performed simultaneously. Implemented in the *LEM* software. Categorical analog to *LISREL* modeling.
NCHS	National Center for Health Statistics. Agency with the Center for Disease Control and Department of Health and Human Services that produces health data for policy and decision making.
OHM	Office of Hazardous Materials. Agency within the DoT that regulates hazardous materials transport.
SDI (Safety Data Initiative)	Effort undertaken by the Bureau of Transportation Statistics to improve safety data collection and empirical analysis.
Three Step Correction Procedure	Procedure involving matrix algebra in which the matrix of the observed and predicted latent variables is corrected, thereby greatly reducing the bias due to the classification error of the latent variables. The result is a corrected matrix of the observed and true latent variables.

Three Step Modeling — Latent structure modeling in which standalone latent class models are built and then used in a structural (loglinear) analysis. Similar to the Modified *LISREL* approach, but the latent class analysis and loglinear modeling are done separately and in succession.



Bayesian network containing two random variables *X* and *Y*, which are represented by circles or ovals. Variable *X* has a direct effect on *Y*, as represented by the arc. The absence of an arc represents independence between two variables.

L^2 — Likelihood ratio chi square statistic. Primary test statistic used in loglinear modeling.

$P(A|B,C)$ — Conditional probability of variable *A* given its parent variables *B* and *C*.

[X] [Y] — Loglinear model notation indicating the lack of interaction, or association, between variables *X* and *Y*. This is also known as the model of mutual independence for *X* and *Y*.

[X Y] — Loglinear model notation indicating an interaction, or direct association, between variables *X* and *Y*.

A⊗B — Kronecker product of matrices *A* and *B*. The super matrix formed from all possible products of the elements of *A* and *B*. Used in the three step correction procedure.

ACKNOWLEDGEMENTS

This dissertation is the result of a desire to understand what research is and was made possible through the guidance and optimism of my advisor, Dr. Mary Besterfield-Sacre. She helped me to formulate this dissertation topic, which was not based on nor an outgrowth of any research being conducted at the University of Pittsburgh or elsewhere.

Researchers at other institutions as well as the University of Pittsburgh were critical to my accomplishment of this research. Dr. C. Mitchell Dayton of the University of Maryland and Dr. Scott Eliason of the University of Minnesota were instrumental to my understanding and application of latent class analysis. Dr. John Kennedy from Ohio State University and his book on loglinear modeling were so important to my understanding of this topic. Tilburg University researchers Dr. Jeroen Vermunt, Dr. Marcel Croon, and Dr. Jacques Hagenaars were very helpful to the Three Step modeling of the problem. Special thanks to Dr. Hagenaars for answering many email communications and assisting me in analyzing large, sparse contingency tables. Finally, thanks to Dr. Marek Druzdzel of the University of Pittsburgh's Decision Systems Lab for spending time with me to ensure the decision modeling was accurate and thorough.

This research was also aided by input and guidance from Doug Reeves at the Office of Hazardous Materials within the Department of Transportation. Thank you to Doug for answering numerous emails and phone messages, in which he always provided his opinion without reservation.

Finally, thank you to Brent for his words of encouragement and numerous acts of support. With great patience and understanding, he saw me spend many evenings at my computer.

1.0 INTRODUCTION

Databases play an important role in today's organizations and may be used both tactically and strategically by businesses and organizations. From a tactical standpoint, they are used to support day to day operations and for reactive decision making. However, data may also be used proactively for business growth or informed governmental policies by applying the decision analysis process. This process dictates that the overall structure of the problem be represented using a model, from which inferences are made for insight and explanation, thereby improving decision making. Although models may be expert or data-driven, they have traditionally been based on expert knowledge. This research takes the non-traditional data-driven approach in constructing a decision model.

In taking a non-traditional approach, a model can be extracted from a database using various statistical, data analysis, or machine learning techniques, including those for categorical data. Large categorical databases are common in today's organizations, as the prevalence of categorical data has increased and categorical data is ubiquitous.[1] However, for the most part, knowledge and use of categorical data methods has remained limited to the social, biomedical, and behavioral sciences as well as education and marketing.[2] Historically, categorical analysis methods were stimulated by research in the social and biomedical sciences, where categorical scales are now pervasive for measuring attitudes, opinions, and medical outcomes.[3] However, as an indicator of the penetration of categorical methods into engineering analysis, only one of the ten top industrial engineering departments for 2005 offers a statistics course focused on categorical data, although most offer courses in basic statistics or continuous data analysis.[4]

Yet, the application of statistical techniques can be challenging when using large categorical databases containing many records, variables, or categories. When the number of variables is large, the number of possible associations between all variable pairs considering all other variables is also large. The large size also leads to problems with convergence, testing, and interpretation of models. In general, when working with a large categorical database, there are challenges in creating a compact, data congruent model for decision making, such as an influence diagram.

A case in point is a database maintained by the Office of Hazardous Materials within the Department of Transportation. This agency develops and recommends regulatory policy changes for the commercial transport of hazardous materials.[5] This transport activity poses risks to life, health, property, and the environment due to the possibility of an unintentional release. This database houses data on hazmat release occurrences, including characteristics such as date, time, location, material type, container failure descriptors, and consequences. The database is largely categorical and contains tens of thousands of records. The use of the database by this hazmat agency has been largely reactive and in support of normal operations, such as investigations surrounding exemptions, occurrence spikes, and cost/benefit analysis. This database has not been used to extract a model of a hazardous materials release. The absence of a model limits the information available during regulatory decision making.

The absence of a modeling approach by this DoT agency is paralleled by and perhaps partly the result of the absence of a similar approach in the hazardous materials transportation literature. This literature base has an operations research focus, with a large number of the

articles involving route optimization or path selection problems. The objective functions in these articles make use of existing, oftentimes simple equations for risk, and the articles do not aim to develop new, multivariate risk models.

Another possible challenge to any previous development of a hazmat release model has been the lack of penetration of categorical data methods into the engineering domain, as mentioned previously. Thus, a methodology for extracting a data-congruent decision model from a large categorical database using statistical methods has not been applied in the engineering arena. Categorical data methods are a recent advance relative to their continuous counterparts and continue to be used mostly by social, behavioral, and biomedical sciences. By the mid 1900's, there was widespread adoption of regression and ANOVA techniques. Conversely, analogs for categorical data received little attention by the social and biomedical research community until the 1960's.[6] Loglinear modeling, which is used to assess associations and can be considered a categorical analog to regression, was mainly developed in the 1970's and gained popularity in behavioral and life sciences in the 1980's.[7] Latent class analysis, which is used for variable simplification and can be viewed as a categorical analog to factor analysis, was developed in the 1950's by sociologist Paul Lazarsfeld for binary survey data. It was extended in the 1970's to include multi-category data and has become a standard tool in social, biomedical, education, and marketing research.[8]

In this research, I provide an approach to the analysis of a large database based on statistical and decision analysis methods from the field of categorical data modeling. Therefore, the contribution made by this research is as follows: using existing categorical data methods, a decision model was extracted from a large categorical database, using the hazmat release

database as the worked example. Since statistical methods were used to build the association structure of the decision model, the relationships among the variables were not based on hunches or assumptions and therefore provide a data-driven basis for decision making.

1.1 PROBLEM STATEMENT

A critical problem related to large categorical databases is effective use of the data for decision modeling. Traditional empirical modeling techniques, such as multiple regression analysis or neural networks, are more conducive to continuous types of data. In addition, a challenge with a very large amount of data is an inability to use significance testing, since the results tend to become significant. Decision analysis in the presence of many categorical variables necessitates extracting a model using a methodology involving exploratory methods for categorical data. Although the application of categorical exploratory methods is present in the literature, a common methodological approach for extracting a decision model, particularly for engineering based problems, is not present. In the case of large amounts of data, a methodology is necessary given the complexity of the data in terms of many records, variables, and categories.

Consider the following related and real situation. The Department of Transportation maintains a large categorical database on hazmat release occurrences in the United States. The database consists of a large number of records and nominal, multi-category variables, whose associations are unknown. Critical information needed from this database includes the identification of influential variables and categories relative to the outcomes of a hazardous materials release incident. This is important because the most influential variables and categories are control points from which operational or policy changes can be made.[9] Another

useful type of information is a characterization of a high-consequence event based on its most likely combination of variables. In this way, the question "What does a high consequence hazmat release most often look like?" can be answered.

This database has not previously been used for proactive decision modeling either by the DoT or researchers in the literature. Existing literature on hazmat releases is unrelated to this, as it focuses on mathematical programming formulations to minimize risk along a transport route. The literature also focuses on risk calculations using analytical equations as part of quantitative risk assessment studies. For the modeling approach taken in this research, the release has already occurred, and the actual consequences and influencing variables are known. Hence, a decision model of the variables and events in a hazmat release can be built to answer questions about the critical variables and their categories. Thus, there are gaps that can be filled through an exploratory analysis of this large database for decision modeling. Given this, a methodology for simplification and measurement of associations among many categorical variables is needed.

In summary, using the DoT database as the worked example, this research endeavors to establish a categorical analysis methodology for developing decision models. In establishing this methodology, critical research based questions about the variables related to the release of hazardous materials can be addressed, which in the past were only speculated within the literature. The application of this methodology to a hazmat transportation problem makes a needed inroad into this policy area as well as other decision problems in the engineering arena. In addition, the questions related to critical variables faced by the DoT are similar to those within other organizations, such as the Center for Disease Control or the Department of Homeland Security, where policies and decisions can be driven by the data collected by these agencies.

1.2 DESCRIPTION OF THE METHODOLOGY

The methodology for development of a decision model consists of three separate analyses of the data. They center on simplification, measurement of associations, and creation of a Bayesian network model. Simplification of the variable domain was accomplished using Pareto analysis, data aggregation, discretization, and latent class analysis. Latent class analysis was used for simplification by combining related variables to form a latent variable.

The determination of the association structure of the decision model began with a temporal layout of the simplified variables. The temporal layout resulted in five distinct stages of a hazardous materials release. These stages are identified as follows: pre-failure initiation, failure initiation, container failure, hazmat release, and realization of consequences, such as dollar loss. After this base structure was created, the associations between the variables were measured using the exploratory technique of loglinear modeling. An exploratory approach was taken in order to create an accurate, data-driven structure. A modeling approach that uses latent class analysis and subsequent loglinear modeling is described in the literature as a three-step, non-simultaneous modeling approach.[10] It is similar to the *LISREL* approach for continuous data that simultaneously combines factor analysis and path analysis.

The variables and data-driven associations determined in the three-step modeling approach were used to build the structure of a Bayesian network, a type of decision model consisting only of random variables and their relationships. Given that a categorical database of uncertain events and variables surrounding a hazardous materials release was the data source for this analysis, a Bayesian network was a natural fit, since its strength exists in modeling complex relations between uncertain variables. In addition, a Bayesian network was a natural fit because the identification of important variables was a key goal of this research. Using the Bayesian

networks developed in this research, the most influential variables relative to two outcomes of a hazardous materials release were identified. In addition, since Bayesian networks allow for computing the impact of some variables on the probabilities of others, desirable policy or operational changes for the explanatory variables were identified. A summary of the methodology developed in this research to analyze a large categorical database is shown below.

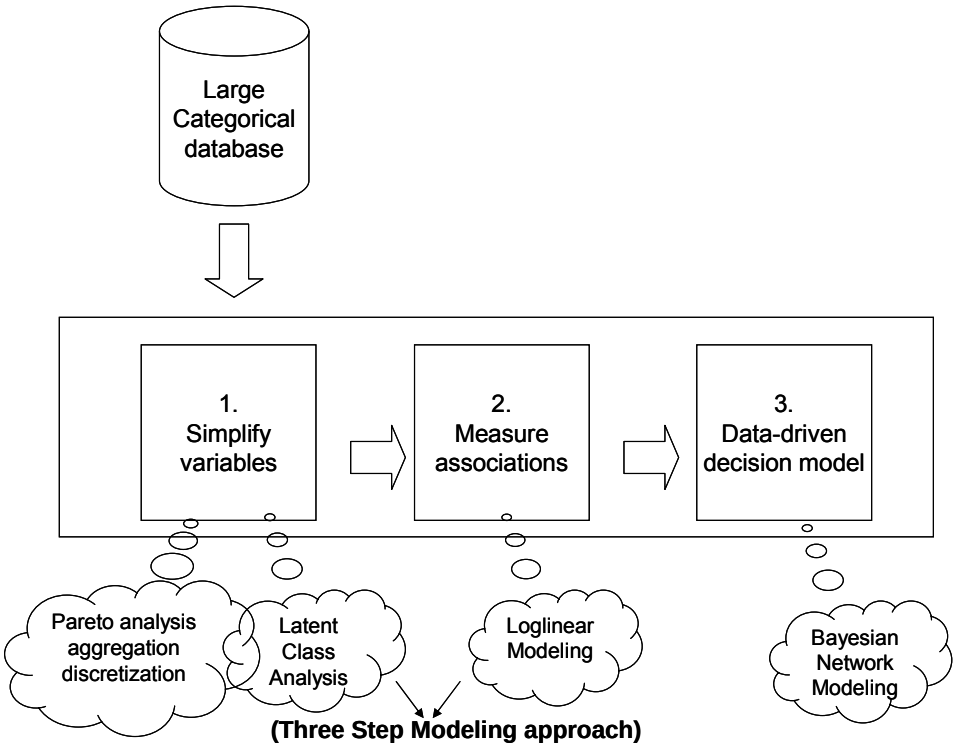


Figure 1: Methodology for Building a Decision Model.

1.3 ADDITIONAL APPLICATIONS OF THE METHODOLOGY

The methodology proposed for building a data-driven Bayesian network using a large categorical database it not limited to the modeling of hazardous materials releases. For example, within the Department of Transportation in general, there is an opportunity and need for growth and

improvement in the analysis of data.[11] This gap was formalized in 1999 by the creation of the Safety Data Initiative (SDI) by the DoT. This program was undertaken by Bureau of Transportation Statistics (BTS) to improve data collection and analysis. Although this program is on hold due to budget cuts resulting from the events of September 11, 2001, its sentiments regarding the need to improve safety data analysis remain the same.[12] One of the goals of this initiative was critical variable or leading indicator identification based on demonstrated correlations with consequences, and this remains a goal within the DoT.[13, 14] The SDI maintained that the DoT's data analysis proficiency was not at the level necessary for good program effectiveness.[15, 16] Within the DoT currently, efficient data collection has achieved proficiency. The next step for sound decision making and enhancement of program effectiveness is improvement of analysis abilities and techniques.[17]

At the state level, transportation authorities could utilize decision support tools when establishing routing designations for hazardous materials. In specifying a routing designation, a state must determine the extent to which certain factors specified by the DoT are incorporated. The DoT maintains that the weighting of these factors is mostly judgmental and should reflect their "expected influence" and the community's consensus.[18] Based on the limited decision support provided for this, the states could benefit from decision models that assist in determining how the factors should be incorporated in routing designations.

Considering areas outside the transportation domain for application of this methodology, the National Center for Health Statistics (NCHS), which is part of the Center for Disease Control (CDC), collects categorical data conducive to the extraction of a decision model representing an overall system or problem network. However, system or network models have not been commonly or formally used by NCHS statisticians, although there is interest in them.[19, 20]

Another potential area for application of this methodology is threat characterization, a required capability within the Department of Homeland Security.[21] These two federal agencies are examples of organizations that could apply this methodology to develop decision models based on categorical data.

2.0 LITERATURE SEARCH: HAZMAT TRANSPORTATION

Risk is an *integral* part of the hazardous materials transportation literature. The majority of articles are operations research studies for minimizing risk on a transport route. The risk equations in the O.R. studies tend to be relatively simple and are often variations on the release probability or the product of release probability and consequences. Other articles focus on calculating risk as part of quantitative risk assessment (QRA) studies of hazmat transport. These articles are typically written by environmental, civil, and chemical engineers who incorporate demographic, meteorological, and chemical databases in calculating risk. The analytical equations in the QRA studies are often mathematically complex and theoretical in nature. These O.R. and QRA studies are focused on releases that occur on the road or along railways. There is not a focus on transport-support activities, such as loading or unloading of containers. Although there are differences in the accident scenarios surrounding these two activities, many of the variables and associations and hence the general Bayesian network structure are the same.

Thus, the great majority of existing studies attempt to minimize or calculate the risk of potential future occurrences. In general, the hazmat literature has not modeled release incidents that have already occurred to determine the influence of the relevant variables. One notable exception is a study by Burns and Clemen in which various sociological, behavioral, and perceptual variables affect the impact of a hazmat release, as depicted using an influence diagram. This study is a continuous-data analog to the present work, and it used a covariance structure, or *LISREL*, model to construct the influence diagram.[22] The present research

contributes to the literature by using categorical data methods to build the decision model, suggested as future work in the Burns article. In addition, the variables considered in the present work are objective variables that describe the accident scenario surrounding a hazmat release to a larger extent.

The decision model by Burns and Clemen is unique within the hazmat transport literature by virtue of its exploratory, statistical nature. In general, this literature lacks a focus on data-driven analyses of outcomes relative to the influencing variables. The literature does not contain multivariate statistical or other exploratory models of risk, due in part to the goals of the researchers. A notable example is the probability of an accident or release, which has traditionally displayed a gap in exploratory modeling of its critical variables. Accident and release probabilities have been estimated for a given road and area type using averaged values, which have limited sensitivity in specific situations.[23] However, some recent empirical work involving fuzzy logic incorporated multiple parameters into a determination of the accident frequency.[24] Additional exploratory work on accident probabilities is still needed.

The primary goal within the hazmat literature has been the use of various risk equations in route optimization and quantitative risk assessment studies. Many of the equations used in the route optimization studies are straightforward in terms of their formulation and are reused across articles. An example of a straightforward risk formulation is the release probability or product of release probability and consequence level. In fact, the straightforward formulations typically contain one or more of the following high-level variables: 1) accident or release probability, 2) consequence level, 3) population count, and 4) exposure amount, such as amount of hazmat transported. Several authors whose risk equations are limited to these high level variables characterize their risk models as "simple."[25, 26]

More complex formulations for risk, which are often used in the QRA studies, include the above high-level variables along with variables such as 5) wind probability or 6) fatality probability, also known as vulnerability. These latter variables are often specified in terms of sub-variables, or input parameters. However, the numerical relationships of the sub-variables to the higher level variables or outcomes are not provided to the reader and are therefore not a discussion focus. For example, in one equation, the release probability calls for the use of vehicle type and material type as sub-variables. However, the exact numerical relationship of vehicle type or material type to release probability is not discussed or provided in the article.[27] In another equation, the following are identified as sub-variables of vulnerability: wind direction, meteorological condition, and final outcome. However, the numerical strength and empirical relevancy of these sub-variables to vulnerability is not demonstrated or a focus of discussion.[28] In general, the determination of high-level variables based on their sub-variables is not described in the literature. This indicates a gap in terms of identifying critical variables.

There are a variety of risk equations used in the risk optimization and QRA studies based on differences in both structure and variables. Thus, there is a lack of agreement on how hazmat transport risk should be represented, as noted in the literature.[29] Based on an analysis of the hazmat transport literature, seven categories for risk were identified, as shown in Table 1.

Table 1: Hazmat Transport Risk Categories.

1	Accident or Release Probability • Probability of a vehicular accident of a hazmat truck • Probability of a vehicular accident that leads to release • Probability of a release
2	Consequence Probability • Individual Risk • Societal Risk
3	Numerical Indices

Table 1 (continued).

4	Consequences
5	Exposure and Product of Exposures
6	Expected Value
7	Variations on Expected Value

In contrast to the various equations, risk is sometimes represented using only statements or definitions versus analytical equations containing variables. In addition, the definitions are often accompanied by assertions of the important variables. The prevalence of these qualitative representations is a further indication of the lack of modeling focus in the hazmat literature. The discussion in the following sections, which is organized based on the different risk representations in Table 1, will elaborate on the issues raised in the previous paragraphs concerning the lack of statistical or exploratory modeling in the hazmat transport literature.

2.1 ACCIDENT/ RELEASE PROBABILITY

Accident, release, and conditional release probabilities have been proposed in the hazmat transport literature as measures for risk. Harwood et al. define risk as the number of releases or vehicular accidents divided by an exposure measure, such as truck miles. Their formula for risk is as follows:

$$Risk = \frac{Events}{Exposure}, \qquad \textbf{Equation 1}$$

where an event is an accident or release. Their accident rates are calculated using truck data from three states, which are combined to produce a weighted average.[30, 31] The Harwood et al.

rates are those most often used to estimate probabilities in risk studies. However, there is concern that application of these rates may lead to inaccuracies in the calculation of risk. For, when using averages, parameters that apply in specific situations cannot be set or altered. For example, in a study performed by Argonne National Lab for the DoT in 2000, the accident and release rates used are stated as a limitation of the study. The study claims that these national averaged rates do not account for local or specific factors that may affect risk.[32] Likewise, Hobeika and Kim suggest that specificity is an important characteristic of accident rates. They feel that state-derived rates should be used instead of national default rates since "each state has unique hazmat transport characteristics."[33] Doug Reeves of the OHM also believes that the use of a general, average rate to calculate risk in specific circumstances, such as for a given highway route, may be inaccurate. However, a challenge in the use of Equation 1, especially for specific scenarios, is the availability of associated data for use in the denominator.[34]

There is a separate group of articles that cover vehicle accidents that do not necessarily involve hazmat. These articles on general accidents contain regression models that use highway geometric and traffic variables to model accidents.[35, 36, 37, 38, 39] The pertinent question is why haven't similar regression models been developed in the hazardous materials transportation literature for modeling accident probabilities?

2.2 CONSEQUENCE PROBABILITY

Models for the probability of a consequence include those identified as Individual or Societal Risk in the hazmat literature. Individual Risk is most commonly defined as the probability of death to an individual due to a hazmat release and is represented by either analytical equations or qualitative statements or definitions. The definitions are often further described by the variables

believed to be important. Societal Risk is represented by means of *F/N* curves, where *F* is the cumulative frequency of an accident with *N* or more fatalities. Analytical equations are used to calculate both *F* and *N*.

2.2.1 Individual Risk

The analytical equations for Individual Risk are often detailed or mathematically complex and have been implemented in software by environmental or chemical engineers for quantitative risk assessment along transport routes. The following high level variables are present in an equation for Individual Risk proposed by Leonelli et al.: 1) frequency of release, 2) probability of final outcome given a release, 3) wind PDF, and 4) vulnerability. This equation, which is implemented in software, is given as

$$Individual\ Risk = \Sigma_j f_{rel}(l, v, j) \int_{L_l} Risk_{UNIT}$$

$$Risk_{UNIT} = \Sigma_i p^{out}(i) \Sigma_k \int_0^{2\pi} p_{wind}(j, k, \theta)\ V(i, k, \theta) d\theta$$

Equation 2

where

$f_{rel}(l, v, j) = release\ frequency\ for\ link\ l, vehicle\ typology\ v, season\ j$

$p^{out}(i) = probability\ of\ final\ outcome\ i\ given\ a\ release$

$p_{wind}(j, k, \theta) = wind\ PDF\ for\ meteorology\ condition\ k, season\ j, wind\ direction\ \theta$

$V(i, k, \theta) = vulnerability\ for\ outcome\ i, meteorology\ condition\ k, wind\ direction\ \theta.$

The various sub-variables, or input parameters, are the road link or segment, season, type of outcome, meteorological condition, wind direction, and vehicle typology, which is a combination of vehicle and material type. The meteorological condition is described by the wind velocity and atmospheric stability class. However, the numerical relationship of the sub-variables to the higher level variables and the details of their calculation are not provided in the article and are thus not a focus of discussion.[40]

As a second example, Bubbico et al. present a similar equation for Individual Risk in their quantitative risk assessment.[41] Their equation contains several of the same factors as Equation 2, such as a meteorological factor (wind direction), the probability of a fatality, and the release probability, as shown below.

$$Individual\ Risk = TA\Sigma_i R_i \Sigma_j L_{ij} W_j \Sigma_k P_{ijk}$$

where

$T = trips / year$

$A = accident\ rate\ (/ km)$

$R_i = release\ probability\ for\ release\ size\ i$

$L_{ij} = length\ of\ release\ location\ zone\ j\ for\ release\ size\ i$

$W_j = probability\ wind\ blows\ in\ direction\ of\ concern\ for\ zone\ j$

$P_{ijk} = probability\ of\ fatality\ at\ zone\ j\ for\ release\ size\ i\ given\ outcome\ k$

$i = \{major, minor, negligible\}$

$k = \{jet\ fire, flash\ fire, fireball, UVCE\}.$

Equation 3

However, there are differences in the sub-variables in Equation 2 versus Equation 3. For example, the probability of a fatality in Equation 3 does not have a meteorological or wind-related sub-variable, as in Equation 2. In addition, season and vehicle typology are not used as sub-variables for the release probability in Equation 3, as in Equation 2. Thus, the question of which equation and sub-variables more-accurately describe Individual Risk can be raised.

In addition to these equations, various authors provide qualitative statements or definitions for Individual Risk. Saccomanno and Shortreed define Individual Risk as the annual probability of death at various distances and suggest that hazmat quantity and traffic level are leading indicators of this risk. Roodbol states that Individual Risk is the annual probability that a 24-hour, unprotected resident at a certain distance from the incident will be killed. According to this

author, risk depends on material, quantity, population density, and traffic safety mechanisms such as speed limit, guidance systems, traffic separation, and infrastructure.[42]

2.3 NUMERICAL INDICES

A risk index of a hazardous materials incident was developed by Scanlon and Cantilli, who approach risk from a transportation and safety engineering perspective. Their risk index consists of numerous independent variables, as shown below.

$$Risk\ Level_{HMI} = RL_{MVI}(5.5P_{ex} + 2.5P_{fl} + 4.0P_{cg} + 1.0P_{c} + 1.0P_{p})L_{V}L_{D}$$

where

P_{ex} = *proportion of explosives vehicles*

P_{fl} = *proportion of flammable liquids vehicles*

P_{cg} = *proportion of compressed gas vehicles*

P_{c} = *proportion of corrosives vehicles*

P_{p} = *proportion of poisons vehicles*

L_{V} = *vehicle level (physical condition, age, packaging)*

L_{D} = *driver level (experience, history, training)*

RL_{MVI} = *Risk Level of a motor vehicle incident*.

Equation 4

The accuracy of this equation and its coefficients and the relevancy of the independent variables are unknown. Evidence of empirical validation is not provided. Potential data sources for several of the independent variables, such as driver level and condition of traffic control devices and medians, are not provided. Therefore, the feasibility of applying this equation, especially in a large area, is questionable. The authors do not demonstrate their risk index in a real-life application. Their equation for the risk level of a motor vehicle incident, a factor in the previous

equation, contains several variables for roadway characteristics and is depicted in the following manner:

$$RL_{MVI} = L_{tv}(N_i \text{ or } N_r + N_{hc} + N_{vc} + C_p + C_m + N_{rh} + C_{tc})$$

where

L_{tv} = *traffic volume level*
N_i = *intersections/mile*
N_r = *ramps / mile*
N_{hc} = *horizontal curves/mile*
N_{vc} = *vertical curves/mile*
C_p = *pavement condition*
C_m = *median condition*
N_{rh} = *roadside hazards/mile*
C_{tc} = *traffic control device condition.*

Equation 5

Although the authors developed these indices using a multi-variable approach, they did not provide justification or rationale for the variables or coefficients chosen.(43)

In addition, recent work has been done in the development of a transportation risk index that incorporates hazard rankings for amount transported, nearest habitation distance from a release, material dispersion characteristics, and chemical properties of the material. These inputs determine a risk index intended to be used as a practical guideline versus a model for transporting various chemicals.(44)

2.4 CONSEQUENCES

Risk is also represented in the literature as the undesirable consequences from a release. Consequences include monetary losses, injuries, and fatalities. The DoT considers consequences

as a measure of risk, but this has the drawback of year-to-year variation. For example, based on data in the HMIRS, there were 120 fatalities in 1996, versus an average of about 11 per year from 1993-2001, excluding 1996. The large number of fatalities in 1996 was due to the crash of a commercial airliner, which caught fire due to the hazmat it was transporting in a non-regulatory manner. However, it is questionable as to whether the overall risk was higher in 1996 versus in other years.

Although the literature does not include equations for calculating consequences, there are various qualitative statements concerning consequences. Erkut and Verter define exposure risk during a release as the undesirable consequences, which are stated to be dependent upon vehicle design, material, geography, and meteorology.[45] In a maritime hazmat article, risk is represented using natural resource restoration costs, which are given as dependent upon the type of material.[46]

Some recent work in railway transportation of hazardous materials considers the initiating events leading to a loss of containment, which is a type of consequence. Although this recent work focuses on collisions and derailments during rail transport, it uses an event, or fault, tree to model a release and therefore has similarities to the present research. However, it does not perform decision modeling, such as Bayesian networking, based on the event tree. Its framework is that of a chain of events leading to an ultimate event, or loss of containment. It also differs from the present work in how it calculates various frequencies or probabilities. Specifically, input from experts is used to determine certain frequencies. In addition, consequence probabilities, including the probabilities of death or injury, are calculated using probit equations based on the effects of concentration level and material type. Derailment frequencies are calculated using analytical equations that make use of detailed rail infrastructure

data. The use of equations to calculate probabilities are in contrast to use of frequency data or counts from a database to calculate probabilities, as was done in this research.[47]

2.5 EXPOSURE AND PRODUCT OF EXPOSURES

Exposure is sometimes defined in the hazmat literature as opportunities for incidents to occur. Therefore, exposure exists in the form of number of shipments, amount of material shipped, or distance traveled. Exposure is also defined as the number of people potentially subjected to a release of material. ReVelle et al. propose their "tons-past-people" measure of perceived risk. This is calculated as the product of tons of waste transported on a link and the population within a certain bandwidth of the links, summed over all links, as given by

$$Risk = \sum_{ij} T_{ij} G_{ij}$$

where **Equation 6**

Tij = tons of waste moving on link ij

Gij = population within a certain bandwidth on link ij.

This equation for risk is used within their multi-objective programming problem for transportation policy analysis. Based on the authors' statement, "tons-past-people" is a simple risk measure, and better measures of risk should be developed in future research.[48] Their statement points to the desirability of detailed modeling of risk.

2.6 EXPECTED VALUE

An expected value representation for risk is advocated by the Department of Transportation and is the most common model for risk in the hazmat literature. In general, the expected value model is calculated as the product of a 1) probability and a 2) consequence or exposure. The DoT defines risk as the product of the probability of an accident that results in a release and the population within the impact area.[49]

There are variations on the probabilities used in the expected value representation for risk. One such probability is the probability of an accident. For example, Sivakumar et al. define risk using the following equation:

$$Risk = \Pr(A)\, C$$
where
$\Pr(A)$ = *probability of an accident*
C = *consequence.*

Equation 7

This model is used in their risk minimization routing problem for the transportation of hazardous materials. Although they indicate that weather conditions and time of day affect the accident probability, it is not their goal to determine the actual relationship. They make use of accident probabilities that are generated randomly and not modeled.[50]

Jin et al. also use the probability of an accident in their expected value calculation of risk, which is used within a risk minimization problem. Despite noting that environmental influences such as design speed, pavement wetness, and visibility influence the accident probability, they calculate accident probability as the product of the segment length and a uniform random number between 0.01 and 1. Thus, the influence of specific environmental factors on accident probability is not taken into account.[51]

Glickman and Sontag use the probability of an accident-causing release in their expected value risk calculation. They acknowledge that their risk calculation does not take factors other than road length and type and population density into account due to the unavailability of such data on a nationwide basis.(52) This may provide some insight into the lack of detailed representation or modeling of probabilities and consequences in hazmat routing studies.

Taking a slightly different approach, Patel and Horowitz, who combine industrial and civil engineering perspectives, view risk as the expected concentration level of a released gas. They model the concentration level using the Gaussian plume model, a common dispersion model that incorporates the mean wind speed, gas emission rate, and atmospheric stability. To calculate the expected concentration, or risk, the concentration level is multiplied by the potential for a vehicle crash.(53)

2.7 VARIATIONS ON EXPECTED VALUE

Perceived risk has been modeled using an exponent on the exposure in the expected value representation for risk discussed previously. This exponent is known as the risk preference parameter. The larger the risk preference parameter, the higher is the decision maker's aversion to risk. Such a model has the following form:

$$Risk = \sum_i (POP_i)^q \Pr(accident \rightarrow release)_i$$

where

$POP_i = exposed\ population\ on\ segment\ i$

$q = risk\ preference\ parameter$

$\Pr(accident \rightarrow release)_i = probability\ of\ an\ accident\ that\ leads\ to\ release\ on\ i.$

Equation 8

A value of $q > 1$ represents risk averse behavior, while a value of $q = 1$ indicates risk neutrality.[54, 55] As with the expected value representation for risk, the factors in this equation are limited to high level variables.

2.8 CONCLUSION

In general, the studies in the hazmat transport literature do not have an exploratory modeling focus. Rather, various analytical equations for risk are used in route optimization or quantitative risk assessment research. The lack of focus on exploratory modeling of risk in terms of its important variables presents a gap or opportunity in the hazmat literature. This research contributes to the literature by introducing a data-driven Bayesian network model of a hazardous materials release during unloading operations.

3.0 LITERATURE SEARCH: CATEGORICAL DATA METHODS

In order to develop a methodology to analyze a large categorical database, a literature search on data analysis topics was done. The specific subjects searched include categorical data analysis, simplification of a large categorical database, identification of critical variables, and exploratory construction of a decision model based on categorical data. The following main topics were uncovered:

- Path Analysis
- Structural Equation Modeling (*LISREL*)
- Modified *LISREL* Approach
- Three Step Latent Structure Modeling
- Loglinear Modeling
- Latent Class Analysis
- Decision Tree Entropy Analysis
- Bayesian Networks

These topics will be discussed in the following sections in order to provide background for the methodology established by this dissertation.

3.1 PATH ANALYSIS

Path analysis is a graphical method used to model the relationships among a group of linearly-related continuous or binary variables. The goal is to measure the direct and indirect paths, or effects, between variables. Thus, path analysis is a means to assess the influence of certain

variables on other variables. An example path diagram, which can be analyzed using path analysis, is provided below in Figure 2. Both *W* and *X* have a direct effect on *Y*, which has a direct effect on *Z*. Both *W* and *X* have only an indirect effect on *Z*.

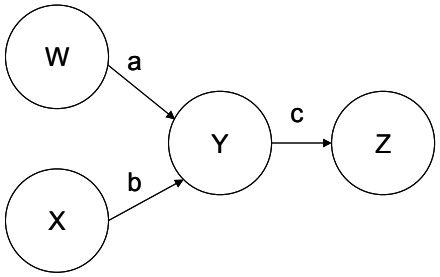


Figure 2: Example Path Diagram.

A *direct effect*, also known as a path coefficient or beta weight, measures the direct influence of the variable at the tail of the arrow on the variable at the head of the arrow, with the other variables held constant. In Figure 2, the path coefficients are *a*, *b*, and *c*. An indirect effect between two variables "passes through," or involves, other variables in the model. Thus, an *indirect effect* is a compound path and is calculated as the product of the path coefficients, or direct effects, along the path. The path coefficients correspond to beta weights in a regression equation, which is one method of solution.[56, 57, 58, 59, 60] A path diagram can be solved as a series of multiple linear regressions, with one equation per dependent variable in the diagram. The second solution method is an algebraic solution based on the decomposition of the overall correlation between two variables into various effects, including direct and indirect.

Path analysis is generally not performed with multi-category nominal variables. First, there is no formal decomposition relationship of the overall association into the various effects, as there is for continuous variables.[61, 62, 63, 64] This decomposition rule forms the basis of the algebraic solution method within path analysis. Second, although various researchers have

proposed the use of dummy variables for accommodating nominal-scaled variables in the regression formulation of path analysis, this is a tedious if not impractical effort in the case of numerous multi-category variables.[65, 66, 67] Dichotomization of the multi-category variables has also been suggested, but this involves subjective judgment about the similarity of the categories.[68]

3.2 STRUCTURAL EQUATION MODELING

Path analysis is one of the major components of structural equation modeling (SEM), also known as *LISREL* or covariance structure modeling. SEM is used with continuous variables to build a structural model consisting of both latent and observed variables. With SEM, path analysis is simultaneously combined with factor analysis, which is used for developing latent variables.[69, 70] *LISREL* is a software product that performs SEM.

3.3 MODIFIED *LISREL* APPROACH

The categorical variant of SEM is known as the Modified *LISREL* approach. A Modified *LISREL* approach simultaneously combines latent class analysis for development of latent variables and loglinear analysis for structural modeling. It is a one-step, simultaneous estimation approach that provides unbiased estimates of the relationships among the observed and latent variables.[71, 72, 73]

The software available for Modified *LISREL* modeling has limitations in terms of the models that can be built, however. The only product available was *LEM*, an academic, non-commercial product.[74] *LEM* does not have some important functionality available in its commercial

successor, *Latent Gold*, such as automated executions of the model using a predetermined number of sets of randomly-generated start values.[75] This functionality is critical when building latent class models of a complex nature. Based on this, *LEM* was not feasible for building the latent class models in this research. In addition, according to *LEM*'s designer and developer, *LEM* has difficulty analyzing large modified *LISREL* models containing multiple observed and latent variables, especially when the latent variables have several indicators.[76]

3.4 THREE STEP MODELS

The *three step approach* to modeling categorical latent and observed variables is similar to the Modified *LISREL* approach in that it involves latent variable development and structural modeling. However, with the three step approach, standalone latent class models are built first and then used in a structural analysis. Thus, the latent variable development and structural modeling are *not* done simultaneously. The latent class models or variables are built using some of the observed variables in the domain, which serve as indicator variables. Then, the latent variables are modeled along with the remaining observed variables in a loglinear analysis. Thus, the latent variables are cross classified with the observed variables. This is done using the latent class scores, which are assigned to the latent variables during the classification stage of the latent class analysis. The latent variables are essentially treated as observed variables in the structural model.[77, 78]

The three-step approach was used in this research, in large part due to the limitations posed by the software available for one-step modeling, in which the latent class analysis and structural modeling are done simultaneously. The only product available for one-step modeling was *LEM*. Dr. Jeroen Vermunt, developer of the product, confirmed that *LEM* would have difficulty

modeling the five latent and six observed variables in this research simultaneously due to the complexity. This recommendation was also based on three to four indicator variables per latent variable and two to nine categories per indicator variable. Dr. Vermunt recommended a stepwise creation of the overall model in this case.[79] This has also been recommended in the literature.[80]

3.4.1 Advantages of a Three Step Approach

There are several additional reasons for using a three step versus a simultaneous approach. First, when a structural model is built in pieces, the possibility for misspecification of the overall model is decreased. This is due to a smaller chance of excluding important associations or masking poor fit in one portion of the model due to good fit in other portions. In addition, a stepwise approach is better suited for cases in which the model building is exploratory, as in this research. In this way, the researcher does not have to specify a priori the complete model with all latent and observed variables. If a correct or best-approximating model is not known beforehand, a one-step or full information method is usually *not* the best approach. The researcher should instead use an approach that divides the global model into different autonomous parts and fits each separately.[81, 82, 83] Disadvantages to the use of a three step approach will be discussed in a future section.

3.5 LOGLINEAR MODELING

Loglinear modeling is a method for detecting associations among multiple categorical variables and is the component of the modified *LISREL* approach that performs structural analysis.[84, 85] Using maximum likelihood estimation, the cell frequencies are estimated based on the specified model. The lambdas (λ), or effect parameters, are then determined as part of the loglinear

modeling. A main effect parameter indicates the effect that an individual variable has on the cell frequencies. An interaction effect parameter indicates the presence of an interaction, or association, between two or more variables.[86]

The expected frequencies are used to assess the goodness of fit of the loglinear model by a comparison to the observed frequencies. Either the Pearson chi square statistic (χ^2) or the Likelihood Ratio chi square statistic (L^2) can be used to assess the fit, although L^2 is the preferred statistic. The Likelihood Ratio statistic has additive properties and can be partitioned for testing conditional independence.[87]

There are two versions of loglinear modeling. In the asymmetric version, also known as a *logit analysis*, a response variable is assumed or chosen, and the effects of the explanatory variables and their associations on the response variable are determined. Specifically, the log of the odds of the expected frequencies of the response variable is modeled in terms of the variables and their associations. In the symmetric version, a response variable is *not* assumed or chosen. Rather, patterns of mutual association among the categorical variables are explored. In a symmetric loglinear model, the log of the expected cell frequency is modeled in terms f the variables and their associations.[88, 89]

3.5.1 Associations in Loglinear Models

The reason for the use of loglinear modeling in this research is to assess the associations among the hazmat variables for development of an accurate network-based model. There are various types of associations among categorical variables that can be determined or measured using loglinear modeling. For example, one can test for either a marginal or partial association between two variables. A *marginal association* between two variables is determined by summing or collapsing over all other variables in the model. The other variables are in essence

ignored, and the association in the two-way table is assessed exclusively. A *partial association* between two variables is an association after removing the effects of other variables. Based on this, it is a conservative test. Partial association is related to the concept of *conditional independence*. If variables *X* and *Y* are conditionally independent given a third variable *Z*, then a partial association does *not* exist between *X* and *Y* given *Z*. This is depicted in the figure below.

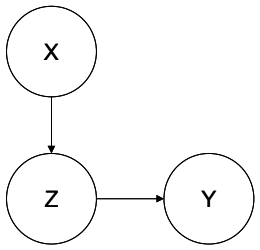


Figure 3: Conditional Independence of *X* and *Y*.

One will notice the absence of an arrow, or arc, from *X* to *Y*, indicating conditional independence, or lack of a direct association.(90) The establishment of conditional independence between two variables simplifies an influence diagram or Bayesian network by reducing the number of needed arcs.(91) If there are no associations among variables *X*, *Y*, and *Z*, then the model of mutual independence holds.

3.5.2 Testing Significance of Associations

Marginal and partial associations between variables are determined based on differences in the L^2 statistics of the pertinent loglinear models. This L^2 difference is known as a *component* and also follows the chi square distribution.(92) For example, suppose one wishes to test the significance of a marginal association between variables *X* and *Y* in a three-way table for *X*, *Y*,

and Z. The statistic for the model of mutual independence among X, Y, and Z (L_0^2) is compared to the statistic for the model that that additionally contains an interaction term for X and Y (L_i^2), thereby obtaining the component L^2. Specifically,

$$L^2 = L_0^2 - L_i^2.$$

In addition, the difference in their degrees of freedom is also calculated, as shown below.

$$df = df_0 - df_i.$$

If the component L^2 is large relative to its component degrees of freedom (*df*), then the association between X and Y is significant.[93] In this test of marginal association between X and Y, the variable Z was ignored, or summed over.[94]

The previous test and loglinear models in general are represented using a conventional notation. For example, for variables X, Y, and Z, the model of mutual independence is represented as follows:

[X] [Y] [Z].

A test of marginal association between X and Y is indicated by a comparison of the above model with the following model, which additionally contains an interaction term for X and Y:

[X] [Y] [Z] [XY].

A test of partial association can also be represented using the conventional notation. To test for a partial association between X and Y in the presence of Z, the following model containing all two-way associations except *[XY]:*

[X] [Y] [Z] [XZ] [YZ],

is compared to the model additionally containing an association term for X and Y

[X] [Y] [Z] [XZ] [YZ] [XY].

If the component L^2 is large enough, then a partial association between X and Y exists.[95] Note that the effect of Z was removed by the inclusion of all two-way associations involving Z in the former model.

3.5.3 Assessing Associations in Large, Sparse Tables

Significance testing is problematic in the case of a large, sparse contingency table as well as a large sample size. However, a large, sparse table is often the type of table that investigators work with.[96] Such tables contain many variables or categories and thus many cells with zeros or small cell counts less than five, despite a large sample size. This is problematic for the use of chi-square statistics, such as L^2. These statistics are suspect under conditions of sparseness because L^2does not follow the chi square distribution in this case.[97] This is also known as Cochran's Rule.[98]

However, although significance testing is suspect with sparse tables, the existence of associations between variables can still be determined using the effects, or *lambda* (λ), parameters.[99, 100] A lambda parameter indicates the strength of an effect, or its importance in explaining any deviation from a flat distribution of the cases among the categories, or cells. Thus, there is a lambda parameter for each combination of the categories of the variables, and a lambda parameter can be positive or negative, with the sign indicating the direction of influence of the effect. For example, in a symmetric loglinear model, a lambda with a positive sign indicates that the effect is responsible for a relative increase in the number of cases in the cell.[101] For an asymmetric model, which assumes a response variable, a lambda with a positive sign indicates increased odds that the response variable equals a given value. In other words, there are a larger proportion of cases associated with the particular value of the response variable.[102, 103] The lambda parameters are much more robust than a chi-square or standardized

lambda test. The lambdas are also insensitive to sample size if the sample size is not small.[104] In general, two variables are considered *not* directly associated if the maximum lambda for the variable pair is less than 0.20 in absolute value. Hence, X and Y are *not* associated if $\max|\lambda_{ij}| < 0.20$, where i and j represent any two categories of X and Y, respectively.[105, 106]

3.6 MORE ON THREE STEP MODELS

3.6.1 Disadvantages of a Three Step Approach

The main disadvantage to the three step approach is the bias introduced in the structural model due to the classification errors of the latent variables. The latent variables are treated as observed variables in the structural model, but they are actually predicted variables with some degree of prediction, or classification, error. The use of latent variables in this manner leads to bias, which causes attenuation, or underestimation, of the strength of the relationship between latent and observed variables.[107, 108, 109]

Two tactics can be used to mitigate the bias. First, greater emphasis can be placed on the classification ability of the latent variable, although this may come at the expense of fit.[110] Second, a correction procedure developed by Bolck, Hagenaars, and Croon can be applied to the (biased) joint distribution of the observed and *predicted* latent variables to obtain the joint distribution of the observed and *true* latent variables.[111] Both tactics were applied in this research. The correction procedure adjusts, or corrects, the biased joint distribution using a transition matrix, which is constructed using characteristics of the latent variable determined during latent class analysis.

3.6.2 Correction Procedures

The correction procedures for three-step modeling developed by Bolck, Croon, and Hagenaars involve adjustment of the matrix containing the joint distribution of the observed and predicted latent variables. This is done using one or more transition matrices, depending on the number of latent variables. The result is a corrected matrix containing the joint distribution of the observed and true latent variables.[112] In the case of one latent variable and one or more observed variables, the relationship between the uncorrected and corrected matrices is given as

$$E = AD$$
where
E = *uncorrected matrix of observed variables and 1 predicted latent variable* (s)
A = *corrected matrix of observed variables and 1 true latent variable* (t)
D = *transition matrix.*

Equation 9

Using matrix algebra, the corrected matrix A is determined as follows:

$$A = ED^{-1}.$$

Equation 10

The corrected matrix A was used for the loglinear modeling versus the uncorrected, or original, table E. The contents of corrected matrix A were rounded to the nearest integer prior to modeling.

The transition matrix D is calculated using the conditional and classification probabilities determined as part of the latent class analysis, as shown below.

$$D = \sum p(y \mid t)\, p(s \mid y).$$

Equation 11

The factor $p(y \mid t)$, in which t represents the true latent variable and y represents the response pattern, is calculated as the product of the conditional probabilities associated with response pattern y in latent class t. The factor $p(s \mid y)$ corresponds to the classification of each response pattern. Assuming modal classification, which was used in this research, $p(s \mid y) = 1$ if response pattern y is assigned to predicted class s and 0 otherwise.[113]

In the case of a joint distribution involving two or more latent variables and one or more observed variables, a more general correction procedure is needed. The previous correction formula (Equation 10) cannot be applied in these cases based on the matrix algebra. Therefore, a more general procedure was developed by Dr. Marcel Croon in January 2005 in response to these more complex joint distributions, which are present in this research.[114] The more general procedure was not part of the published correction procedures by Bolck et. al. The general procedure involves concepts from advanced matrix algebra, such as the Kronecker Product.

In order to present this general correction procedure, the Kronecker Product will be defined for the case of two matrices, although it can be extended to more than two. Assume $A$is an $n \times m$ matrix and B be an $r \times s$ matrix. Their *Kronecker Product* $A \otimes B$ is the $nr \times ms$ super matrix formed from all possible products of the elements of A with those of B.[115] Also, the *vectorization* operation (*vec)* for a matrix consists of writing the elements of the matrix as a single vector by stacking the columns. Using the case of three latent and two observed variables as an example, the relationship between uncorrected matrix Q and corrected matrix P is given as

$$vec(Q) = A \otimes B \otimes C \; vec(P)$$

where

Q = *uncorrected joint distribution of observed and predicted latent variables* **Equation 12**

$A \otimes B \otimes C$ = *Kronecker Product of transition matrices A, B, and C*

P = *corrected joint distribution of observed and true latent variables*.

The transition matrices *A, B*, and *C* are associated with the three latent variables and are determined as previously using Equation 11.[116] The general procedure (Equation 12) can be extended to include additional latent and observed variables. For each additional latent variable, there is an additional transition matrix. Equation 12 is solved algebraically for the corrected matrix *P*, which is used within the loglinear modeling, as given by

$$vec(P) = (A^{-1} \otimes B^{-1} \otimes C^{-1}) \; vec(Q).$$

3.7 LATENT CLASS ANALYSIS

The second component of the modified *LISREL* approach is latent class analysis, which performs the measurement portion of the modeling. *Latent Class Analysis (LCA)* is a technique used to determine a categorical latent variable from an analysis of the relationships among cross-classified categorical indicator variables. A latent variable is an unobserved variable that cannot be measured directly. An example of a latent variable is a person's attitude as portrayed through a survey. A latent variable can be measured only indirectly using observed or manifest variables, which are also referred to as *indicator variables*. An example of an indicator variable is a survey question.[117] The basic premise of a latent variable is that it explains or accounts for the relationships among the indicator variables.[118]

Latent class analysis is often referred to as a categorical analog to factor analysis and was originally conceived as a method for survey analysis in the social sciences.[119] Factor analysis and LCA are similar in that both methods explore the latent structures among a group of observed variables. Within three step modeling, a latent variable enables the researcher to work with one simple "predicted variable" versus many indicator variables.[120] Since a latent variable explains the associations among its indicator variables, the indicators are simplified to a more basic and general latent construct.[121, 122, 123, 124] In essence, various associated nominal variables are "combined." Thus, in this research, latent class analysis was used as a variable simplification and reduction tool. The following sections describe the various latent class analysis fundamentals necessary used in applying this technique within this research.

3.7.1 Model Building Strategy

The outcome of a latent class analysis is a latent variable, which contains a number of categories, or latent classes. The objective is to choose the simplest model, or the model with the fewest classes, that has acceptable fit and classification ability.[125, 126] Thus, the model builder must attempt to balance simplicity with fit and classification ability. In choosing the number of classes for the latent variable, the first model that is tested is the model of independence, which has one class. If this model is acceptable, the indicator variables are *not* associated, and a latent variable is *not* necessary. However, if a one-class model is not acceptable, then models containing several class are evaluated, starting with two.[127] To compare models containing a different number of classes, the best run for each model is used.[128] Based on these best runs for different models, a final model is chosen based on a comparison of fit, classification ability, and parsimony.

3.7.2 Output Parameters

There are two types of parameters estimated as part of a latent class analysis. *Latent class probabilities* describe the sizes, or distribution, of the classes of the latent variable and sum to one. A *conditional probability* parameter is the probability of a particular category of an indicator variable given the latent class. In other words, a conditional probability is the probability that an indicator variable has category *i* given the latent variable has class *t*. It indicates the degree of the relationship between the category and the latent class. The conditional probability parameters are used to interpret and name the latent classes.[129] The conditional probabilities for an indicator variable within a latent class sum to one.[130]

3.7.3 Max Likelihood Estimation of Parameters

The latent class and conditional probabilities are typically estimated using a max likelihood *(ML)* approach.[131, 132, 133] However, there is no closed-form *ML* solution for these parameters, and most software packages use an iterative procedure known as the Expectation Maximization *(EM)* algorithm to estimate the parameters.[134, 135] The *EM* Algorithm begins with trial values for the parameters and iterates until the change in the estimated parameters is less than a pre-defined tolerance or until the maximum number of iterations is reached.[136] A caution with the use of the *EM* Algorithm is its tendency to converge to local maximums. However, performing many runs of a model using different start values for the parameters allows a determination of the optimal solution with a high degree of certainty.[137]

3.7.4 Goodness of Fit

One criterion used in choosing the best model is fit. There are various statistics and measures used to assess goodness of fit. These include the Pearson Chi Square statistic (X^2) and Likelihood Ratio Chi Square statistic (L^2), which are used for significance testing. In addition,

the Index of Dissimilarity (I_d), Normed Fit Index (*NFI*), Bayesian Information Criterion (*BIC*), Akaike Information Criterion (*AIC*), and the Consistent Akaike Information Criterion (*CAIC*) are other measures that can be used to assess fit. The statistics X^2 and L^2 have the drawback of being dependent on the sample size. They tend to reject a model when the sample size is large, even though the model is reasonable. If X^2 or $L^2$is used to assess the fit of a latent class model, then the model is accepted as fitting the data if the chi square statistic is small enough relative to the degrees of freedom. This is opposite of the traditional goal or rejecting the null hypothesis of independence by obtaining a large test statistic. In finding the best fitting model, we hope to accept the hypothesized model.(138)

Therefore, other measures are used to assess fit when the sample size is large, as in this research. For instance, the Index of Dissimilarity (I_d) takes sample size N into account and is defined as follows:

$$I_d = \frac{\sum_s |n_s - \hat{n}_s|}{2N}. \qquad \textbf{Equation 13}$$

As a general rule, values of I_d less than 0.05 are considered small and indicate good fit. Thus, $I_d \leq 0.05$ provides a target range for good fit when the sample size is large.(139) In this research, the I_d is one of the primary measures for assessing fit.

The Normed Fit Index (*NFI*) is another measure that can be used to assess fit with large sample sizes. This index is calculated by comparing the likelihood ratio chi square of the model being tested (L_i^2) with that of a baseline model (L_0^2), such as a one-class model, as shown in Equation 14.(140)

$$NFI = \frac{L_0^2 - L_i^2}{L_0^2}$$

where

$L_0^2 = likelihood\ ratio\ chi\ square\ for\ baseline\ model$

$L_i^2 = likelihood\ ratio\ chi\ square\ for\ tested\ model.$

Equation 14

When the *NFI* is between 80% and 90%, then goodness of fit is suggested. In other words, when a model begins to account for 80-90% of residuum variation, then the model has good fit.[141, 142]

The Akaike Information Criterion (*AIC)* is a measure of model fit based on concepts from information theory. The *AIC* accounts for the number of independent parameters and is a parsimony index because it favors models with fewer parameters. However, a criticism of the *AIC* is that it does not take *N* into account. There are no critical values or targets, but smaller is better.[143] It is calculated as shown in Equation 15.[144]

$$AIC = -2\ln(L) + 2m$$

where

$L\ (Likelihood) = \prod_s P(Y_s)^{n_s}$

$m = number\ independent\ parameters$

$P(Y_s) = probability\ of\ response\ pattern\ Y_s$

$n_s = number\ observed\ cases\ for\ response\ pattern\ Y_s.$

Equation 15

The Bayesian Information Criterion (*BIC*) takes both *N* and the number of independent parameters into account. Relative to the *AIC*, it tends to select less complex models, since it heavily penalizes for the number of parameters when *N* is large. Similarly, there are no critical

values for *BIC*, but smaller is better.[145] The Consistent Akaike Information Criterion (*CAIC*) is similar to the *BIC* in that it penalizes for both sample size and number of parameters.[146] The smaller the *AIC*, *BIC*, or *CAIC*, the better the model.

3.7.5 Classification

Each pattern of the indicator variables is assigned to a class of the latent variable. Each pattern is assigned based on the modal conditional probability. The *modal conditional probability* is the largest probability of membership in a class of the latent variable given the particular pattern.[147, 148]

Since modal assignment is probabilistic, measures of classification performance are calculated. These include the classification error (P_e) and Goodman and Kruskal's Lambda (λ).[149] Lambda is a proportional reduction in error (*PRE*) measure that determines the proportional decrease in the error rate when modal assignment is used versus assignment of all patterns to the largest latent class in the model. Specifically, λ is calculated as given in Equation 16.[150, 151]

$$\lambda = \frac{E_1 - E_2}{E_1}$$

where **Equation 16**

$E_1 = error\ rate\ for\ assigning\ all\ patterns\ to\ largest\ latent\ class$

$E_2 = error\ rate\ for\ modal\ assignment\ of\ patterns.$

The closer λ is to one, the better the classification performance, or predictive ability, of the model.[152] An LCA model should be judged not only on its fit but also on its ability to classify the patterns.[153, 154]

3.7.6 Identifiability

An LCA model must have the property of being identified. When a model is *identified*, there is a unique set of parameters associated with a value of L^2.[155] Thus, when a model is *not* identified, more than one set of latent class or conditional probabilities exists for the same value of L^2. *Local identifiability* applies to any given run of a model, while *global identifiability* applies to the optimal run, or the run with the minimum L^2. Local identifiability indicates whether there are additional parameter solutions for the same L^2 in the same neighborhood.[156] A necessary condition for identifiability is non-negative degrees of freedom.

3.7.7 Local Maximum Solutions

As indicated in section 3.7.3, a latent class model often converges to local maximum solutions, which have a larger L^2 than that of the optimal model.[157] A local maximum solution can differ substantially from the optimal solution in terms of the parameters. Convergence to local max solutions is a noted problem in latent class analysis. Therefore, it's imperative to run a model many times using random parameter start values to arrive at the minimum L^2 for the model, as was done for the latent variables in this research.[158, 159] The automation of this process by software is advantageous and was necessary to build the latent class models in this research. Once two separate runs having the same minimum L^2 are found, they are then verified to have the same parameters (latent class and conditional probabilities). If their parameters are equal, the model is globally identified.

3.7.8 LCA Software

There are various software products available to perform latent class analysis. Some are academically developed and/or freely-downloadable, such as *LEM* and *MLLSA*. The product used in this dissertation was *Latent Gold 3.0*, a commercial product by Statistical Innovations.

Latent Gold allows a model to be automatically run many times, each time using a random set of start values, to ensure the optimal run is found. The best of these runs is reported as the resultant model. The functionality within *Latent Gold* enabled easy and fast determination of the globally identified solution for a given number of latent classes.

3.8 BAYESIAN NETWORKS

The final topic on categorical data modeling to be introduced is the Bayesian network. A *Bayesian network* is a graphical decision model consisting of variables, represented by nodes, as well as the direct dependencies or associations between the variables, which are represented by arcs. It is a directed graph that does not contain cycles. A Bayesian network is used for probabilistic *inference*, or querying the probabilities of certain variables when the values of other variables are known. For example, one of the main applications of a Bayesian network is determining the most likely cause for a given effect, also known as *diagnostic*, or bottom-up, reasoning. Top down reasoning can also be performed, in which the probability of effects given causes is computed.[160, 161, 162] Within diagnostic reasoning, the explanatory variables can be ranked based on their value of information and the degree to which they reduce the uncertainty of the effect.[163, 164] For the hazmat release model, this was used to identify the variables that should be the top priorities for policy change. In general, the Bayesian network has become a popular means of modeling expert or decision support systems, such as for medical diagnosis or other trouble-shooting applications.

The dependency, or association, structure of a Bayesian network is one of its two major components. Only two-way associations and conditional independencies are depicted in directed graphs. The absence of an arc indicates conditional independence between two variables.

Three-way and higher-order associations are represented only indirectly through multiple arcs. With a directed graph, if two variables are connected by an arc to a third variable, the three-variable interaction is automatically represented in the graph by the connecting arcs.[165] Therefore, even if the exact form of the relationships among the variables is not known, it does not matter because the uncertainty is represented probabilistically.

A typical method of building the structure of a small to moderate sized Bayesian network is manually with the assistance of an expert. Newer methods, which are often applied to larger networks or in the absence of a readily available expert, are machine learning or algorithmic approaches involving inductive inference or search for the most probable structure.[166] Learning modules were implemented in academic Bayesian network software beginning in the early 1990's.[167] A learning module was just implemented in *GeNIe*, the decision model software used in this research, in the summer of 2005. An opportunity for future research is a comparison of the results of loglinear modeling with those of learning algorithms for building the structure of the network.

The second major component of a Bayesian network is the quantitative portion, and it represents the joint probability distribution among all the variables. The joint probability distribution is calculated using the conditional probability distribution associated with each node in the network. The conditional probability distribution of a node is the probability that the node takes on each of its possible values given every combination of values of its parent nodes. The joint and conditional probabilities are related according to the chain rule. The *chain rule* states that for a Bayesian network over the variables $U=\{A_1,...,A_m\}$, the *joint probability* distribution $P(U)$ is the product of all conditional probability distributions specified in the network.

Specifically,

$$P(U) = \prod_i P(A_i \mid pa(A_i)), \qquad \textbf{Equation 17}$$

where $pa(A_i)$ is the parent node set of node A_i. When a variable has no parents, the probability distribution is the prior distribution.[168] In order to determine the quantitative portion of the Bayesian network, the conditional probability distribution for each variable, or node, must be calculated. The *conditional probability* distribution for a variable A given its parents B and C is calculated according to Equation 18.[169]

$$P(A \mid B, C) = \frac{P(A \cap B \cap C)}{P(B \cap C)}. \qquad \textbf{Equation 18}$$

This equation is easily extended to include additional parent variables by including them in the numerator and denominator in the same manner as B and C. To calculate $P(A \mid B, C)$ for category combination $A=i$, $B=j$, and $C=k$, the number of records in which $A=i$ and $B=j$ and $C=k$ is divided by the number of records in which $B=j$ and $C=k$.[170] The conditional probability distribution for A contains i x j x k probabilities, so there is a probability associated with each category combination i,j,k.

Conditional probabilities can be determined based on record counts from a database or subjective data or beliefs from an expert. All probabilities calculated for the Bayesian network in this research were calculated using record counts, or frequency data, from the HMIRS database. Frequency data can be used when dealing with repetitive events that have been recorded. However, a database may not be available, or the event may not be repetitive, for

instance a nuclear war. In these cases, the conditional probabilities must be assessed subjectively by an expert. The subjectivist view considers probability as a measure of personal belief. Hence, Bayesian networks are also known as belief networks.[171]

The foundation of inference in Bayesian networks is Bayes Theorem, which enables inference in any direction in the network. Using Bayes Theorem, some probabilities are updated based on new *evidence*, or specific values, of other probabilities.[172] Several algorithms exist for performing inference in a Bayesian network. The clustering algorithm, in which the directed graph is converted to a junction tree where the probabilities are then updated, is the fastest known exact algorithm. The clustering algorithm is the default algorithm implemented in *GeNIe,* which is discussed next.[173]

3.8.1 Bayesian Network Software

The decision model software used in this research was *GeNIe*, a graphical decision-theoretic package developed at the Decision Systems Lab at the University of Pittsburgh. *GeNIe* is a development environment for Bayesian networks and influence diagrams and is available to the community at no cost. Using *GeNIe*, the modeler builds the network structure using circular nodes and arcs in an intuitive, graphical environment. Conditional probability distributions can be copied into *GeNIe* for each node, making the construction of the network very efficient. Once this is complete, the various forms of inference discussed previously can be performed.

Decision models can be studied in terms of value of information, which refers to the information value of a parent variable relative to the outcome variable. The information value of a parent variable can also be viewed as its ability to influence or reduce uncertainty in the outcome variable. An entropy-based value function is used in *GeNIe* to rank the parent variables based on their information content in relation to the outcome variable. This function determines

the decrease in entropy, or uncertainty, by observing a given parent variable. Entropy is a concept from the field of information theory and is used to measure the information value of a variable, which represents the expected amount of information needed to classify a new instance involving the variable.[174]

4.0 METHODOLOGY

A methodology for determining a data-directed decision model from a categorical dataset is being proposed and demonstrated in this research. The major components of this methodology include simplification, determination of associations, and construction of a Bayesian network model, as shown in Figure 4 by way of review.

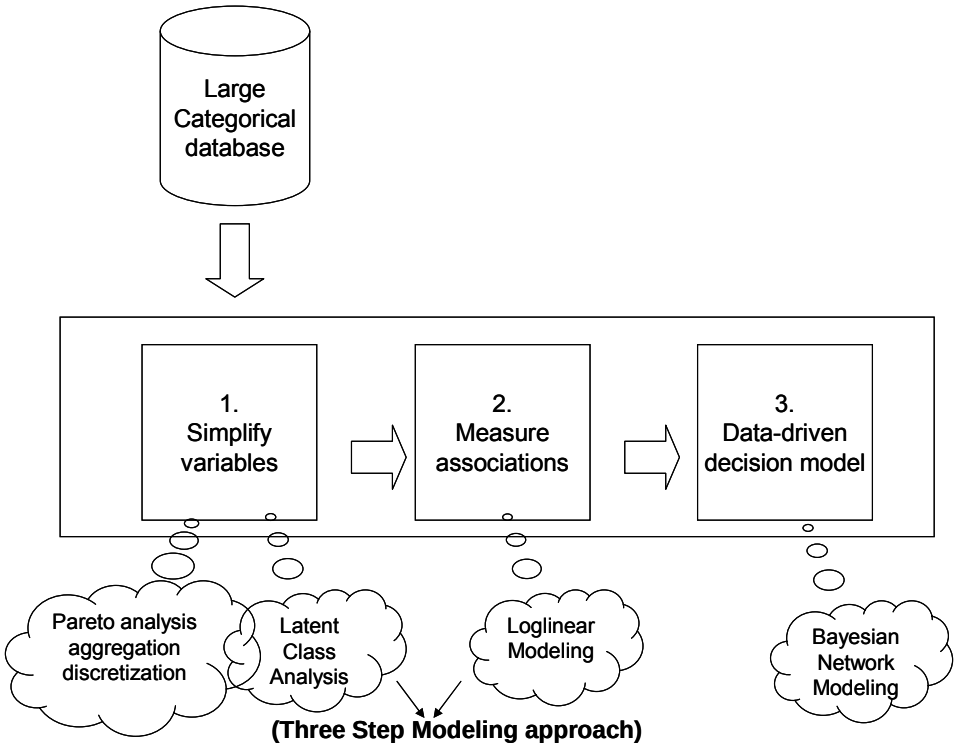


Figure 4: Methodology for Building a Decision Model.

The variable domain was simplified for purposes of model building. The simplification was accomplished using Pareto analysis, data aggregation, discretization, and latent class analysis.

Pareto analysis was used to eliminate infrequent categories and variables from consideration to decrease the sparseness of the contingency tables used for latent class analysis and loglinear modeling. Data discretization was applied to the outcome variables release quantity and dollar loss, which were continuous variables. These variables were made discrete based on their distribution as well as expert input. This was done to simplify the data and to enable these variables to be used within loglinear and Bayesian network modeling. After the Pareto analysis and data aggregation, latent class analysis was used to further simplify the variable domain by combining related variables to form latent variables. Using latent class analysis, the number of variables in the decision model was reduced from 24 to 11. The simplification strategy applied to the domain of variables is summarized below in Figure 5.

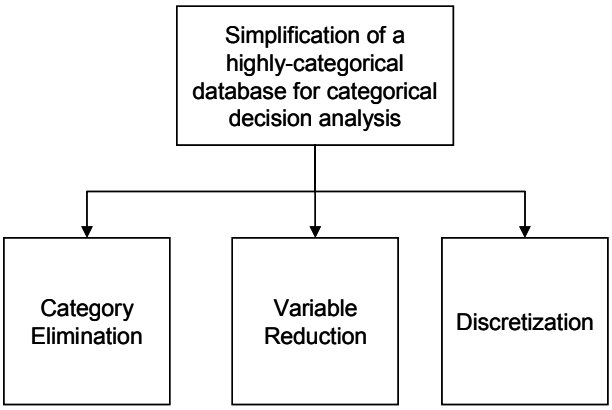


Figure 5: Simplification Strategy for a Highly-Categorical Database.

The 11 simplified variables were used to construct a time-ordered, base network structure of a hazardous materials release. In order to determine accurate relationships, or associations, between the variables, an exploratory loglinear modeling approach was taken, as part of a three-step approach for the modeling of categorical latent variables. In determining these associations, the downstream variables in the network served as response variables to the upstream variables, as the loglinear, or logit, modeling proceeded from left to right in the network. An exploratory

analysis was considered the best approach based on the gaps in the literature, the possibility for non-obvious relationships, and the large amount of available data. Loglinear modeling has been used previously by social science researchers to build data-congruent path diagrams.[175, 176]

The associations determined using loglinear modeling were used to construct the structural, or qualitative, portion of a Bayesian network decision model. The joint probability distribution among the 11 variables, which forms the quantitative portion of the Bayesian network, was obtained from the database using incident counts. The Bayesian network was used for making inferences on the variables, including ranking the explanatory variables and analyzing desirable changes for them. Starting with the simplification techniques of Pareto analysis, data aggregation, and discretization, the overall methodology is demonstrated using the DoT's hazardous materials release database as the worked example.

4.1 WORKED EXAMPLE

A general, high level methodology for development of a data-driven decision model based on categorical data was proposed in the previous section. This methodology is demonstrated in the following sections using an engineering problem as the worked example. The problem is the decision model of a hazardous materials release during transportation-related unloading of containers. The decision model will be used for identification of critical variables and operational change analysis related to these types of hazmat releases. In general, the decision model can be used to gain a better understanding of the hazmat release problem in order to decrease the severity of incidents. In the following sections, the proposed methodology for

decision model construction is carried out. Specifically, the following sections describe the Pareto analyses, data aggregation, discretization, three-step modeling procedure, and Bayesian network construction, as applied to the large, categorical hazmat release database.

4.1.1 Simplification

The simplification of the variable domain for the hazmat release problem using the techniques described previously is demonstrated in the following sections. The application of Pareto analysis, data aggregation, discretization, and latent class analysis to the hazmat release database is demonstrated.

4.1.1.1 Data Sources and Incident Types

The most complete source of data on hazardous materials releases is the HMIRS, the database maintained by the DoT's Office of Hazardous Materials (OHM). If hazardous materials are unintentionally released during commercial interstate or intrastate transport, a written report must be submitted for entry into the HMIRS. The HMIRS is readily available on the internet in the form of downloadable datasets covering years 1993 to the present. There are approximately 149,000 records from January 1993 to July 2002, the time period being considered in this research.

The HMIRS was compared to a state database that also tracks hazmat releases. The state of Ohio, which has the largest number of off-road highway incidents according to the HMIRS, maintains a database for commercial transport releases. However, based on a comparison of Ohio's database to the HMIRS, the federal reporting requirement indicates the desirability of using federal data for off-road incidents. In Ohio, it is not mandatory for commercial carriers to report unintentional releases. The entities that typically report include regulatory and local

emergency response agencies. As a result, only 30-50 releases were recorded annually between 2000 and 2001 in Ohio's database.[177] This contrasts with an annual average of 1356 highway releases recorded for Ohio in the HMIRS between 2000 and 2001.

The majority of hazmat incidents in the United States are related to the highway mode of transport as opposed to air, water, and rail. The highway mode is associated with 86% of incidents. Highway-related incidents occur both on and off the road, but 88% occur off the road. Of these, 73% occur during the unloading of hazardous materials. This compares to 22% during loading and 5% during storage operations. Due to the prevalence of incidents that occur during unloading, the incident type considered is limited to unloading release incidents within the United States. Based on an analysis of the HMIRS, there are approximately 80,000 incidents meeting these criteria in the HMIRS. However, after applying Pareto, approximately 40,000 were used in constructing the association structure of the Bayesian network. As shown in Table 2, the percentage of unload incidents from 1993 to mid 2002 has remained fairly constant. However, one will notice a relative increase in 2001. This may be the result of a requirement beginning in October 1998 to report intrastate as well as interstate incidents.[178] Therefore, this research does not consider a trend in incidents over this time period.

Table 2: Unload Incidents by Year.

YEAR	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Unload Incidents	7394	8701	8145	7200	7174	7613	9241	9225	10451	4834
Total Highway Incidents	11074	13984	12762	11909	11852	12995	14963	15012	14921	6852
Unload Percentage	0.668	0.622	0.638	0.605	0.605	0.586	0.618	0.615	0.7	0.706

4.1.1.2 Pareto Analysis and Data Aggregation

The need to simplify the variable domain was apparent at the outset of the data analysis. The variables were multi-valued, ranging from two to thirty-six categories. Despite the large sample size, the large number of categories created sparse contingency tables containing many zeros or small cell counts of less than five. Unfortunately, the use of chi square statistics for significance testing in sparse tables is suspect. Therefore, various categories were eliminated from consideration.

Another important reason for reducing the categories considered was an increased chance of convergence of the loglinear models. Loglinear models often do not converge when the contingency table contains many zeros, which is driven by a large number of categories.[179] For example, an early model involving 13 container types did not converge.

In order to determine the categories to retain for modeling, a frequency analysis of each variable based on involvement in incidents was done. In general, the categories associated with 80% of the incidents were included. Thus, the 80/20, or Pareto Principle, was applied when possible. This principle focuses on the top 20% of the factors or categories that are associated with 80% of the outcome.[180] For example, material type approximately follows the Pareto Principle since two of the nine material types, corrosives and flammable liquids, are associated with 80% of the unloading incidents. The Pareto analysis of each variable and the categories retained for latent class modeling and loglinear analysis are described in the following sections.

In addition, there were several variables that were natural candidates for data aggregation, or generalization, as a means of simplification. These variables included date, time, and U.S. state. Data aggregation or generalization was desirable and possible with these variables because they included natural groupings.

A note on the selection of categories and variables as part of the simplification is in order. Although Pareto was applied to most of the variables and sets of binary variables, it was not applied to all of them due to modeling constraints. The variables and sets to which Pareto was not applied were container type, failure item, and failure area, as will be discussed. This resulted in the dataset of unloading incidents being reduced by 50% from approximately 80,000 records to 40,000 records. The issues of model testing and convergence, which are affected by the size and sparseness of the contingency table, became the overriding factors for determining the categories and variables retained for container type as well as the failure item and area binary sets. In addition, the binary variables in these two sets were combined to form one latent variable versus a latent variable for each set. This was done so that the largest loglinear model would have a maximum of ten variables, due to an *SPSS* limitation. Also, the ability to interpret latent class models is enhanced when the number of indicator variables and categories is kept small. A summary of the approach used in this research for category elimination is provided in Figure 6.

Based on this, the infusion of subject matter knowledge can contribute to the methodology by removing some of the arbitrariness in selecting categories and variables and in general adding "art" to the science. The modeler in this case would likely have better or additional reasons for retaining or eliminating certain categories or variables.

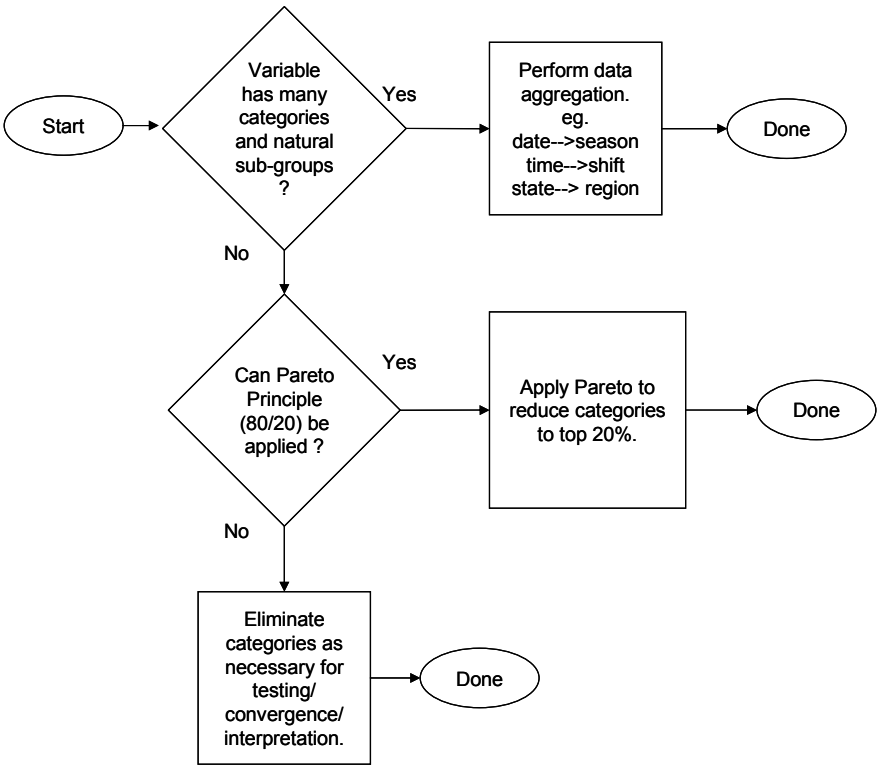


Figure 6: Strategy for Category Elimination.

The non-simplified variables to be discussed in the following sections are summarized below.

Table 3: Non-Simplified Variables in the Hazmat Release Network.

Variable	Data Type	Number of Categories	Range
Area Type	nominal	3	
Container Type	nominal	36	
Dollar Loss	continuous		$0-$43,760
Geographic Division (State)	nominal	9	
Land Use	nominal	5	
Material Type	nominal	9	
Release Quantity	continuous		0-2000 gal. 0-200 lb.
Season (Date)	nominal	4	
Shift (Time)	nominal	3	

Table 3 (continued).

Variable Set	Data Type	Number of Variables	Range
Causing Object	binary	9	
Contributing Action	binary	18	
Failure Area	binary	8	
Failure Item	binary	8	
Failure Mode	binary	8	

Area Type Area type describes the location of the incident in terms of a suburban, urban, or rural setting. Suburban incidents occurred most frequently, with urban incidents following closely behind, as shown in Table 4. If an area type was not reported, as in 2% of the incidents, the record was excluded from the analysis. Since suburban and urban accounted for over 80% of the incidents, the rural area type was not included in the analysis.

Table 4: Unload Incident Count by Area Type.

	Area Type	Incident Count	Incident Percentage	Cumulative Percentage
1	Suburban	34,809	0.44	0.44
2	Urban	32,591	0.41	0.84
3	Rural	11,043	0.14	0.98
4	Not Reported	1,535	0.02	1.00

Land Use Land use is another location-related variable that describes the scene of the incident in terms of a commercial, industrial, residential, agricultural, or undeveloped setting. Commercial and industrial incidents were about equally prevalent and accounted for the great majority of incidents at 96%, as shown in Table 5. Residential settings were associated with only 2% of incidents, while agricultural and undeveloped accounted for less than 1% each.

Approximately 2% of incidents had a non-reported land use and were not included. Based on the large percentage of commercial and industrial incidents, the remaining categories were eliminated.

Table 5: Unload Incident Count by Land Use.

	Land Use	Incident Count	Incident Percentage	Cumulative Percentage
1	Commercial	39,561	0.49	0.49
2	Industrial	37,224	0.47	0.96
3	Residential	1,626	0.02	0.98
4	Not Reported	1,237	0.02	1.00
5	Agricultural	187	0.00	1.00
6	Undeveloped	143	0.00	1.00

State/Geographic Division The third variable that describes the location of the incident is the U.S. state. However, for simplification purposes, state was generalized to geographic division, which is based on the nine U.S. Census Bureau divisions.[181] These nine divisions and their constituent states are shown in Table 6. The East North Central division, which consists of Illinois, Indiana, Michigan, Ohio, and Wisconsin, accounted for the largest number of incidents with 21%. Next, the South Atlantic and Middle Atlantic divisions were associated with 16% and 15%, respectively. All nine divisions were included in this research.

Table 6: Unload Incident Count by Geographic Division.

	Division	States	Incident Count	Incident Percentage	Cumulative Percentage
1	East North Central	IL, IN, MI, OH, WI	17,145	0.21	0.21
2	South Atlantic	DC, DE, FL, GA, MD, NC, SC, VA, WV	12,536	0.16	0.37
3	Middle Atlantic	NJ, NY, PA	12,058	0.15	0.52
4	Pacific	AK, CA, HI, OR, WA	8,704	0.11	0.63
5	West South Central	AR, LA, OK, TX	8,080	0.10	0.73
6	West North Central	IA, KS, MN, MO, ND, NE, SD	6,721	0.08	0.82
7	East South Central	AL, KY, MS, TN	5,686	0.07	0.89

Table 6 (continued).

8	Mountain	AZ, CO, ID, MT, NM, NV, UT, WY	5,473	0.07	0.96
9	Northeast	CT, MA, ME, NH, RI, VT	3,575	0.04	1.00

Date/Season In order to create a simplified and discrete variable for modeling, incident dates were aggregated based on the season, using the ranges shown in Table 7. Incidents occurred most frequently during the summer season (29%) but were nearly as prevalent in the spring (28%). The fall and winter seasons ere associated with 22% and 21% of incidents, respectively. All four seasons were analyzed in this research due the proximity of their percentages.

Table 7: Unload Incident Count by Season.

	Season	Begin Date	Incident Count	Incident Percentage	Cumulative Percentage
1	Summer	21-Jun	23,189	0.29	0.29
2	Spring	20-Mar	22,567	0.28	0.57
3	Fall	22-Sep	17,513	0.22	0.79
4	Winter	21-Dec	16,709	0.21	1.00

Time/ Shift The occurrence times of unloading incidents were also aggregated, as shown in Table 8, by the work shift. The largest number of incidents occurred during the daytime shift, followed by the midnight and twilight shifts. All three shifts were considered in this research due to the high prevalence of each. If the time was invalid or not reported, the incident was excluded.

Table 8: Unload Incident Count by Shift.

	Shift	Times	Incident Count	Incident Percentage	Cumulative Percentage
1	Day	7 AM – 2:59 PM	33,851	0.42	0.42
2	Midnight	11 PM - 6:59 AM	23,872	0.30	0.72
3	Twilight	3 PM - 10:59 PM	20,339	0.25	0.98

Table 8 (continued).

4	Not Reported or Invalid		1,919	0.02	1.00

Material Type There are nine hazmat classes that group materials based on their dangerous characteristics, as shown in Table 9. Hazmat classes 8 and 3 (corrosives and flammable liquids, respectively) were associated with 80% of the incidents and therefore followed the Pareto Principle. Applying the 80/20 Principle, only classes 8 and 3 were included in this research. Since a given incident may involve more than one material type, an incident may be represented more than once in Table 9.

Table 9: Unload Incident Count by Hazardous Material Class.

Hazard Class	Description	Incident Count	Incident Percentage	Cumulative Percentage
8	Corrosives	32954	0.41	0.41
3	Flammable Liquids	31530	0.39	0.80
6	Toxic and Infectious materials	6140	0.08	0.88
2	Gases (Flammable, Non-Flammable and Toxic)	3851	0.05	0.92
5	Oxidizers and Organic Peroxides	3142	0.04	0.96
9	Miscellaneous	2250	0.03	0.99
4	Flammable Solids, Spontaneously Combustibles, Dangerous When Wet	786	0.01	1.00
1	Explosives	25	0.00	1.00
7	Radioactive Materials	18	0.00	1.00
Not Reported		1	0.00	1.00

Container Type There were 36 container types associated with unloading incidents, as shown in Table 10. The top two container types, fiber box and bottle, were considered in this research. Although they represent only 45% of the incidents, the other types were eliminated to reduce the sparseness of the stage one contingency table. Container types were *not* combined for

simplification since this could be done based on either structure or material. For example, should all drums be combined, or should all plastic containers be combined? The desirability of one criterion versus the other was unknown. An incident can involve more than container type, and so an incident may be represented more than once in Table 10.

Table 10: Unload Incident Count by Container Type.

	Container Type	Incident Count	Incident Percentage	Cumulative Percentage
1	BOX FIBER	42,735	0.34	0.34
2	BOTTLE	13,764	0.11	0.45
3	DRUM METAL	12,887	0.10	0.55
4	TANK	11,953	0.09	0.64
5	JUG	11,432	0.09	0.73
6	DRUM NON-METAL	8,465	0.07	0.80
7	INSIDE CONTAIN	6,888	0.05	0.85
8	CAN	5,246	0.04	0.90
9	CONTAINER	2,804	0.02	0.92
10	BAG PAPER	1,708	0.01	0.93
11	DRUM	1,392	0.01	0.94
12	PAIL	1,292	0.01	0.95
13	JAR	1,156	0.01	0.96
14	CYLINDER	1,000	0.01	0.97
15	BAG PLASTIC	940	0.01	0.98
16	BOX	820	0.01	0.98
17	BAG	747	0.01	0.99
18	JERRICAN	617	0.00	0.99
19	COMPOSITE	253	0.00	1.00
20	TUBE	99	0.00	1.00
21	IBC	92	0.00	1.00
22	OTHER	68	0.00	1.00
23	BAG CLOTH	58	0.00	1.00
24	BOX WOOD	37	0.00	1.00
25	BOX PLASTIC	28	0.00	1.00
26	CARBOY	25	0.00	1.00
27	HOPPER	23	0.00	1.00
28	BATTERY	21	0.00	1.00
29	CYLINDER BULK	9	0.00	1.00
30	BOX METAL	7	0.00	1.00
31	KEG METAL	7	0.00	1.00
32	TANK INTERMODAL	6	0.00	1.00
33	RAM CONTAINER	4	0.00	1.00
34	TANK CRYO	4	0.00	1.00

Table 10 (continued).

35	BARREL/KEG WOOD	1	0.00	1.00
36	TANK CAR	1	0.00	1.00

Container Failure Variables There are several sets of binary variables that describe the failure of the container and subsequent release of hazmat. Each set contains between eight and eighteen binary variables, which represent yes/no responses, such as the container was punctured (yes), or the container was *not* dropped (no). Each set consists of several binary variables that are grouped on the incident reporting form. For example, there is a section on the form for "Action Contributing to Packaging Failure," and it includes yes/no variables such as dropped, improper loading, and loose fitting. Any number of variables within a set may have a "yes" response, allowing for the joint action of various factors.

In order to simplify the variable domain, two general actions were taken relative to the binary container failure variables. First, only the top binary variables in each set were included in the analysis. Second, these top variables were used as indicator variables for a latent variable characterizing the set. For example, using the top variables in the section "Action Contributing to Packaging Failure," a latent variable named Contributing Action was developed. The following sections discuss the various sets of binary variables that describe the failure of the container of hazardous materials.

Contributing Action There are 18 binary variables in the set called Contributing Action, as shown in Table 11. These variables describe the factors and actions that contributed to the failure of the container, such as loose fitting or improper loading. Based on their grouping on the incident form, the variables were used to create a latent variable called Contributing Action.

To simplify the analysis and reduce sparseness, the top four contributing actions, which represent almost 80% of the incidents, were utilized. Thus, the following variables in the set served as indicator variables for the latent variable: other, loose fitting/valve, improper loading, and dropped. Despite its lack of information, the variable "other" was included due to its large association with monetary and human consequences. Based on an analysis of the HMIRS, "other" was associated with 39% of monetary damages, 41% of fatalities, 38% of injuries, and 50% of evacuees, relative to total amounts. However, elimination of the "other" variable related to contributing action and the container failure variables in general, is an item for future research, as will be discussed in section 5.3.

Table 11: Unload Incident Count by Container Failure Contributing Action.

	Contributing Action	Incident Count	Incident Percentage	Cumulative Percentage
1	Other	44276	0.32	0.32
2	Loose Fittings/Valves	25068	0.18	0.51
3	Improper Loading	20979	0.15	0.66
4	Dropped	17041	0.12	0.78
5	Struck/Rammed	12925	0.09	0.88
6	Improper Blocking	5148	0.04	0.92
7	Defective Fittings/Valves	4310	0.03	0.95
8	Overload/Overfill	2563	0.02	0.97
9	Metal Fatigue	1335	0.01	0.98
10	Friction	1293	0.01	0.99
11	Corrosion	715	0.01	0.99
12	Venting	537	0.00	0.99
13	Incompatible Materials	351	0.00	1.00
14	Freezing	185	0.00	1.00
15	Fire/Heat	90	0.00	1.00
16	Vehicle Overturn	63	0.00	1.00
17	Vehicle Collision	57	0.00	1.00
18	Vandalism	25	0.00	1.00

Causing Object Objects that caused the container to fail are represented by the binary variables in Table 12. For example, a combination of the ground and water may have caused the container

to fail. The top four variables, which are associated with 78% of incidents, were used to create a latent variable for causing object. Thus, the indicator variables consisted of the following: none, other, floor/ground, and water/liquid. The relationship of "none" and "other," the top two variables, to consequences was large. Based on this, these variables were not excluded from the analysis, despite the limited information they provide. Relative to total fatalities, evacuees, damages, and injuries, respectively, "other" was associated with 53% of fatalities, 50% of evacuees, 40% of monetary damages, and 43% of injuries. "None" was associated with 47% of fatalities, 32% of evacuees, 29% of monetary damages, and 31% of injuries. However, as discussed previously, "none" and "other" should be removed as part of future research.

Table 12: Unload Incident Count by Container Failure Causing Object.

	Causing Object	Incident Count	Incident Percentage	Cumulative Percentage
1	None	33521	0.26	0.26
2	Other	29550	0.23	0.49
3	Floor/Ground/Roadway	18441	0.14	0.64
4	Water/Other Liquid	18236	0.14	0.78
5	Other Freight	17802	0.14	0.92
6	Forklift	7933	0.06	0.98
7	Nail/Protrusion	2249	0.02	1.00
8	Roadside Obstacle	225	0.00	1.00
9	Other Transport Vehicle	157	0.00	1.00

Failure Mode A set of variables describes the manner in which the container failed, as shown in Table 13. For example, a container may have been crushed, punctured, and/or cracked. The variables other, punctured, and crushed were associated with 82% of the incidents. Consequently, these three variables were used as indicators for a latent variable called Failure Mode. The remaining variables were eliminated from consideration. The "other" failure mode

variable was analyzed for possible exclusion from the analysis. However, despite its lack of information, it will remain in the analysis, as it was related to 68% of monetary damages, 88% of fatalities, 62% of injuries, and 73% of evacuees relative to the totals.

Table 13: Unload Incident Count by Container Failure Mode.

	Failure Mode	Incident Count	Incident Percentage	Cumulative Percentage
1	Other	76093	0.59	0.59
2	Punctured	16591	0.13	0.72
3	Crushed	12719	0.10	0.82
4	Cracked	5937	0.05	0.86
5	Burst/Internal Pressure	5920	0.05	0.91
6	Ripped	5677	0.04	0.95
7	Ruptured	4880	0.04	0.99
8	Rubbed/Abraded	1272	0.01	1.00

Failure Item The item or items on the container that failed are described by the set of binary variables in Table 14. For example, the basic package material itself and/or a closure may have failed. For simplification purposes, the failure item variables were combined with a second group of variables related to the physical aspect of the container to form a latent variable. This second group, the failure area, will be discussed in the next section. The top two indicator variables from each of these two sets of variables were used to develop a latent variable called Failure Item-Area. Thus, in the failure item set, basic package material and closure, which represent 67% of the incidents, were included in the analysis. In the failure area set, the binary variables top and bottom, which represent 63% of the incidents, were used in the analysis. Pareto was not applied to these two sets of variables due to the need to create one latent variable versus two latent variables based on an *SPSS* limitation of 10 total variables. In addition, by limiting the number of indicator variables, the number of classes for the latent variable was also minimized for convergence of the largest loglinear model.

Table 14: Unload Incident Count by Container Failure Item.

	Failure Item	Incident Count	Incident Percentage	Cumulative Percentage
1	Basic Package Material	55438	0.43	0.43
2	Closure	30963	0.24	0.67
3	Other	29199	0.23	0.89
4	Fitting/Valve	7224	0.06	0.95
5	Weld/Seam	4005	0.03	0.98
6	Inner Liner	1065	0.01	0.99
7	Hose/Piping	966	0.01	1.00
8	Chime	357	0.00	1.00

Failure Area The eight areas of the container that may fail are given by the set of variables shown in Table 15. For example, the most frequent area that failed was the top of the container, while the forward, or front, of the container failed the least. The top and bottom areas of the container were used to develop the latent variable Failure Item-Area discussed previously. The "other" failure area was not analyzed due to simplification needs.

Table 15: Unload Incident Count by Container Failure Area.

	Failure Area	Incident Count	Incident Percentage	Cumulative Percentage
1	Top	49907	0.37	0.37
2	Bottom	34916	0.26	0.63
3	Other	29666	0.22	0.86
4	Right	7148	0.05	0.91
5	Left	7124	0.05	0.96
6	Center	3043	0.02	0.98
7	Rear	1090	0.01	0.99
8	Forward	973	0.01	1.00

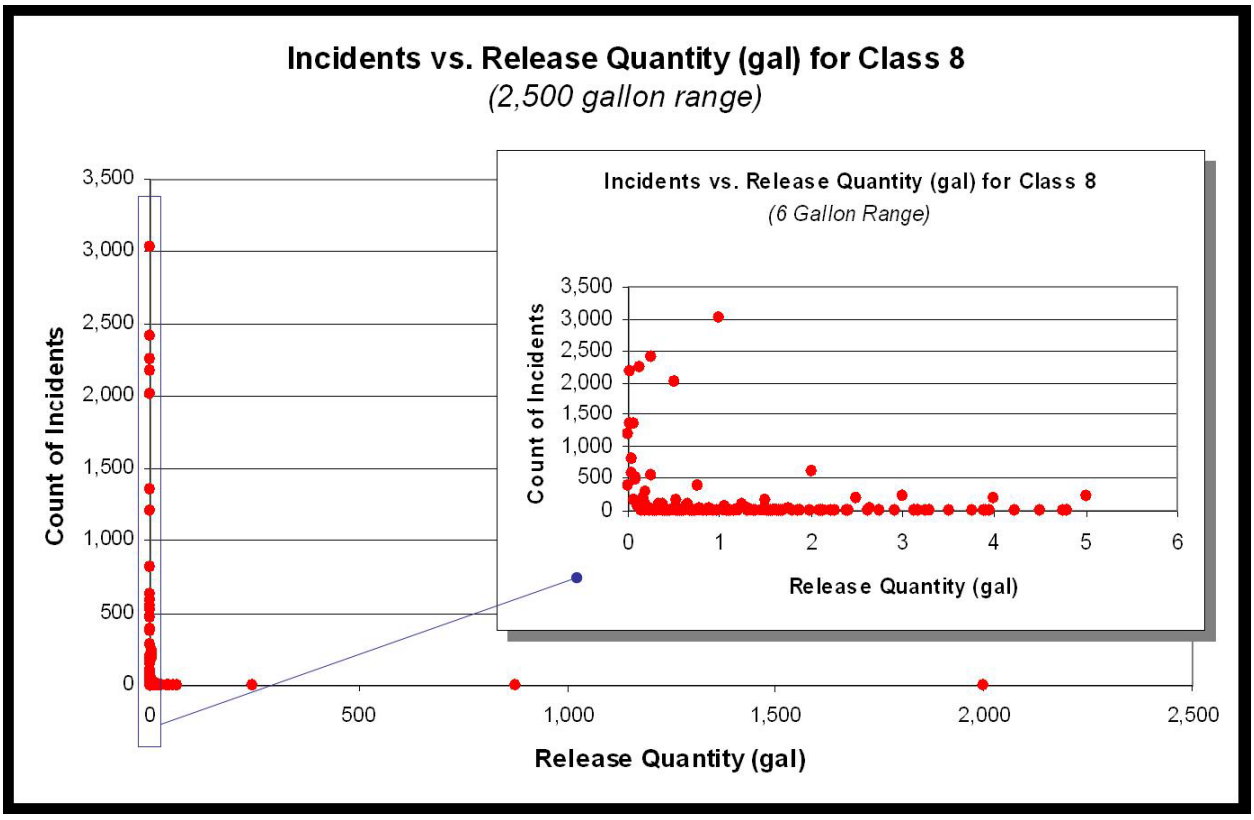
4.1.1.3 Discretization

The two outcome variables release quantity and dollar loss are continuous variables, as reported on the incident form. In order to use them within loglinear and Bayesian network modeling, they were transformed to discrete variables based on both the distribution of the data and expert input. In discretizing these variables, the number of categories was minimized so the ten-variable models in stages four and five, which contain release quantity and dollar loss, would converge. The method and rationale for discretizing these variables is described in the following two sections.

Release Quantity Unlike the variables previously discussed, the quantity of hazmat released is a continuous variable**.** It was captured on the incident form in a free-form fashion as numeric, non-categorical data. A unit of measure was provided by the user, including gallons and pounds. For simplification as well as for usage within categorical analyses such as loglinear modeling, release quantity was converted to a discrete variable. Simplification was necessary because the range of the data was very large. For corrosives, the release amount ranged from 0 to 2000 gallons or 0 to 200 pounds, depending on the unit. For flammable liquids, the range was 0 to 4,827.74 gallons. The unit of gallons was much more prevalent than pounds, being associated with more than 39,000 corrosives and flammable liquids incidents. Since only 273 incidents involved pounds as the unit of measure, these records were discarded from the analysis.

A discrete version of release quantity was created based on the data itself as well as expert input. The distributions of the incidents were wide, skewed, and multi-modal, as shown in Figure 7. As an indication of the skewed nature of the data, 91% of the corrosives incidents involved 1 gallon or less, and 99% involved a maximum of 5 gallons, although the maximum amount recorded was 2,000 gallons. For flammable liquids, 90% of the incidents were 1 gallon or less, and over 99% involved 5 gallons or less, compared to a maximum amount of 4,827.74

gallons. The use of discretization techniques such as equal interval or equal frequency binning is problematic with such distributions. When applied to the data given the need to keep the number of bins small, the equal interval technique led to gradations that were too coarse. The incidents were also unevenly distributed among the bins due to being heavily skewed to the left. Equal frequency binning led to the placement of incidents with the same release quantity in different bins due to the skewed nature of the data.



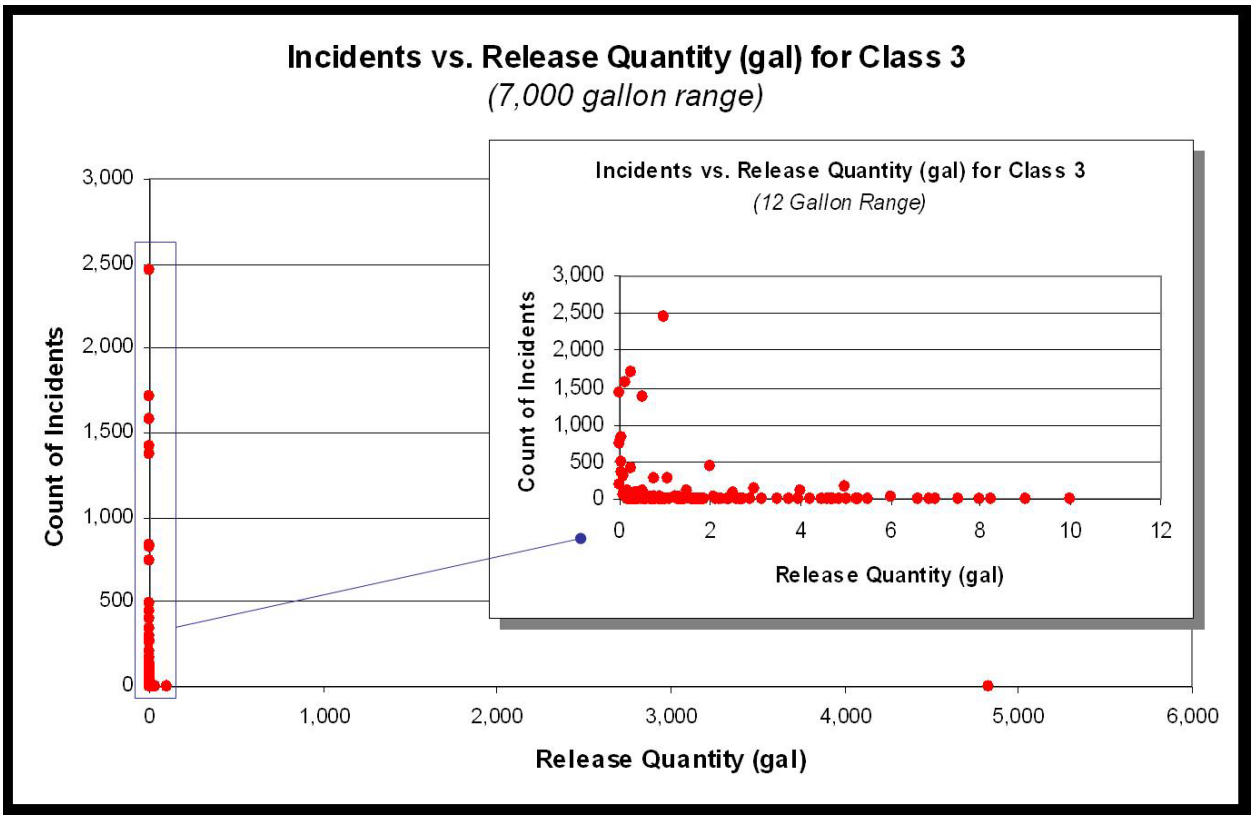


Figure 7: Incidents vs. Release Quantity for Classes 8 and 3.

Despite the large ranges for corrosives and flammable liquids incidents, there were very few releases greater than 100 gallons. Specifically, there were only five releases involving more than 100 gallons in the HMIRS. Therefore, for further simplification, the non-zero range considered by this research was narrowed to 0.01 to 100 gallons. Applying the log (base 10) transform to this range, the range was further narrowed to -2 to 2. As evident, there were two approximately-equal interval bins based on the logarithm, or exponent, as shown below:

- 10^{-2} to 10^{0}
- $10^{.005}$ to 10^{2},

which are equivalent to the following:

- 0.01 to 1 gallon
- 1.01 to 100 gallons.

These equal interval bins very closely coincided with input provided by Doug Reeves of the OHM on appropriate categories for release quantity. Reeves felt that a category for zero was desirable, since an incident may involve no release of material. For example, an incident must be reported if a road closure results, regardless of the amount released. Reeves also felt that carriers tend to report a zero quantity when the amount is too minor to quantify. For a "small" release, Reeves felt that a 1 gallon upper limit was appropriate based on a new policy initiated in January 2005. This policy maintains that a carrier is not required to report a release if it involves fewer than 5 gallons. Finally, since the OHM identifies a "large" or bulk release as 119 gallons or more, Reeves felt that an upper limit of 100 gallons was appropriate for a "medium" release.[182] Based on the small number of releases greater than 100 gallons, a category for "large" was not considered in this study. In summary, the following categories were defined for release quantity for this research:

- 0 gallons
- 0.01 to 1 gallon (small)
- 1.01 to 100 gallons (medium).

Dollar Loss Dollar Loss, which is also a continuous variable as captured on the incident form, was analyzed in a similar fashion to release quantity. Dollar loss is one of several types of consequences associated with a hazardous materials release, and for this research, the total dollar loss associated with the incident was used. The range for dollar loss was large at $0 - $43,760, and the distribution was multimodal and skewed. For example, there were concentrations of incidents at the values $0, $50, $100, $125, and $525, as shown in Figure 8. In addition, 90% of all incidents involved $470 or less, and 99% involved $550 or less. As discussed above, the equal frequency and equal interval binning techniques are difficult to apply when the distribution is wide and skewed.

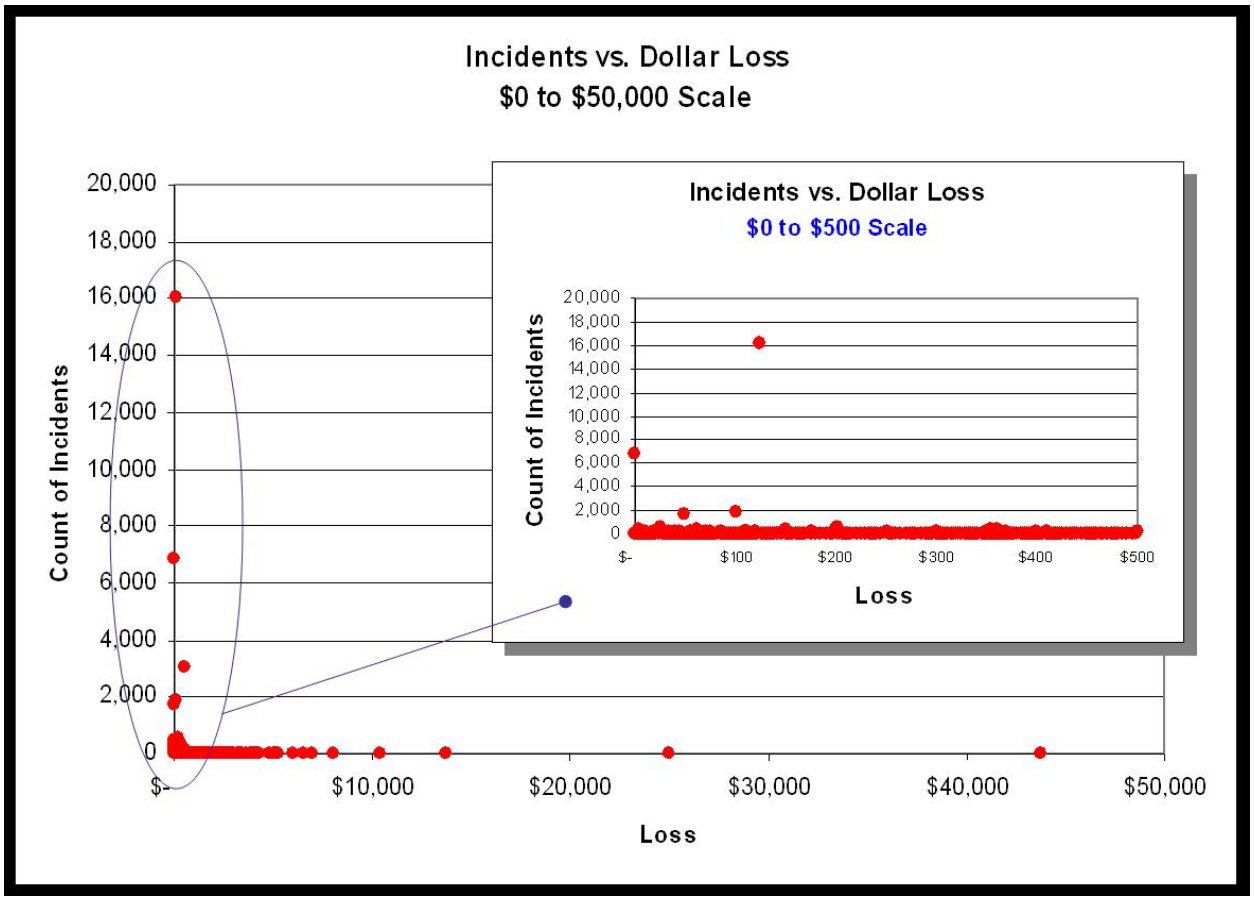


Figure 8: Incidents vs. Dollar Loss.

Expert opinion provided by Doug Reeves was primarily used to identify the categories for dollar loss. Reeves recommended the following categories for zero, small, and medium dollar loss:

- $ 0
- $1 to $500 (small)
- $501 to $25,000 (medium).

As with release quantity, Reeves recommended a separate category for $0. In fact, 17% of the incidents were associated with a zero dollar loss. The "small" category consisting of losses of

$500 or less corresponds to a new DoT guideline as of January 2004. This guideline states that the dollar loss must be reported only if it exceeds $500.[183] Reeves recommended an upper limit of $25,000 for a "medium" loss because any incident above this amount would likely receive a great deal of attention and be considered "large" by the OHM.[184] There were only two incidents in the HMIRS that exceed $25,000. Since very few "large" releases occurred (0.005%), a category for them was not considered by this research.

In an effort to apply the binning techniques, the dollar loss range was narrowed using the logarithmic function, and equal-interval bins based on the logarithm were identified. However, the results did not coincide well with the categories suggested by Reeves. Therefore, the decision was made to place greater weight on the expert's recommendations versus the data-driven categories, since they were in part based on DoT guidelines.

Since the release incidents considered occurred from 1993 to 2002, the dollar loss values were discounted to 2002 dollars for standardization prior to categorizing the incidents based on dollar loss. This was done using annual inflation rates for 1993 through 2002. Dollar loss categorizations were thus based on the 2002-equivalent amounts.

The approach taken in this research for discretizing release quantity and dollar loss, which includes a combination of expert input and data-driven analysis, is summarized in flowchart form in Figure 9.

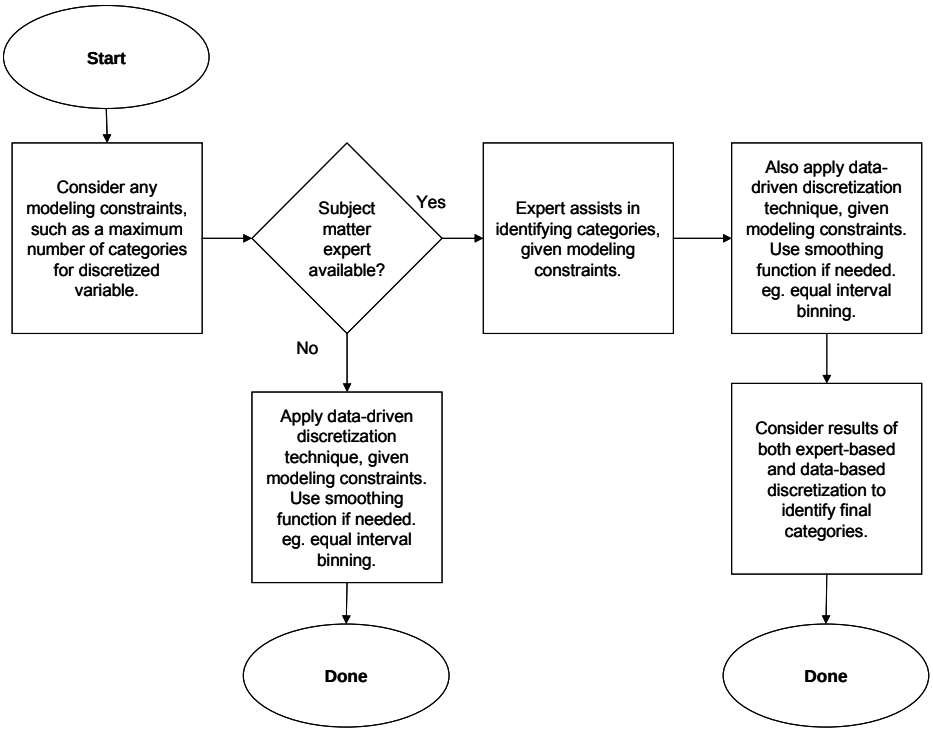


Figure 9: Strategy for Discretization.

4.1.1.4 Latent Variable Development

The last step in the simplification of the variable domain was the development of latent class variables. Five latent class models were developed using the simplified variables and variable sets discussed previously. Several of the variables and sets that were simplified using Pareto analysis and data generalization were used as indicator variables to build simplifying latent variables. In developing the latent variables, an affinity diagramming approach was taken to group the variables believed to be related. For example, geographic division (based on the state), land use, and area type were used as indicators for a latent variable describing the location of the hazmat incident. This approach of building manageable, autonomous models in the form of latent class models is advocated in the literature.[185]

In the case of the container failure variables, the use of latent variables served to simplify the network by replacing many binary variables with a fewer number of latent variables. The latent

variables also served to summarize the failure events. For example, a latent variable called contributing action replaced the top four binary variables that describe the actions that contributed to the failure of the container. The failure item-area latent variable was developed using variables related to both the physical item and area of the container that failed. The four latent variables that were developed to describe the failure of the container served to simplify the network, which otherwise would contain 15 binary variables. The strategy taken for reducing the number of variables in order to simplify the domain that was modeled is shown in Figure 10.

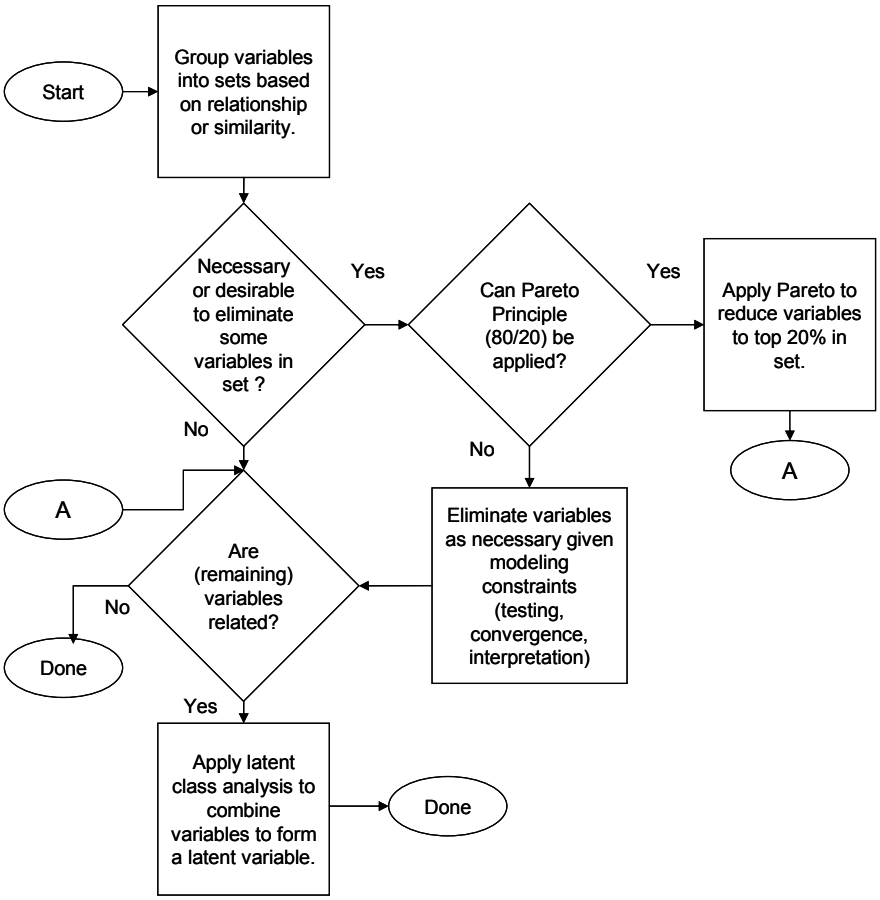


Figure 10: Strategy for Variable Reduction.

In addition to simplifying the domain, the latent variables also resolved cyclic relationships among the indicator variables. This is necessary when developing influence diagrams or

Bayesian networks. For example, when constructing a Bayesian network, any mutual, or two-way, associations among the variables must be converted to one-way relationships. This process can be subjective, especially in cases where there is no clear temporal ordering among the variables.[186] For example, there are mutual associations among the binary indicator variables for contributing action. Since these variables are not ordered in time, determining the direction of influence between them would be subjective.

Due to the availability of data, a large sample size was used to develop each of the latent class models.[187] Measures and indices for determining goodness of fit for large sample sizes were employed, since significance testing typically results in rejection of models with large *N*. These indices include the Index of Dissimilarity (I_d), the Normed Fit Index (*NFI*), and the information criterion measures Bayesian Information Criterion (*BIC*), Akaike Information Criterion (*AIC*), and the Consistent Akaike Information Criterion (*CAIC*).

Location A latent class model describing the location of the incident in terms of its area type, land use, and geographic division was developed. These three variables served as the indicator variables for the latent variable and have the categories shown in Table 16.

Table 16: Categories for Geographic Division, Land Use, and Area Type.

Geographic Division	Land Use	Area Type
Northeast	Industrial	Urban
Middle Atlantic	Commercial	Suburban
East North Central		
West North Central		
South Atlantic		
East South Central		
West South Central		
Mountain		
Pacific		

The latent class analysis for location using these three variables as indicators resulted in a two-class model. This model was chosen based on considering its fit and predictive ability relative to

stable, optimal models with one and three classes. This information is provided in Table 17. As classes were added, the L^2, I_d, *BIC, AIC,* and *CAIC* decreased, and the Normed Fit Index (*NFI)* increased, as shown in the table. The stability of a given model, which contains a certain number of classes, is determined by comparing parameters in two different runs that have the same minimum (optimal) L^2.

Table 17: Measures for Location Models.

Classes	I_d	L^2	BIC	AIC	CAIC	NFI (%)	Classification Error	PRE	DF	Runs
1	0.111	6,255.02	5,977.65	6,205.02	5,952.65	0	0.00	1.00	25	6000
2	0.072	1,999.24	1,843.92	1,971.24	1,829.92	68	0.0009	0.9981	14	6000
3	0.037	809.23	775.95	803.23	772.95	87	0.2315	0.5868	3	6000

Since the sample size was large, L^2 was not used to test goodness of fit. Rather, the I_d, *NFI,* and information criterion values were used to assess the fit of the models. The I_d for the two-class model was close to the target value of 0.05, and its information criterion indices (*BIC, AIC, CAIC*) were smaller relative to the one-class model, indicating the desirability of more than one class. In addition, the predictive ability of the two-class model was very good relative to the three-class model, as shown by the classification error. Since the latent variables were to be used as part of a 3-step modeling approach, considerable weight was placed on a model's classification performance so as to minimize bias in the structural model. Although the I_d of the two-class model was slightly above the target value, this model was chosen based on its lower classification error of 0.0009 versus 0.2315 for the three-class model. Thus, based on its fit, classification ability, and parsimony, the two-class model was chosen.

The parameters of the two class model are given in Table 18. These include the latent class probabilities of 51% and 49%, which indicate the sizes of the classes. The conditional

probabilities are also given in Table 18. For example, the conditional probability that the area type is suburban in the first latent class is 99.9%, indicating a strong association of the first latent class with a suburban area type.

Table 18: Parameters of Location Model.

Class	1	2
Class Size	0.51	0.49
Manifest Variables		
AREA TYPE		
Urban	0.0010	0.9991
Suburban	0.9990	0.0009
LAND USE		
Industrial	0.3754	0.5404
Commercial	0.6246	0.4596
DIVISION		
New England	0.0590	0.0280
Middle Atlantic	0.1574	0.0941
East North Central	0.2691	0.1937
West North Central	0.0795	0.0960
South Atlantic	0.1462	0.1488
East South Central	0.0487	0.0924
West South Central	0.0847	0.1277
Mountain	0.0592	0.0818
Pacific	0.0963	0.1376

A class is interpreted by examining its conditional probabilities. There are no standards for naming latent classes, and the process is subjective on the part of the model builder. However, naming or interpreting the latent classes should reflect how the classes differ from one another. In addition, names should be based on the conditional probabilities that provide the greatest differentiation of the classes.[188] The first class represents suburban locations, the majority of which are commercial. It favors the eastern portion of the U.S., including both central and coastal states. Specifically, the East North Central, Middle Atlantic and South Atlantic divisions

are most prevalent. An example of a likely class one location is Monroeville, PA. In contrast, class two represents urban settings that can be either industrial or commercial. Class two differs from class one in that both eastern and western divisions are prevalent. Specifically, the East North Central, South Atlantic, Pacific, and West South Central divisions are most prevalent. Two likely examples are Norfolk, VA or Detroit, MI. This interpretation of the location variable is summarized in Table 19.

Table 19: Interpretation of Location Model.

Class	1	2
Class Size	0.51	0.49
Area Type	Suburban	Urban
Land Use	Commercial favor	Industrial or Commercial
Geographic Division	Eastern favor (ENC, MA & SA)	Eastern or Western (ENC, SA, PAC & WSC)

Contributing Action The binary variables that comprise the set contributing action describe the actions that contributed to the failure of the container, such as improper loading or dropped. The binary variables in this set, which are grouped in a specific section on the incident form, were used to create a latent variable for contributing action. The indicator variables were as follows:

- Other
- Loose Fitting or Valve
- Improper Loading
- Dropped.

The latent class analysis of these indicator variables resulted in a three-class model for contributing action, with class sizes 46%, 35%, and 19%, as shown in Table 20. Based on the

conditional probabilities, class two represents some "other" contributing action not listed on the incident form. The third latent class corresponds to loose fitting or valve. The first class differs from the others in that it represents a combination of improper loading and dropped, which work in combination to define a type of contributing action.

Table 20: Parameters of Contributing Action Model.

Class	1	2	3
Class Size	0.46	0.35	0.19
Manifest Variables			
LOOSE FITTING OR VALVE			
Y	0.0012	0.0177	0.9999
N	0.9988	0.9823	0.0001
DROPPED			
Y	0.2827	0.0117	0.0029
N	0.7173	0.9883	0.9971
IMPROPER LOADING			
Y	0.3519	0.0096	0.0084
N	0.6481	0.9904	0.9916
OTHER			
Y	0.0026	1.0000	0.0008
N	0.9974	0.0000	0.9992

A summary of the interpretation of the contributing action latent variable is given below in Table 21.

Table 21: Interpretation of Contributing Action Model.

	Class		
	1	**2**	**3**
Class Size	0.46	0.35	0.19
	Improper Loading and Dropped	Other	Loose Fitting or Valve

The three-class model was chosen based on its fit and excellent classification error (P_e=0.0018), as shown in Table 22. Its I_d was close to the target value of 0.05, its *NFI* was above the 80%

threshold, and its information criterion values were lower than those of the one and two class models. The measures for competing models with one and two classes are given in Table 22. A four-class model was not feasible due to negative degrees of freedom.

Table 22: Measures for Contributing Action Models.

Classes	I_d	L^2	BIC	AIC	CAIC	NFI (%)	Classification Error	PRE	DF	runs
1	0.380	84,201.60	84,072.68	84,179.60	84,061.68	0.00	0.0000	1.0000	11	6000
2	0.199	39,805.04	39,734.72	39,793.04	39,728.72	52.73	0.0010	0.9970	6	6000
3	0.086	13,705.21	13,693.49	13,703.21	13,692.49	83.72	0.0018	0.9966	1	6000

Causing Object A latent variable for causing object, which describes the objects that caused the failure of the container, was developed using four binary variables found in the section "Object Causing the Failure" on the incident form. By way of review, these binary variables are as follows:

- None
- Other
- Floor/Ground
- Water/Liquid.

The model chosen for causing object was a three-class model. The three-class model was chosen based on its I_d, which is close to the target value of 0.05, as well as the *NFI* of 83.6%. Its information criteria values were also less than those of the one and two-class models. In addition, it had excellent predictive ability, with a classification error $P_e = 0.0006$ and proportional reduction of error $PRE = 0.9987$. The four-class model, which had slightly better fit, was not feasible due to its negative degrees of freedom. The measures of competing models for causing object with various classes are shown below in Table 23.

Table 23: Measures for Causing Object Models.

Classes	I_d	L^2	BIC	AIC	CAIC	NFI(%)	Classification Error	PRE	DF	runs
1	0.361	81,896.78	81,767.86	81,874.78	81,756.86	0.00	0	1	11	6000
2	0.206	41,482.02	41,411.70	41,470.02	41,405.70	49.35	0.0005	0.9981	6	6000
3	0.079	13,451.70	13,439.98	13,449.70	13,438.98	83.57	0.0006	0.9987	1	6000

The parameters for the three-class model are shown in Table 24. These include the latent class probabilities of 50%, 26%, and 24% as well as the various conditional probabilities indicating the association of the indicator variables to each latent class. For example, class two is heavily associated with "none," or no causing object.

Table 24: Parameters of Causing Object Model.

Class	1	2	3
Class Size	0.50	0.26	0.24
Manifest Variables			
WATER/ LIQUID			
Y	0.2736	0.0019	0.0029
N	0.7264	0.9981	0.9971
FLOOR / GROUND			
Y	0.2890	0.0000	0.0051
N	0.7110	1.0000	0.9949
NONE			
Y	0.0010	0.9999	0.0079
N	0.9990	0.0001	0.9921
OTHER			
Y	0.0003	0.0005	0.9999
N	0.9997	0.9995	0.0001

The first class or type of causing object is a combination of the floor and water (or other liquid). Classes two and three differ in nature from class one. They are characterized as "none" and

"other," respectively, as opposed to a combination of variables. Unfortunately, classes two and three provide limited information on the causing object yet together account for 50% of the cases.

Table 25: Interpretation of Causing Object Model.

	Class		
	1	**2**	**3**
Class Size	0.50	0.26	0.24
Class Description	Floor and Water/ Liquid	None	Other

Failure Mode Three indicator variables related to the manner of container failure were used to create a latent variable for Failure Mode. These indicator variables are as follows:

- Other
- Punctured
- Crushed.

The model chosen for failure mode was a two class model, and its parameters are shown in Table 26. The first class, having a probability of 59.5%, corresponds to a failure mode of "other." However, the interpretation of the second class differs from the first in that it corresponds to a combination of punctured and crushed.

Table 26: Parameters of Failure Mode Model.

Class	1	2
Class Size	0.595	0.405
Manifest Variables		
OTHER		
Y	1.0000	0.0020
N	0.0000	0.9980
PUNTURED		
Y	0.0024	0.3264
N	0.9976	0.6736
CRUSHED		
Y	0.0024	0.2467
N	0.9976	0.7533

The interpretation for failure mode is summarized in Table 27.

Table 27: Interpretation of Failure Mode Model.

	Class	
	1	**2**
Class Size	0.595	0.405
Class Description	Other	Punctured and Crushed

The two-class model for failure mode was chosen based on its fit and classification performance relative to a one-class model, as shown in Table 28. Its I_d was close to the target value of 0.05, and its *NFI* exceeded 80%. Its classification ability was excellent (P_e=0.0007, λ=0.9982). A model with three classes was not possible due to negative degrees of freedom, as apparent from Table 28.

Table 28: Measures for Failure Mode Models.

Classes	**I_d**	**L^2**	**BIC**	**AIC**	**CAIC**	**NFI(%)**	**Classification Error**	**PRE**	**DF**	**runs**
1	0.270	61,915.58	61,868.70	61,907.58	61,864.70	0.00	0.0000	1.0000	4	6000
2	0.059	8,462.42	8,462.42	8,462.42	8,462.42	86.33	0.0007	0.9982	0	6000

Failure Item-Area Two sets of binary variables pertaining to the physical aspects of the container were combined to develop a latent variable to describe the failed item and area of the container. The indicators for this latent variable, which include the failed item and area, are as follows:

- Basic Package Material
- Closure
- Top
- Bottom.

The model chosen for failure item-area was a three-class model, having the parameters shown in Table 29. The largest latent class, which has a probability of 49%, can be characterized as the top of the basic package material. The second latent class, with a size of 27%, corresponds to the bottom of the basic package material. The third class identifies closures on the top of the container as a possible item-area of failure. These interpretations are summarized in Table 30.

Table 29: Parameters of Failure Item-Area Model.

Class	1	2	3
Class Size	0.49	0.27	0.24
Manifest Variables			
BASIC PACKAGE MATERIAL			
Y	0.4187	0.8355	0.0160
N	0.5813	0.1645	0.9840
CLOSURE			
Y	0.0008	0.0161	0.9999
N	0.9992	0.9839	0.0001
TOP			
Y	0.3230	0.0334	0.9445
N	0.6770	0.9666	0.0555
BOTTOM			
Y	0.0012	0.9999	0.0027
N	0.9988	0.0001	0.9973

Table 30: Interpretation of Failure Item Model.

	Class		
	1	**2**	**3**
Class Size	0.49	0.27	0.24
Class Description	Basic Package Material on Top of Container	Basic Package Material on Bottom of Container	Closure on Top of Container

As in previous models, the three-class model was chosen based on its I_d and *NFI*, both of which exceeded their target values, and its excellent classification performance (P_e=0.0012).

Table 31: Measures for Failure Item-Area Models.

Classes	I_d	L^2	BIC	AIC	CAIC	NFI(%)	Classification Error	PRE	DF	runs
1	0.366	137,362.93	137,234.01	137,340.93	137,223.01	0.00	0.0000	1.0000	11	6000
2	0.191	28,939.01	28,868.70	28,927.01	28,862.70	78.93	0.0354	0.8770	2	6000
3	0.041	2,954.08	2,942.36	2,952.08	2,941.36	97.85	0.0012	0.9976	1	6000

4.1.2 Associations

The next task in construction of the decision model was a determination of the associations among the variables, which were unknown. There was no previous exploratory analysis in the literature, and the possibility for non-obvious relationships was a concern. For example, could shift and material type be associated? Therefore, to create a network model depicting accurate associations among the variables, an exploratory analysis was pursued. Using the large amount of available data, the direct associations between the variables were measured using a series of loglinear models, which proceeded from left to right in the network. The following discussion begins with the establishment of the base structure of the network, which takes the temporal ordering among the variables into account. Based on this, five distinct stages of a release event

were identified. Next, taking an overall perspective, the systematic analysis of the network is described. Finally, the methodology and results of the modeling of each stage are presented, including the direct associations identified.

4.1.2.1 Temporal Layout of Network

After simplifying the variable domain through Pareto analysis, data aggregation and discretization, and latent class analysis, a network containing these simplified variables was constructed. By way of review, these simplified variables are shown in Table 32:

Table 32: Simplified Variables in the Hazmat Release Network.

Variable	Number of Categories
Causing Object	3
Container Type	2
Contributing Action	3
Dollar Loss	3
Failure Item-Area	3
Failure Mode	2
Location	2
Material Type	2
Release Quantity	3
Season	4
Shift	3

As a first step, the temporal ordering among the variables was considered. Therefore, the variables were positioned left to right according to their time of occurrence or determination. This technique for network construction is suggested in the literature.[189, 190] As a result of positioning the variables, five distinct stages of a hazardous materials release emerged. Specifically, these five stages and their constituent variables are given in Table 33.

Table 33: Stages of a Hazmat Release.

Stage	Name of Stage	Constituent Variables
1	Pre container-failure initiation	Container Type, Material Type, Location, Season, Shift
2	Container failure initiation	Contributing Action, Causing Object
3	Container failure	Failure Item-Area, Failure Mode
4	Hazmat release	Release Quantity
5	Consequences	Dollar Loss

These stages approximately coincide with Elisabeth Pate-Cornell's System-Action-Management (SAM) framework for catastrophic accidents. In the SAM Framework, there are time-ordered stages leading up to an accident. Specifically, management and organizational factors influence human decisions and actions, which influence the failure or accident events.[191] Roughly speaking, hazmat stage one corresponds to management and organizational factors, such as shift and container type. Stages two and three correspond to human decisions and actions, including the contributing action and failure mode. Stages four and five, which contain the release quantity and dollar loss, correspond to an accident event. The base structure of the network containing these time-ordered stages is shown below in Figure 11.

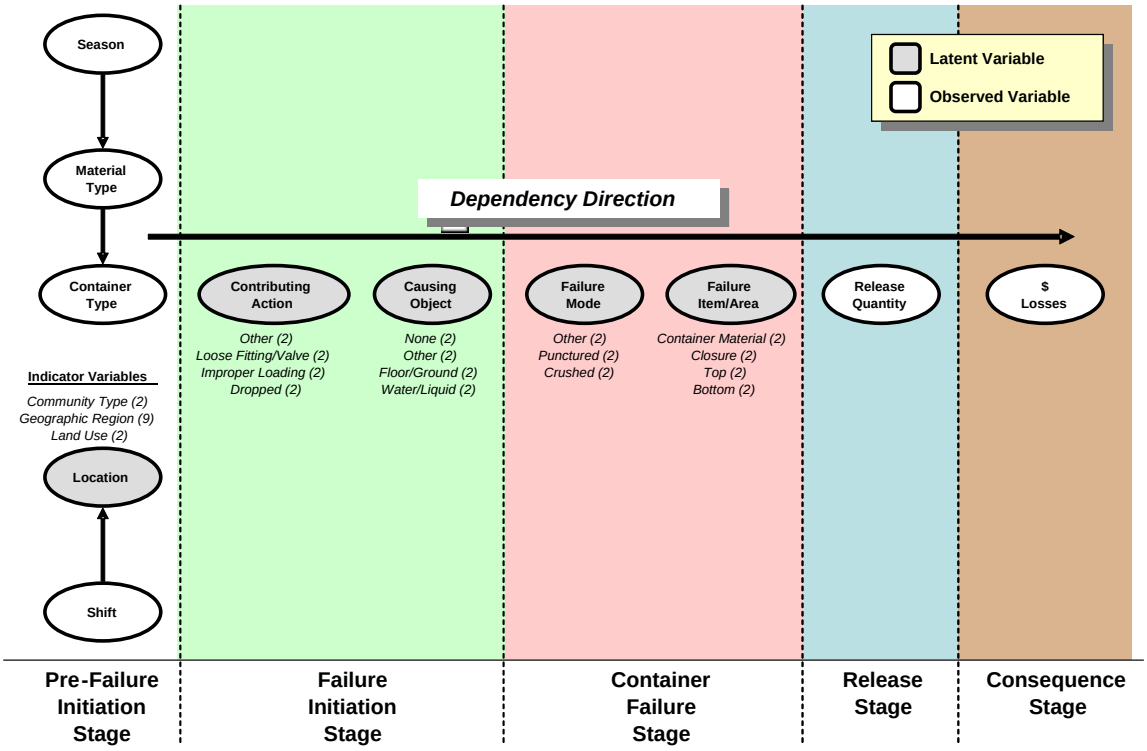


Figure 11: Timed-Ordered Stages of a Hazardous Materials Release.

4.1.2.2 Systematic Analysis of Network

The loglinear analyses for assessing associations and independencies proceeded from left to right in the network, starting with a symmetrical analysis of the variables in stage one, the pre-failure initiation stage. These variables consist of container type, location, material type, season, and shift and are known or determined prior to unloading of containers or the initiation of failure. A symmetric loglinear analysis was performed for this stage since a direction of influence among the variables was not known or assumed. When the stage one analysis was completed, a temporal order was then applied to the pairs of variables found to be associated. In this way, the

mutual, or bi-directional, associations were converted to unidirectional associations, which are necessary for influence diagrams or Bayesian networks. The following direct associations were uncovered among the stage one variables:

- Material Type and Container Type
- Season and Material Type
- Shift and Location.

A panel of engineers and scientists at the Office of Hazardous Materials provided some interpretations of these associations. Material type and container type are likely directly associated because regulations dictate the type of container for transporting a particular type of material. The seasonal usage of materials is an explanation for the direct association between season and material.[192]

Given these three direct associations, a temporal order was applied to each pair, as discussed above. Season, which is based on the incident date, can be considered a predetermined or general variable of a fundamental nature. Several authors of loglinear modeling texts, including Knoke, Burke, and Hagenaars, identify the concept of a predetermined or fundamental variable and recommend its use as the preceding variable in a causal chain.[193, 194] Based on this, season was assumed to precede material type. It was then determined that material type influences and therefore should precede the container type, based on input from the OHM.[195] Finally, shift, which is based on the incident time, was assumed to be more fundamental than the location. Therefore, shift was assumed to precede location in the model.

Proceeding to stage two, container failure initiation is characterized by the latent variables contributing action and causing object. In this stage, failure is initiated by a combination of actions and objects that contribute to or cause the failure. The indicator variables for

contributing action are as follows: dropped, loose fitting/valve, improper loading, and other. For causing object, they are floor/ground, water/liquid, none, and other. It was assumed that contributing action precedes causing object, based on these indicator variables. For example, an improperly loaded container may drop and impact the ground and encounter water or other liquid.

Since stage two is downstream in the network relative to stage one, the stage two variables served as response, or logit, variables.[196] Therefore, an asymmetric loglinear analysis was performed, with the stage one variables serving as explanatory variables. A loglinear analysis was first performed between contributing action, the first variable in stage two, and the variables in stage one. Then, a second loglinear analysis was performed between causing object, the second variable in stage two, and its explanatory variables. These explanatory variables consist of contributing action and the stage one variables. This forward analysis of the network, each time utilizing a new logit variable further downstream in the chain, is suggested by Agresti, and Knoke and Burke.[197, 198] Thus, the associations and conditional independencies among the variables are determined using a forward series of loglinear models.

Within stage three, where container failure occurs, the failure mode was assumed to precede the item-area of failure on the container. For example, the container might be punctured, leading to a failure of the bottom of the basic package material. The indicators for failure mode are punctured, crushed, and other. For failure item-area, they are as follows: basic package material, closure, top, and bottom. The variables in stages one and two served as explanatory variables for the variables in stage three. An asymmetric loglinear analysis was first performed between

failure mode and the variables in stages one and two. Moving forward, an analysis was then performed between item-area and its explanatory variables, failure mode and the variables in stages one and two.

At the next stage, the release of hazardous material to the surrounding area occurs. The quantity of material released is represented by a discrete variable with categories zero, small, and medium, which have upper limits of 0, 1, and 100 gallons, respectively. An asymmetric loglinear analysis was performed between release quantity and the variables of the preceding stages. In this way, the associations and conditional independencies between release quantity and each of its explanatory variables were determined.

Finally, in stage five, the ultimate consequences of a release are realized. The consequence considered in this study was the total dollar loss, which is represented by a discrete variable with categories zero, small, and medium with upper limits of $0, $500, and $25,000, respectively. Dollar loss served as the ultimate logit variable for the network. Therefore, in this final loglinear analysis, the variables in the first four stages served as explanatory variables.

The high-level approach described in the previous sections for measuring the associations as part of step two of the methodology is summarized in Figure 12. This methodology is demonstrated in the following sections, and the results for each stage are presented in detail.

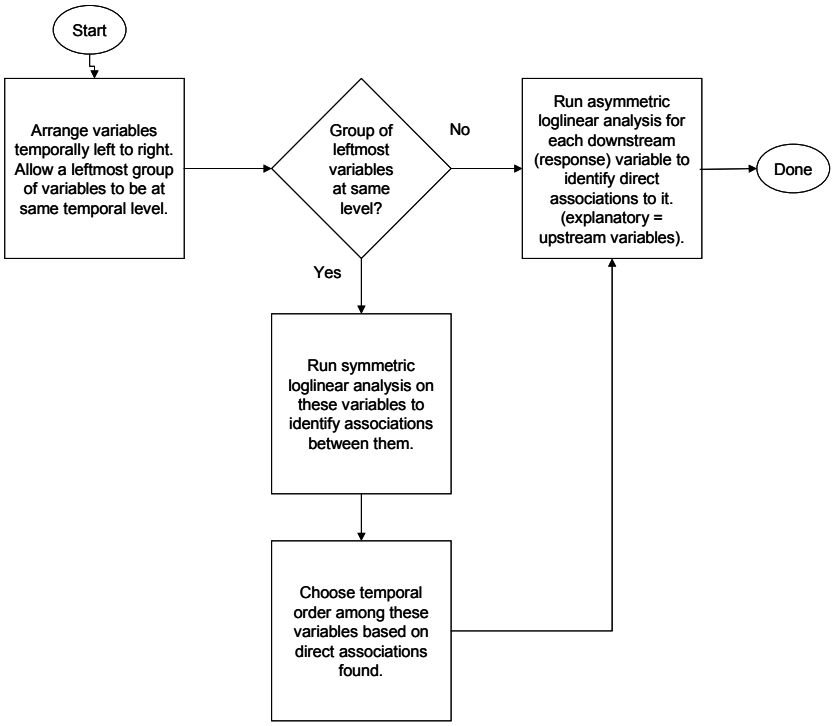


Figure 12: High Level Approach to the Measurement of Associations.

Stage One The exploratory analysis in stage one consisted of a symmetric loglinear analysis of the variables given in Table 34. These variables were determined or known prior to unloading of the containers and before any initiation of failure. The abbreviation used in the loglinear notation and number of categories for each variable are also provided below.

Table 34: Stage One Variables.

Stage 1 Variable	Variable Abbreviation	Number of Categories
Container Type	C	2
Location	L	2
Material Type	M	2
Season	SE	4
Shift	SH	3

The contingency table for these five variables consisted of 96 cells, and the sample size was primarily determined by examining the number of cells with zeros and small counts of less than five. The zeros and small cell counts associated with N=1500 and N=2500 based on five random samples are given below.

N	Sampling Zeroes	Small cell count
1500	0%	13.75%
2500	0%	2.92%

The sample size of N=2500 was chosen based on the low percentage of cells having a small count. Due to the availability of data, five random samples of N=2500 each were used for the significance testing. Therefore, conclusions about the associations among the stage one variables were based on results from five different samples.

Prior to beginning this analysis, an additional preparatory step was taken. This consisted of applying the correction procedure developed by Bolck et. al., as discussed previously in section 3.6.2. In stage one, the correction procedure for a table with one latent variable was applied using Equation 10. However, since the classification error associated with the latent variable location was small (P_e=0.0009), the correction procedure had a very small effect, resulting in no differences between the uncorrected and the corrected table for each sample.[199, 200] Despite its lack of impact in this case, the correction procedure was nonetheless applied for purposes of demonstration. For, future application of the methodology may entail variables with higher classification error. Sample transition, uncorrected, and corrected matrices based on a sample in stage one are given in Table 79 in Appendix A. Each row of the uncorrected and corrected matrices represents a different combination of the categories of the observed variables. Each column represents a different class of the latent variable. The source code written to calculate

the transition matrix for location is provided in Figure 24 in Appendix E. In addition, the Excel spreadsheets used as input to this source code are provided in Table 91 and Table 92 in Appendix E. As a note for future application of the correction procedure, an error was made within this research in calculating the elements of the transition matrices for the various latent variables. However, this error had a very small, if not negligible, impact on the corrected contingency tables. The error was made in the choice of the particular classification probabilities to use to calculate the elements of the transition matrix. Only two possible values for the classification probability should have been used, namely 1 or 0. These probabilities indicate whether or not a response pattern was classified into a particular latent class. The classification probabilities that were erroneously used were the actual calculated values. However, most of the actual calculated classification probabilities were close to 0 or 1 anyway, as evidenced in the low classification errors associated with the latent variables. Thus, the impact of this error was very small on both the correction procedure and the overall loglinear analyses.

After the correction procedure was applied, loglinear models using each of the five corrected tables were built. A test of marginal association was used to determine the presence of a significant association between a given pair of variables.[201] In this test, the model of mutual independence of the five variables in stage one is compared to a model that differs by the presence of a two-way term. In stage one, ten different variable pairs were evaluated using ten marginal association tests. Each pair and the *p-value* associated with each marginal test are given in Table 35. For further reference, the values of the residual and component L^2 for each test are given in Table 78 in Appendix A. Three pairs were assessed using two additional samples due to borderline values for the component L^2 and associated *p-values*. The additional samples are reported in columns S6 and S7 in Table 35.

Table 35: P-values for Marginal Associations in Stage One.

Variable Pair	S1	S2	S3	S4	S5	S6	S7	AVG
[SH L]	0.0241	0.0064	0.0225	0.0790	0.0003			0.0265
[L SE]	0.5707	0.4588	0.5709	0.6120	0.1044			0.4634
[L M]	0.5999	0.4934	0.0186	0.3286	0.0049	0.1287	0.6244	0.3141
[SH SE]	0.1734	0.2886	0.0247	0.0766	0.8431	0.5096	0.8295	0.3922
[SH M]	0.1135	0.5860	0.3460	0.8797	0.1236	0.0966	0.5368	0.3832
[SE M]	0.0040	0.0219	0.0040	0.0006	0.0177			0.0096
[L C]	0.3527	0.5167	0.5486	0.5906	0.7989			0.5615
[SH C]	0.9176	0.9112	0.9731	0.8782	0.9559			0.9272
[SE C]	0.3472	0.2962	0.8836	0.3044	0.2619			0.4187
[M C]	0.0001	0.0001	0.0001	0.0009	0.0001			0.0003

Based on these results, *[M C], [SE M],* and *[SH L]* were the significant associations at α=0.05. The association of material type and container type, *[M C]*, was characterized by consistently small *p-values*. In four samples, p = 0.0001, and for one sample, p = 0.0009. The association between season and material type, *[SE M],* was also consistently strong, with *p-values* between 0.0006 and 0.0219. The final significant association was between shift and location, *[SH L],* as indicated by *p-values* between 0.0003 and 0.0790.

Three of the pairs had borderline values for the component L^2 and associated *p-values*, based on five samples. These associations were *[L M], [SH SE],* and *[SH M]*. Therefore, two additional random samples were used to draw a conclusion about them. Based on these additional samples, however, these three associations were determined to be non-significant, as shown in Table 35 in columns S6 and S7. Therefore, material type and container type *[M C],* season and material type *[SE M],* and shift and location *[SH L]* were identified as the only direct associations in stage one. Since the associations were to be used to build a directed graph, only two-way, as opposed to three or four-way, associations were considered.

The lambda (λ), or effect, parameters provided similar evidence that *[M C], [SE M],* and *[SH L]* were directly associated. A direct association between a pair of variables is present when any

of its lambda values exceeds 0.20 in absolute value, as discussed previously in section 3.5.3. The lambda of maximum absolute value for each variable pair, which represents the largest or strongest effect associated with the pair, is given in Table 36, along with the sign. The sign indicates the direction of influence of the effect, as discussed in 3.5.3. In many cases, there were equally-strong effects in opposite directions for a variable pair, as indicated by the positive/negative (±) signs in this table. The lambdas were based on the full first-order model, which contained all two-variable associations.[202] For the three associations, *[M C], [SE M],* and *[SH L],* which were consistently significant, the maximum lambda for the variable pair was also consistently above 0.20 in absolute value. For the associations [*L SE*], [*SE C*], [*SH M*], and [*SH SE*], several of the maximum lambdas exceeded 0.20 in absolute value. However, since the significance tests for these associations were inconsistent and sometimes notably in favor of a non-significant relationship, these variable pairs were determined not to be directly associated. For example, for the association between shift and season [*SH SE*], the *p-values* in Table 35 for samples S5-S7 were very large.

Table 36: Lambdas of Max Absolute Value in Stage One.

Variable Pair	S1		S2		S3		S4		S5		S6		S7	
[M C]	±	0.4070	±	0.4330	±	0.3713	±	0.3123	±	0.4805				
[SE M]	±	0.2413	±	0.5101	±	0.6271	±	0.2380	±	0.5243				
[SH L]	±	0.3029	±	0.3543	±	0.3781	±	0.3868	±	0.6940				
[L C]	±	0.0955	±	0.0542	±	0.0422	±	0.0546	±	0.0568				
[L M]	±	0.0573	±	0.0487	±	0.1858	±	0.0817	±	0.2394	±	0.1269	±	0.0389
[L SE]	±	0.3749	±	0.1516	±	0.1808	±	0.3175	±	0.6517				
[SE C]	±	0.1125	±	0.4527	±	0.0531	±	0.2258	±	0.2586				
[SH C]	±	0.1016	±	0.0864	±	0.0431	±	0.0410	±	0.0702				
[SH M]	±	0.3598	±	0.0704	±	0.0840	±	0.0706	±	0.3619	±	0.2886	±	0.1614
[SH SE]	+	0.3765	-	0.6808	-	2.1407	-	0.9240	+	0.7006	-	0.3484	-	0.1723

The lambda parameters provide insight into the categories that were responsible for the largest effects. The categories associated with the largest positive and negative effects, or lambdas, for *[M C], [SE M],* and *[SH L]* are provided in Table 37. These category combinations had the largest influence on incident counts for the three variables pairs. As shown in Table 37, the combination of flammable liquids in an (outer) fiber box as well as corrosives in a bottle had the (equally) strongest positive effects for material type and container type [*M C*]. Thus, these two combinations were associated with relatively more incidents. For season and material type *[SE M],* there were relatively more incidents involving flammable liquids during the winter based on 3/5 samples. Conversely, there were relatively fewer corrosives incidents during the winter months based on 3/5 samples. Also, based on the maximum positive lambda for shift and location [*SH L*] in 5/5 samples, there were relatively more incidents that occurred during the twilight shift in suburban/commercial/eastern locations in the United States.

Table 37: Interpretation of Largest Effects in Stage One.

Variable Pair	Largest Positive Effect	Based on Samples	Largest Negative Effect	Based on Samples
[M C]	- Flammable liquids in fiber box - Corrosives in bottle	5/5	- Flammable liquids in bottle - Corrosives in fiber box	5/5
[SE M]	Flammable liquids during winter	3/5	Corrosives in winter	3/5
[SH L]	Twilight and suburban/ commercial/ eastern	5/5	Twilight and urban/ industrial or commercial/ eastern or western	5/5

The three direct associations uncovered in stage one, which consist of season to material type, material type to container type, and shift to location, are summarized graphically in Figure 13.

Figure 13: Graphical Results of Stage One Analysis.

As apparent, season directly influences material type, which directly influences container type. Shift directly influences location.

Stage Two Moving downstream in the network, the associations between stages one and two were investigated next. Stage two consists of catalyst-type variables that describe the initiation of container failure, specifically the contributing action and causing object. Contributing action was assumed to precede causing object based on their categories, as shown in Table 38. For example, a container was improperly loaded and therefore dropped and impacted the floor. Thus, contributing action was a response variable relative to the stage one variables. In turn, causing object was a response variable relative to contributing action and the stage one variables.

Table 38: Stage Two Variables.

Stage 2 Variable	Abbreviation	Categories
Contributing Action	CA	1) Improper Loading and Dropped 2) Other 3) Loose Fitting or Valve
Causing Object	CO	1) Floor and Water/Liquid 2) None 3) Other

The first logit analysis in stage two had contributing action as the response variable. In addition, there were two groups of explanatory variables for contributing action based on their independence. The first group consisted of season, material type, and container type, and the second group consisted of shift and location. The variables in the first group were independent of those in the second group, based on the direct associations uncovered in stage one, as shown previously in Figure 13. Based on their independence, the first group of variables was analyzed separately from the second group relative to contributing action, as given by the Collapsibility Theorem.[203] Thus, two separate logit analyses for contributing action were done. Application of the Collapsibility Theorem was relevant because it permitted the use of smaller tables and sample sizes, which enabled significance testing.

First Logit Analysis for Contributing Action To perform significance testing in the first logit analysis, the sample size had to be chosen appropriately. Based on the zero and small cell counts associated with various sample sizes as shown below, a sample size of N=4000 was chosen.

N	Sampling Zeroes	Small cell count
3000	0%	5.02%
3500	0%	4.58%
4000	0%	1.66%

Five samples containing 4000 records each were used to investigate the presence of significant associations among the variables season, material type, container type, and contributing action.

Prior to beginning the analysis, the correction procedure was applied to the table for these four variables. However, since the latent variable (contributing action) had low classification error, the correction procedure did not have a large impact. Specifically, across the five samples, the maximum difference between the counts in the uncorrected and corrected tables was six, with a maximum cell difference of one.

Nonetheless, the corrected matrix was used to perform the loglinear, or logit, analysis, the results of which are shown in Table 39. The marginal component L^2 associated with *[C CA]* had a consistently small *p-value* (< 0.0001), indicating a significant association between container type and contributing action. The partial component associated with *[M CA]* was significant at α=0.05 in three out of five samples, as demonstrated by its component L^2 values. As discussed previously in sections 3.5.1 and 3.5.2, the partial association *[M CA]* is the association after adjustment for the effects of container type *C*. Since the partial component was significant at α=0.05 in 3/5 samples and at α=0.10 in 4/5 samples, material type *M* and contributing action *CA* were determined to be directly associated. In other words, *M* and *CA* were *not* conditionally independent given *C*. The partial association between season *SE* and contributing action *CA* as measured by its component L^2 was *not* significant at α=0.05 in four out of five samples. Based on this, season did not appear to be directly associated with contributing action given container and material type. However, as will be discussed, the lambdas (λ) suggested a different result for *SE* and *CA*.

Table 39: Significance Tests for Logit *CA*. (*SE, M, C*)

		RESIDUAL			COMPONENT			
Sample	Model	L^2	Df	p	L^2	df	p	Significant Component
1	[SE M C][CA] (null logit)	1463.5744	30	2.00E-289				
	[SE M C][CA][C CA]	31.3128	28	0.3034	1432.2616	2	<0.0001	Y
	[SE M C][CA][C CA][M CA]	23.9555	26	0.5785	7.3573	2	0.0253	Y
	[SE M C][CA][C CA][M CA][SE CA]	20.8198	20	0.4078	3.1357	6	0.7916	
2	[SE M C][CA] (null logit)	1338.4577	30	1.00E-262				
	[SE M C][CA][C CA]	37.7725	28	0.1028	1300.6852	2	<0.0001	Y
	[SE M C][CA][C CA][M CA]	26.9944	26	0.4096	10.7781	2	0.0046	Y
	[SE M C][CA][C CA][M CA][SE CA]	18.6818	20	0.5426	8.3126	6	0.2161	
3	[SE M C][CA] (null logit)	1461.7462	30	6.00E-289				
	[SE M C][CA][C CA]	31.6474	28	0.2891	1430.0988	2	<0.0001	Y
	[SE M C][CA][C CA][M CA]	30.2672	26	0.2567	1.3802	2	0.5015	
	[SE M C][CA][C CA][M CA][SE CA]	16.4394	20	0.689	13.8278	6	0.0316	Y
4	[SE M C][CA] (null logit)	1403.7126	30	1.00E-276				
	[SE M C][CA][C CA]	33.9977	28	0.2009	1369.7149	2	<0.0001	Y
	[SE M C][CA][C CA][M CA]	25.9030	26	0.4684	8.0947	2	0.0175	Y
	[SE M C][CA][C CA][M CA][SE CA]	18.7515	20	0.5380	7.1515	6	0.3071	
5	[SE M C][CA] (null logit)	1543.7182	30	2.00E-306				
	[SE M C][CA][C CA]	24.7975	28	0.6388	1518.9207	2	<0.0001	Y
	[SE M C][CA][C CA][M CA]	19.5840	26	0.8108	5.2135	2	0.0738	
	[SE M C][CA][C CA][M CA][SE CA]	10.4514	20	0.9592	9.1326	6	0.1663	

In addition to significance testing, the lambdas, or effect parameters, were also examined. The lambdas were based on the full first-order model containing the main effects of *C*, *M*, and *SE* on the logit variable *CA*, as were the lambdas in all subsequent stages in this analysis.[204, 205] The lambdas were therefore (conservative) partial effects, serving as measures of association after adjusting for all other variables. For *[C CA]*, the lambda of maximum absolute value for each sample was consistently large, as shown in Table 40. This coincided with the consistently small *p-values* (< 0.0001) for this association. The lambdas for *[M CA]* were greater than 0.20 in 5/5 samples. This generally coincided with the significant association between material type and

contributing action at α=0.05 in 3/5 samples and at α=0.10 in 4/5 samples. Regarding the third and final association, the maximum lambdas for *[SE CA]* were similar in magnitude to those for *[M CA]* and were greater than 0.20 in absolute value in 5/5 samples. Although season *SE* and contributing action *CA* were not directly associated based on significance testing, the decision was made to assume a direct association between them based on the values of their maximum lambdas and their proximity to the maximum lambdas for [*M CA*].

Table 40: Lambdas of Max Absolute Value for Logit *CA*. (*SE, M, C*)

Variable Pair	S1		S2		S3		S4		S5	
[SE CA]	+	0.3768	+	0.3488	+	0.4806	+	0.3592	-	2.1656
[M CA]	±	0.5287	±	0.3668	±	0.2636	±	0.6953	±	0.3813
[C CA]	±	7.2263	±	6.9781	±	7.4654	±	7.4152	±	7.6799

Interpretations of the strongest effects based on the maximum lambdas for each variable pair are given in Table 41. For example, based on five out of five samples, a higher proportion of corrosives were released as the result of a loose fitting or valve. In conjunction with this, there were proportionally more incidents of loose fittings or valves on bottles based on 5/5 samples.

Table 41: Interpretation of Largest Effects for Logit *CA*. *(SE, M, C)*

Variable Pair	Largest Positive Effect	Based on Samples	Largest Negative Effect	Based on Samples
[SE CA]	Winter and other contributing action	2/5	- Winter and loose fitting or valve - Fall and loose fitting or valve	2/5 2/5
[M CA]	Corrosives and loose fitting or valve	5/5	Flammable liquids and loose fitting or valve	5/5
[C CA]	Bottle and loose fitting or valve	5/5	Fiber box and loose fitting or valve	5/5

The logit analysis of the first group of explanatory variables for contributing action is summarized graphically in Figure 15 and consists of three direct associations to the response.

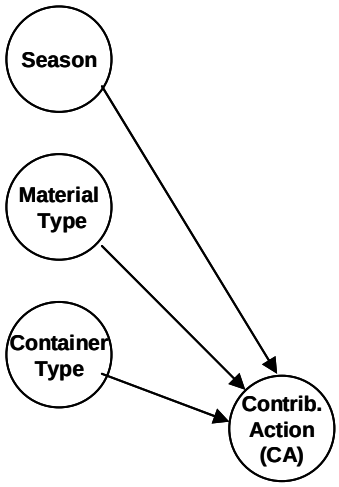


Figure 14: Graphical Results of Logit Analysis for *CA*. (*SE, M, C)*

Second Logit Analysis for Contributing Action The second logit analysis for contributing action had shift and location as the explanatory variables, which were independent of the explanatory variables in the first logit analysis. A sample size of 1200 was chosen based on the absence of zeros or small cell counts in five different samples.

Prior to the loglinear modeling, the correction procedure was not applied because the literature did not address the case of a joint distribution involving two or more latent variables. I contacted the developers of the correction procedure concerning this gap. Dr. Marcel Croon provided a more generally-applicable correction procedure later in time, which will be demonstrated in stage four. Since the classification errors associated with location and contributing action were low (P_e=0.0009 and P_e=0.0018, respectively), the inability to apply a correction at this point in time was not a concern.

Significance testing and evaluation of lambdas were again used to assess the presence of direct associations. The marginal association between location and contributing action *[L CA]* was consistently significant at α=0.05 across the five samples, as indicated by the significant

values of the component L^2 in Table 42. The partial association between shift and contributing action *[SH CA]*, which was obtained after correcting for the main effect of location, was significant in 2/5 samples, as demonstrated by its component L^2 in Table 42.

Table 42: Significance Tests for Logit *CA*. (*SH, L*)

		RESIDUAL			COMPONENT			
Sample	Model	L^2	df	P	L^2	df	P	Significant Component
1	[L SH][CA] (null logit)	36.5235	10	7.00E-05				
	[L SH][CA][L CA]	9.6272	8	0.2922	26.8963	2	<0.0001	Y
	[L SH][CA][L CA][SH CA]	4.6614	4	0.3238	4.9658	4	0.2908	
2	[L SH][CA] (null logit)	32.1485	10	4.00E-04				
	[L SH][CA][L CA]	16.5212	8	0.0355	15.6273	2	0.0004	Y
	[L SH][CA][L CA][SH CA]	2.4770	4	0.6488	14.0442	4	0.0072	Y
3	[L SH][CA] (null logit)	21.5129	10	1.78E-02				
	[L SH][CA][L CA]	15.4588	8	0.0508	6.0541	2	0.0485	Y
	[L SH][CA][L CA][SH CA]	1.3828	4	0.8472	14.0760	4	0.0071	Y
4	[L SH][CA] (null logit)	26.2131	10	3.50E-03				
	[L SH][CA][L CA]	14.0863	8	0.0795	12.1268	2	0.0023	Y
	[L SH][CA][L CA][SH CA]	8.6398	4	0.0708	5.4465	4	0.2445	
5	[L SH][CA] (null logit)	16.9713	10	7.50E-02				
	[L SH][CA][L CA]	7.9825	8	0.4352	8.9888	2	0.0112	Y
	[L SH][CA][L CA][SH CA]	1.5442	4	0.8188	6.4383	4	0.1687	

The significant association between location *L* and contributing action *CA* coincided with the maximum lambdas for [*L CA*], as shown in Table 43, which were all above 0.20 in absolute value. Although the *[SH CA]* partial association was significant at α=0.05 in just two out of five samples, the largest lambdas for *[SH CA]* ranged from 0.6088 to 1.5114 in absolute value. Based on this, the decision was made to assume a direct association between shift and contributing action.

Table 43: Lambdas of Max Absolute Value for Logit *CA*. (*SH, L*)

Variable Pair	S1		S2		S3		S4		S5	
[SH CA]	+	0.7421	-	0.8297	-	0.6088	+	1.5114	-	0.7967
[L CA]	±	0.8838	±	0.5530	±	0.6852	±	0.3235	±	0.3775

As explanations of the largest effects, a greater proportion of incidents on the twilight shift involved loose fittings or valves, based on two of five samples, as shown in Table 44. For the [*L CA*] direct association, proportionally more incidents of loose fittings or valves occurred in urban/industrial or commercial areas, based on 3/5 samples.

Table 44: Interpretation of Largest Effects for Logit *CA*. *(SH, L)*

Variable Pair	Largest Positive Effect	Based on Samples	Largest Negative Effect	Based on Samples
[SH CA]	Twilight and loose fitting or valve	2/5	Midnight and loose fitting or valve	3/5
[L CA]	Urban/ industrial or commercial/ eastern or western and loose fitting or valve	3/5	Suburban/ commercial/ eastern and loose fitting or valve	3/5

The results of the logit analysis for contributing action involving the second group of explanatory variables are graphically shown in Figure 15.

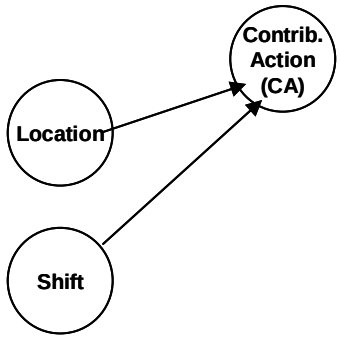


Figure 15: Graphical Results of Logit Analysis for *CA*. (*SH, L)*

Logit Analysis for Causing Object The second variable in stage two and the next response variable to analyze was causing object, which describes the physical objects, such as the floor and water, leading to failure of the container. Therefore, the explanatory variables for this analysis consisted of contributing action, also in stage two, and the variables in stage one. The Collapsibility Theorem no longer applied to this model for causing object or others downstream, due to the associations among the preceding variables. Thus, for this model involving causing object, all variables had to be analyzed as part of one model or table. The logit model for causing object involved seven variables and 864 cells. Due to the sample size needed to create a non-sparse contingency table, significance testing was not feasible. Rather, for this model and those downstream in the network, associations had to be assessed using the lambda, or effect, parameters only, which are insensitive to large sample size and sparseness.[206] The total sample of 40,474 records was used to build the model. The table had 1.4% sampling zeros and 23.1% small cell counts.

The largest lambda parameters in absolute value for this seven-variable model are shown in Table 45, along with interpretations of these largest effects. Contributing action had the largest effect on causing object, with $\lambda = 4.5500$, followed by location ($\lambda = \pm 0.9462$) and container type ($\lambda = \pm 0.8962$). Several of the largest effects on causing object involved the "other" category and therefore provide limited insight at this time into the exact influence on causing object.

Table 45: Lambdas of Max Absolute Value and Interpretation of Largest Effects for Logit *CO*.

Variable Pair	S1	
[CA CO]	+	4.5500
[C CO]	±	0.8962
[M CO]	±	0.3036
[L CO]	±	0.9462
[SE CO]	-	0.2111
[SH CO]	+	0.6801

Table 45 (continued).

Variable Pair	Largest Positive Effect	Largest Negative Effect
[CA CO]	Loose fitting or valve and no causing object	Loose fitting or valve and floor and water
[C CO]	Bottle and no causing object	Fiber box and no causing object
[M CO]	Flammable liquids and other causing object	Corrosives and other causing object
[L CO]	Urban/ industrial or commercial/ eastern or western and other causing object	Suburban/ commercial/ eastern and other causing object
[SE CO]	Winter and other causing object	Summer and other causing object
[SH CO]	Twilight and other causing object	Day and other causing object

The direct associations and conditional independencies are graphically summarized in Figure 16.

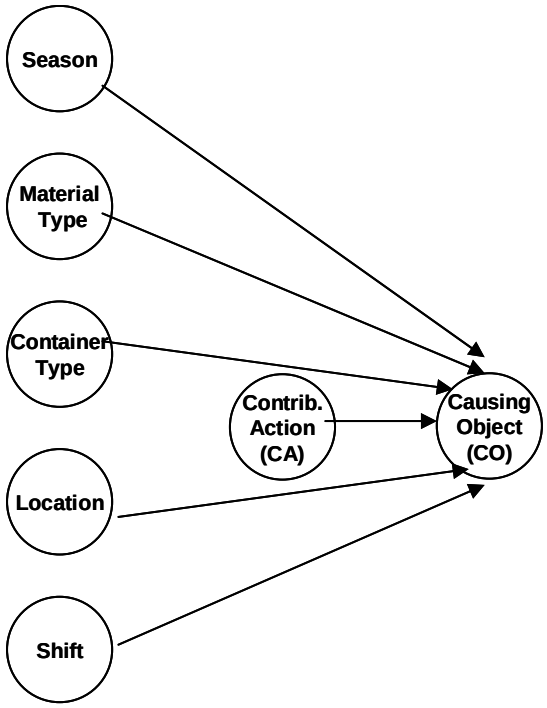


Figure 16: Graphical Results of Logit Analysis for *CO*.

Stage Three Continuing downstream in the network, actual failure of the container occurs in stage three. Stage three is similar to stage two in that it contains two variables and hence two separate logit models. The failure mode describes the manner of failure, and the failure item-

area defines the physical point of failure on the container. Failure mode was assumed to temporally precede the item-area, based on their categories. For example, perhaps the container was punctured, causing the bottom side of the basic package material (e.g. cardboard) to fail. The stage three variables are summarized below.

Table 46: Stage Three Variables.

Stage 3 Variable	Abbreviation	Categories
Failure Mode	FM	1) Other 2) Punctured and Crushed
Failure Item-Area	FIA	1) Basic Package Material on Top 2) Basic Package Material on Bottom 3) Closure on Top

The logit analysis for failure mode consisted of seven preceding, explanatory variables. With a total of 1728 cells, the contingency table had 13.1% zeros and 51.7% small cell counts, using the total sample size of 40,474. Evaluation of the lambda parameters was the method used to determine the direct associations of the explanatory variables with failure mode.

Based on the lambda parameters shown in Table 47, contributing action was again the variable with the strongest effect on the logit variable, with $\lambda = \pm 4.3000$. The positive lambda parameter indicates that a larger proportion of loose fittings or valves were associated with some "other" failure mode not listed on the incident form, as shown in Table 47. Conversely, a smaller proportion of loose fittings or valves led to puncture and crush of the container ($\lambda = -4.3000$). Causing object and container type had the next largest effects on failure mode, with $\lambda = \pm 1.3657$ and $\lambda = \pm 1.3598$, respectively. Location and season were also directly associated with failure mode albeit to lesser extents ($\lambda = \pm 0.2589$ and $\lambda = \pm 0.2548$, respectively). However, shift and material type were not directly associated with failure mode, since their lambdas were

below 0.20 in absolute value. Material type was the explanatory variable least associated with failure mode with $\lambda = \pm 0.0103$. Thus, arcs were not drawn from shift and material type to failure mode in the network.

Table 47: Lambdas of Max Absolute Value and Interpretation of Largest Effects for Logit *FM*.

Variable Pair	S1	
[CO FM]	±	1.3657
[CA FM]	±	4.3000
[C FM]	±	1.3598
[M FM]	±	0.0103
[SE FM]	±	0.2548
[L FM]	±	0.2589
[SH FM]	±	0.1490

Variable Pair	Largest Positive Effect	Largest Negative Effect
[CO FM]	Floor and water and punctured and crushed	Floor and water and other failure mode
[CA FM]	Loose fitting or valve and other failure mode	Loose fitting or valve and punctured and crushed
[C FM]	- Fiber box and other failure mode - Bottle and punctured and crushed	- Bottle and other failure mode - Fiber box and punctured and crushed
[M FM]	*Not directly associated*	*Not directly associated*
[SE FM]	Winter and other failure mode	Winter and punctured and crushed
[L FM]	- Urban/ industrial or commercial/ eastern or western and other failure mode - Suburban/ commercial/ eastern and punctured and crushed	- Urban/ industrial or commercial/ eastern or western and punctured and crushed - Suburban/ commercial/ eastern and other failure mode
[SH FM]	*Not directly associated*	*Not directly associated*

The results are summarized graphically in Figure 17 and show that four of the seven explanatory variables are directly associated with the response variable failure mode in stage three of a release.

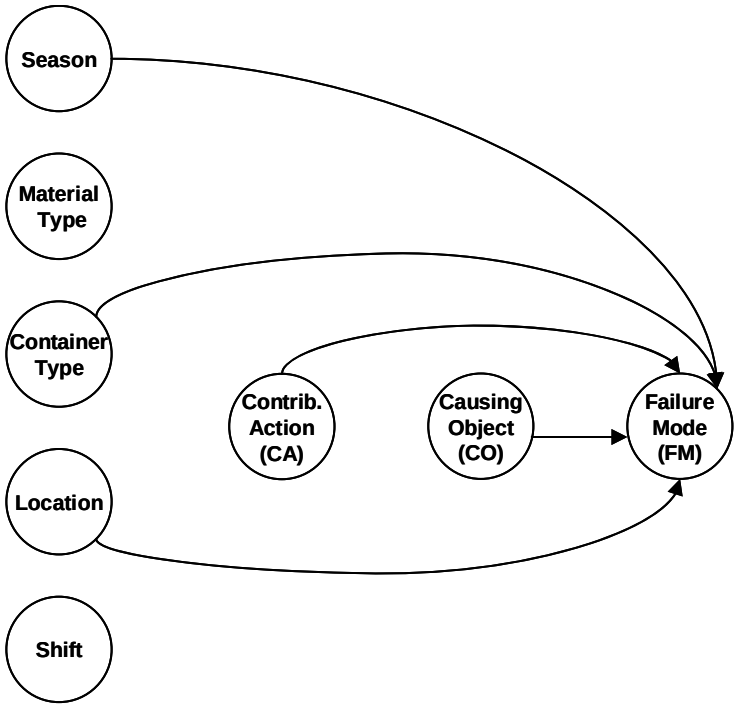


Figure 17: Graphical Results of Logit Analysis for *FM*.

Direct associations to the failure item-area were analyzed next and involved eight explanatory variables, which consisted of failure mode and the variables of stages one and two. There were 5,184 cells in this table, which was sparsely populated with 41.8% zeros and 80.8% small cell counts.

As with causing object and failure mode, contributing action also had the largest effect on failure item-area with $\lambda = 7.4955$, followed by container type ($\lambda = \pm 6.2782$), as shown in Table 48. Based on this, there were proportionally more loose fittings or valves that led to the failure of a closure on the top of the container ($\lambda = 7.4955$) as well as more bottles in which the closure on the top failed ($\lambda = 6.2782$). Causing object exerted a strong influence with $\lambda = 3.1437$, as it did in the previous logit model. Based on this, some "other" causing object, for example as opposed to the floor and water, led to proportionally more failures of the basic package material on top of the container. Failure mode was also directly related to failure item-area with $\lambda = \pm$

1.2031. This lambda indicated that a higher proportion of incidents involved some "other" failure mode of closures on top of the container, as opposed to puncture and crush, for example.

Table 48: Lambdas of Max Absolute Value and Interpretation of Largest Effects for Logit *FIA*.

Variable Pair	S1	
[FM FIA]	±	1.2031
[CO FIA]	+	3.1437
[CA FIA]	+	7.4955
[C FIA]	±	6.2782
[M FIA]	±	0.3367
[SE FIA]	-	0.5586
[L FIA]	±	0.2222
[SH FIA]	+	0.3389

Variable Pair	Largest Positive Effect	Largest Negative Effect
[FM FIA]	Other failure mode and closure on top	Punctured and crushed and closure on top
[CO FIA]	Other causing object and basic package material on top	Other causing object and basic package material on bottom
[CA FIA]	Loose fitting or valve and closure on top	Other contributing action and closure on top
[C FIA]	Bottle and closure on top	Fiber box and closure on top
[M FIA]	Corrosives and closure on top	Flammable liquids and closure on top
[SE FIA]	Winter and basic package material on bottom	Winter and closure on top
[L FIA]	Suburban/ commercial/ eastern and closure on top	Urban/ industrial or commercial/ eastern or western and closure on top
[SH FIA]	Twilight and closure on top	Twilight and basic package material on bottom

A visual summary of this logit analysis, in which five out of eight variables were directly associated to failure item-area as the response variable in stage three, is presented in Figure 18.

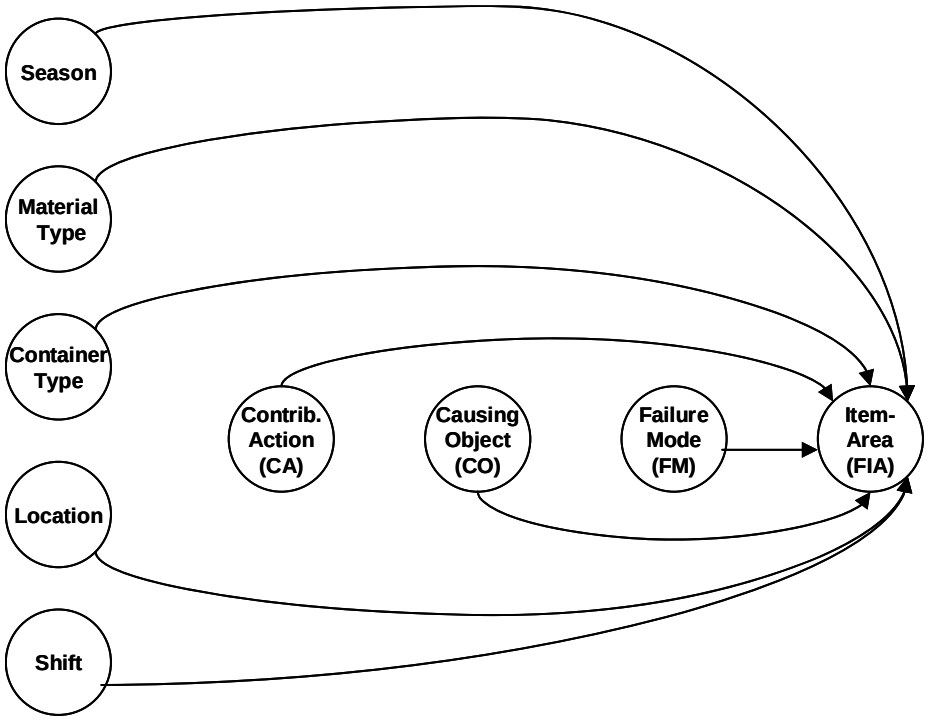


Figure 18: Graphical Results of Logit Analysis for *FIA*.

Stage Four Stage four of a release event involves the actual release of hazardous material to the environment, as quantified by a three-category variable for a zero, small, or medium-amount release, as described in Table 49.

Table 49: Stage Four Variables.

Stage 4 Variable	Abbreviation	Categories
Release Quantity	RQ	1) Zero 2) Small (≤1 gal) 3) Medium (≤100 gal)

The stage one variables and the container failure variables in stages two and three served as explanatory variables to release quantity. With nine explanatory variables, the table was large and sparse, with a total of 15,552 cells.

At stage four, the general correction procedure described in section 3.6.2 for a table with two or more latent variables and one or more observed variables was applied based on its availability at the time of the stage four analysis. This procedure corrected for the bias introduced by the five latent variables, which consisted of location, contributing action, causing object, failure mode, and failure item-area. Due to their low classification errors, the correction procedure had a small effect. However, the procedure was nonetheless applied to demonstrate its application as part of three-step model building. As discussed previously in section 3.6.2, this procedure involved calculations involving the Kronecker product of the transition matrices for the five latent variables. The Kronecker product of the transition matrices was large (108 *X* 108) and exceeded the limits of Microsoft Excel 2002. As such, source code was written to determine the corrected matrix, using text files for output. The individual transition matrices for the five latent variables are shown in Table 80 in Appendix B. Due to space considerations, the Kronecker product and the uncorrected and corrected matrices are not shown. This correction procedure had a limited effect, as expected, due to the low classification errors of the latent variables. The difference between the total counts in the uncorrected and corrected contingency tables was 97, which represented 0.24% of the total count in the uncorrected table. The maximum difference in any cell was 4. Nonetheless, the corrected table was used for the stage four loglinear analysis.

Based on this analysis, the direct effects on release quantity tended to be smaller than the direct effects on previous logit variables. The container failure variables continued to exert the largest influences. Causing object had the largest effect on release quantity with $\lambda = -1.1925$. Based on this lambda parameter, a smaller proportion of incidents with some "other" causing object were associated with a medium release quantity. This "other" causing object is opposed to the floor and water, for example. Causing object was followed by failure mode, container

type, and item-area, with λ = ± 0.9578, ± 0.7952, and 0.7260, respectively. Contributing action, the most influential variable in several prior models, was much less influential on release quantity (λ = 0.5785). This lambda indicates that proportionally more improperly loaded and dropped containers led to a medium release amount. Material type was not directly related to release quantity, based on λ = ± 0.1730.

Table 50: Lambdas of Max Absolute Value and Interpretation of Largest Effects for Logit *RQ*.

Variable Pair	S1	
[FIA RQ]	+	0.7260
[FM RQ]	±	0.9578
[CO RQ]	-	1.1925
[CA RQ]	+	0.5785
[C RQ]	±	0.7952
[M RQ]	±	0.1730
[SE RQ]	-	0.2208
[L RQ]	±	0.4004
[SH RQ]	+	0.2784

Variable Pair	Largest Positive Effect	Largest Negative Effect
[FIA RQ]	Closure on top and small release quantity	Closure on top and medium release quantity
[FM RQ]	Punctured and crushed and medium release quantity	Other failure mode and medium
[CO RQ]	Other causing object and zero release quantity	Other causing object and medium
[CA RQ]	Improper loading and dropped and medium release quantity	Other contributing action and medium release quantity
[C RQ]	Fiber box and medium release quantity	Bottle and medium release quantity
[M RQ]	*Not directly associated*	*Not directly associated*
[SE RQ]	Summer and medium release quantity	Spring and medium release quantity
[L RQ]	Suburban/ commercial/ eastern and medium release quantity	Urban/ industrial or commercial/ eastern or western and medium release quantity
[SH RQ]	Twilight and medium release quantity	Twilight and zero release quantity

The network for release quantity, which consists of six direct associations with the explanatory variables in stages one through three of a release, is presented graphically in Figure 19.

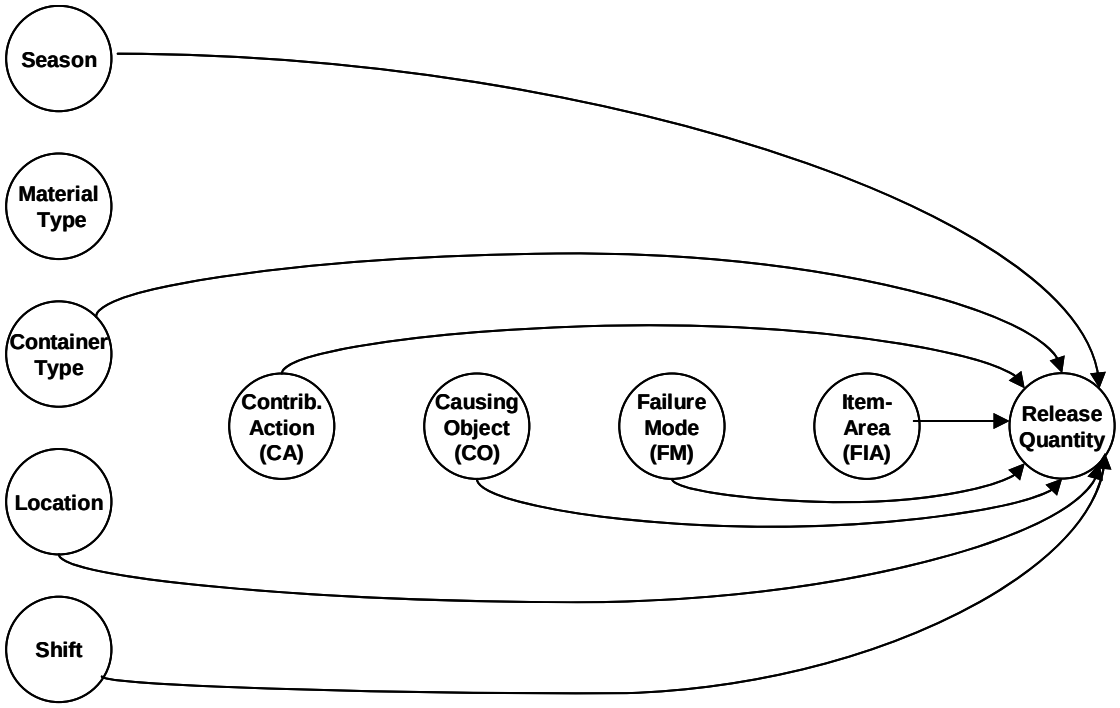


Figure 19: Graphical Results of Logit Analysis for *RQ*.

Stage Five The final, or ultimate, logit variable in the network was dollar loss, one possible consequence of a hazmat release and a measure for risk identified in the literature. Dollar loss has the following categories: zero, small, and medium, as summarized in Table 51 .

Table 51: Stage 5 Variables.

Stage 5 Variable	Abbreviation	Categories
Dollar Loss	D	1) Zero 2) Small (≤$500) 3) Medium (≤$25K)

There were ten variables that preceded dollar loss in the network. However, *SPSS 11.0* has a limitation of nine explanatory variables. To address this limitation, the contingency table

containing all 11 variables (which included dollar loss) was collapsed over one of the explanatory variables that had a small impact in previous stages, as suggested by Hagenaars.[207] Consequently, the table was collapsed over shift. This led to a model containing nine explanatory variables, in which shift was not modeled. This table had 15,552 cells and a sample size of 40,282, after the general correction procedure was applied. The correction procedure again had a small effect, with a difference in total count between uncorrected and corrected tables of 91, which represented 0.23% of the uncorrected total count. There was a maximum cell difference of 8. In order to investigate possible effects of shift on dollar loss, a second model was run. In this model, the contingency table was collapsed over season, another variable that had a small impact on previous response variables. This table was also corrected, resulting in a total difference of 70 (0.17%) between the uncorrected and corrected tables and a maximum cell difference of 10.

The lambda values for the two models were very similar, as shown in Table 52. In this table, the model on the left was collapsed over shift, and the model on the right was collapsed over season. In addition, interpretations of the lambda parameters are given in Table 53. Causing object had the largest influence on dollar loss ($\lambda \cong - 4.3$), as it did on release quantity. Based on this as shown in Table 53, a lesser proportion of some "other" causing object was associated with medium dollar loss. This is in contrast to the largest positive effect by causing object, in which a greater proportion of floor and water incidents that led to medium dollar loss ($\lambda \cong 3.2$). Contributing action had the next largest effect ($\lambda \cong 3.6$), followed by release quantity ($\lambda \cong 3.4$) and failure item-area ($\lambda \cong -2.3$). The largest positive effect by the contributing action indicated that a greater proportion of loose fittings or valves were associated with medium dollar loss. The largest positive effect by release quantity led to proportionally more incidents in which a medium

release quantity was associated with medium dollar loss, as might be expected. In summary, the container failure variables and release quantity had strong effects on the ultimate logit variable. The only variable which was not directly associated with dollar loss was container type ($\lambda \cong \pm 0.11$). Both shift and season showed direct effects on dollar loss ($\lambda = 0.4749$ and $\lambda = 0.5654$), which points to the value of having run both models.

Table 52: Lambdas of Max Absolute Value for Logit *D*.
Collapsed over Shift (left) and Season (right).

Variable Pair	S1	
[RQ D]	+	3.4386
[FIA D]	-	2.2529
[FM D]	±	0.5156
[CO D]	-	4.4261
[CA D]	+	3.5947
[C D]	±	0.1130
[M D]	±	0.5511
[SE D]	+	0.5654
[L D]	±	1.3196

Variable Pair	S1	
[RQ D]	+	3.3845
[FIA D]	-	2.2660
[FM D]	±	0.5262
[CO D]	-	4.2793
[CA D]	+	3.6344
[C D]	±	0.1113
[M D]	±	0.5160
[L D]	±	1.3434
[SH D]	+	0.4749

The interpretations of the strongest effects on dollar loss are presented below in Table 53.

Table 53: Interpretation of the Largest Effects for Logit *D*.

Variable Pair	Largest Positive Effect	Largest Negative Effect
[RQ D]	Medium release quantity and medium dollar loss	Zero release quantity and medium dollar loss
[FIA D]	Closure on top and medium dollar loss	Basic packaging material on top and medium dollar loss
[FM D]	Punctured and crushed and medium dollar loss	Other failure mode and medium dollar loss
[CO D]	Floor and water and medium dollar loss	Other causing object and medium dollar loss
[CA D]	Loose fitting or valve and medium dollar loss	Loose fitting or valve and zero dollar loss
[C D]	*Not directly associated*	*Not directly associated*
[M D]	Corrosives and medium dollar loss	Flammable liquids and medium dollar loss
[SE D]	Fall and medium dollar loss	Summer and medium dollar loss
[L D]	Suburban/ commercial/ eastern and medium dollar loss	Urban/ industrial or commercial/ eastern or western and medium dollar loss
[SH D]	Day and medium dollar loss	Twilight and medium dollar loss

As presented in Figure 20, there are direct associations between dollar loss and all the explanatory variables in stages one through four except for container type, as shown by the absence of an arc from this variable.

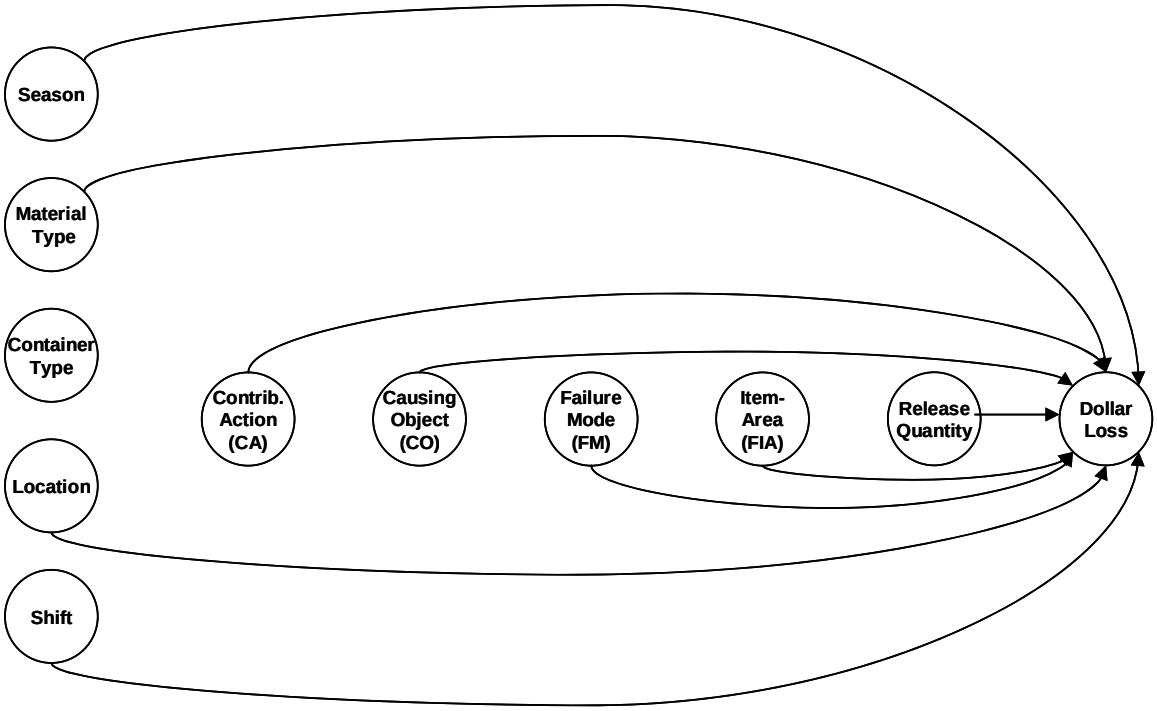


Figure 20: Graphical Results of Logit Analysis for *D*.

Summary of Stages 1-5 A flowchart that summarizes the method used for constructing and using the individual loglinear models to analyze each stage of a hazmat release is provided in Figure 21.

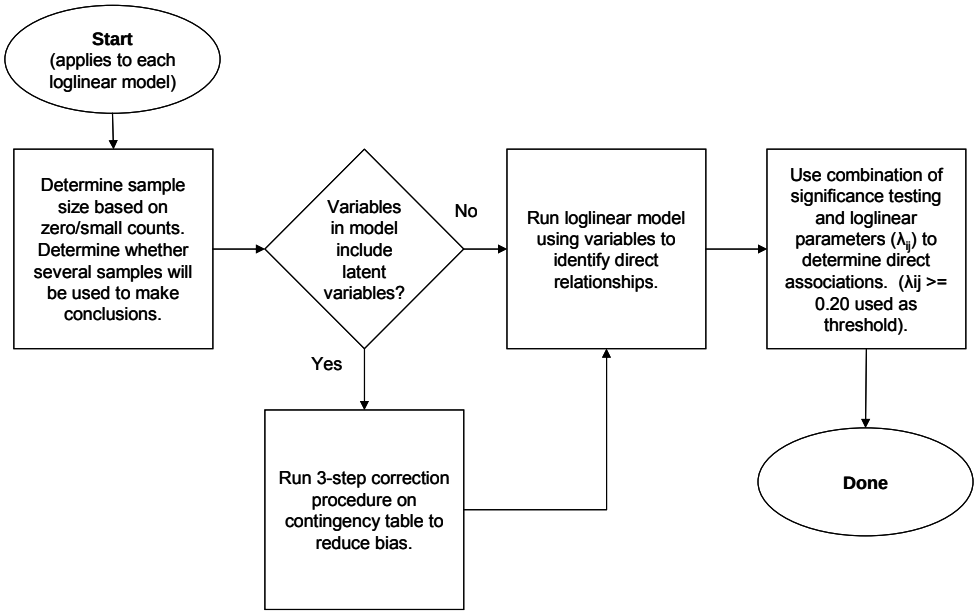


Figure 21: Approach to Constructing and Using a Loglinear Model.

A complete structure showing all the direct associations determined among the variables in stages one through five is given in Figure 22. In this diagram, the stage one variables are positioned on the outer ring of the circle, while dollar loss is in the center. This diagram shows the high degree of interconnectedness of the network. This structure served as the association structure for the Bayesian network decision model of the hazmat release variables, to be constructed next in the methodology.

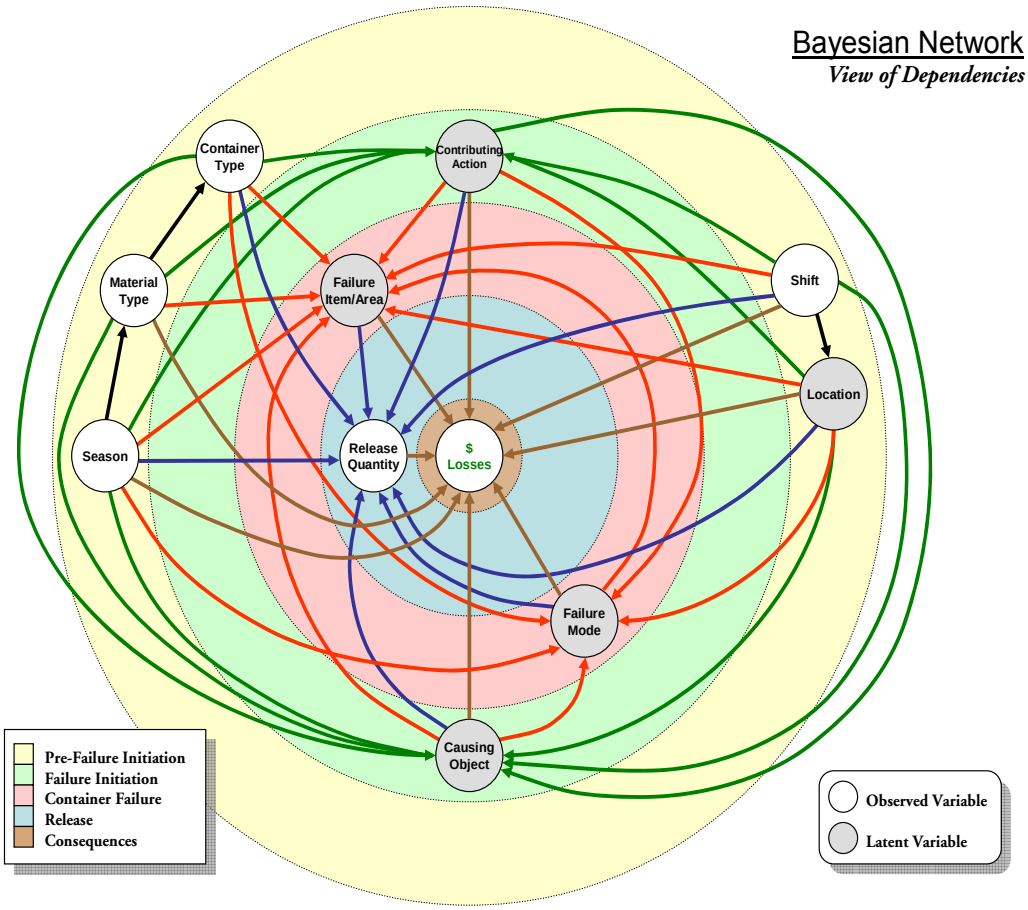


Figure 22: Bayesian Network for Stages 1-5.

4.1.3 Three Step Modeling Assumptions

An assumption of the three step modeling approach, which was just used to develop the association structure, is that an observed variable does *not* have a direct effect on an indicator variable of a latent variable. This assumption is stated in various publications by the authors of the correction procedure, namely Bolck, Croon, and Hagenaars.[208, 209, 210, 211] For example, this assumption holds that container type, an observed variable, should not have a direct effect on improper loading, an indicator variable for the contributing action latent variable. If there is a direct relationship, as measured by a significant L^2 or $|\lambda_{ij}| \geq 0.20$, then the indicator variable should be converted to an observed variable.[212]

Efforts at the beginning of the loglinear modeling to investigate these direct effects among the hazmat variables in order to comply with the assumption were not met with success. Early efforts were not successful because the contingency tables for the required loglinear models were sparse. At this time, significance testing was the method being used to determine the presence of associations, or direct effects. The ability to use lambda parameters to assess associations in sparse tables was not known until later in time. In addition, in the literature, violation of this assumption was described as leading to issues with interpretation of the final model versus issues of model accuracy or estimation of lambda parameters.[213, 214] Based on this and the inability to perform significance testing, the loglinear modeling proceeded without an investigation of the possible direct effects.

After completion of the loglinear modeling, I learned through personal communication with Dr. Jacques Hagenaars that a "large" direct effect of an observed variable on an indicator can be influential on the estimation of the lambda parameters in the model.[215] However, the value of a "large" direct effect is uncertain. Dr. Hagenaars believed that a value of λ=0.75 may *not* be a large direct effect, especially if latent variables are involved.[216] At this point, I ran the necessary loglinear models to investigate the direct effects of the observed variables on the indicator variables based on the lambda parameters. This was done for each indicator variable influenced by an observed variable in the hazmat network. The presence of a direct effect was determined through a three-variable loglinear model consisting of the observed, indicator, and latent variable.

Of the four observed variables that may directly influence the indicator variables, container type had the largest direct effects. The direct effects of container type ranged from λ=0.01 to λ=5.75, with an average of λ=1.07. Shift in general had smaller direct effects, which ranged

from λ=0.0002 to λ=10.21, with an average of λ=0.74. Finally, the direct effects of season and material type were smaller than those of container type and shift, as presented below in Table 54.

Table 54: Direct Effects of Observed Variables on Indicator Variables.

Observed Variable	Range		Average
Container Type	0.0100	5.75	1.07
Shift	0.0002	10.21	0.74
Season	0.0093	8.06	0.51
Material Type	0.0100	6.79	0.45

The indicator variables upon which the largest direct effects occurred were "other" and "none." For example, shift had a direct effect of λ =10.21 on the indicator variable "other" of the latent variable contributing action.

Since the value of a "large" direct effect is not known, the impact of the direct effects of the observed variables on the indicator variables in this research is not known. Since the "other" and "none" indicators were in general associated with the largest direct effects, an opportunity for future research is elimination of these indicator variables from the analysis. Although "other" and "none" were associated with many incidents and consequences, they provide limited information and contribute to violation of the three-step modeling assumption. As a consideration for future research, if a threshold for "large" had been chosen and the three step assumption followed by converting indicators to observed variables based on large λ's, then the same degree of simplification using latent class analysis would not have been possible. For, many of the indicator variables would have been included as observed variables and would not have been combined as indicators to form the latent variables. In addition, not all of the latent variables that were developed in this research would have been possible. Thus, there is a tradeoff between the ability to simplify and the three-step modeling assumption.

4.1.4 Bayesian Network Construction

Using the association structure determined from the three step modeling approach, a Bayesian network decision model of a hazardous materials release was developed as the final step of the methodology. A Bayesian network was chosen as the type of decision model since only random variables are considered in this research. In addition, a Bayesian network was a natural fit due to its ability to perform inference on the variables, including ranking the explanatory variables based on their influence on the outcome variable. Also, changes in the distribution of the outcome variable could be determined based on changes in the categories of the explanatory variables, suggesting desirable operational or policy changes. Dollar loss, the ultimate response variable, and release quantity, the stage-four response variable of potential concern to environmentalists, were analyzed as network outcome variables. The following sections describe the training, testing, and inferential results. The most influential variables and recommended policy changes for each outcome variable are discussed in detail.

4.1.4.1 Bayesian Network Training

The second component of a Bayesian network, the conditional probability distribution of each node or variable, was determined using incident counts from the HMIRS database. Calculation of these conditional probability distributions constitutes the training of the network, or learning, at which point inference can be performed.

Some of the conditional probabilities were zero due to the large number of parent, or explanatory, variables, such as those of dollar loss and release quantity. Conditional probabilities of zero were replaced by a small constant equal to 0.0001, as suggested in the

literature as well as by an expert.[217, 218] The rationale for this was that the zeros were probably not structural zeros but sampling zeros. And, using Bayes Theorem, a zero probability can never be updated to another value.[219]

There were also instances in which the probability in the denominator of the conditional probability calculation was zero, thereby preventing calculation of the conditional probability. In these cases, the uniform distribution was applied to the conditional probabilities that make use of this particular probability in the denominator.[220] Using the uniform distribution, each of these conditional probabilities was assigned an equal value such that the values summed to one. For example, for a three-category variable, the values *1/3, 1/3, 1/3* were assigned.

A total of 40,191 records were available for network training and testing. These records were divided into five non-overlapping sets of equal size so that five-fold cross validation, or testing, could be done. *Cross validation* involves comparing model predictions with actual values using a set of test records in order to assess model accuracy. Although there is no general or gold standard for assessing the accuracy of Bayesian networks, cross validation was performed on the hazmat release networks to document their accuracy.[221] Thus, five Bayesian networks were trained and tested using the available data. The method for doing this is described in Table 55, where the five sets of records are represented by T1 through T5. For example, for the first Bayesian network, sets T2-T5 were used for training, and T1 was used for testing, or cross validating. For the second network, sets T1 and T3-T5 were used for training, and T2 was used for testing. The plan described in this table resulted in five sets of testing and inference results, which will be described in the following sections.

Table 55: Bayesian Network Training and Test Plan.

Network	Training Sets	Test Set
1	T2,T3,T4,T5	T1
2	T1,T3,T4,T5	T2
3	T1,T2,T4,T5	T3
4	T1,T2,T3,T5	T4
5	T1,T2,T3,T4	T5

4.1.4.2 Dollar Loss Outcome – Testing and Quality

Cross-validation is commonly done after training to evaluate the resulting network by comparing model predictions to actual values using an independent set of test records. For example, to evaluate the ability of the network to predict the correct category for dollar loss, the values for season, material type, and the other explanatory variables were set as evidence based on their test record values. The probability for dollar loss was then updated using the network, leading to a most likely category for this outcome variable. The most likely category was then compared to the actual category per the test record to identify the existence of a match. This procedure was done for all records in the test set, leading to a calculation of the accuracy. Specifically, a count of the records in which the actual category matched the most likely category was made, along with a count of matches on the next most likely category. These two counts gave an indication of the accuracy of the network, as described in Onisko et. al.[222] In predicting dollar loss, the accuracies of the five networks are given in Table 56. The most likely category matched the actual category approximately 70% of the time, ranging from 68.9% to 70.8% across the five networks. Thus, the accuracies associated with the five test sets were close in value, likely due to the large training and testing datasets of approximately 32,000 and 8,000, respectively. When there was not a match on the most likely category, the next most likely category matched the actual approximately 23% of the time, ranging from 22.5% to 24.2% across the test sets.

Although the networks did not make perfect predictions for dollar loss, the accuracy appears to be reasonable. The inability to make perfect predictions reflects the difficulty of the problem and may be the result of other factors or variables that are not part of the model.[223]

Table 56: Prediction Accuracies for Dollar Loss.

	Network				
	T1	**T2**	**T3**	**T4**	**T5**
Most Likely = Actual	69.6%	68.9%	69.4%	70.8%	70.2%
Next Most Likely = Actual	23.3%	24.2%	23.8%	22.5%	22.8%

Tests were also performed in the reverse direction to assess the ability to predict values of the explanatory variables based on the category of dollar loss. To do this, the value for dollar loss was set as evidence based on the test record value. The probabilities for all the explanatory variables were then updated using the network, and counts of the most likely and next most likely matches for each parent variable were determined, as shown in Table 57. For example, for season, the percentage of records in which the actual category matched the most likely category was approximately 31%, ranging from 30.2% to 32.2% across the five test sets. The accuracies for each explanatory variable given in Table 57 were always better than random based on the number of categories of the variable. In addition, for several of the variables, the predictions were notably better than random, such as those for container (*C*), contributing action (*CA*), causing object (*CO*), failure mode (*FM*), failure item-area (*FIA*), and release quantity (*RQ*). Note the excellent prediction of *RQ* at approximately 88%.

Table 57: Prediction Accuracies for Dollar Loss Explanatory Variables.

		Matches for Dollar Loss by Network									
		T1		T2		T3		T4		T5	
Explanatory Variable	Cat.	Most Likely	Next Most Likely	Most Likely	Next Most Likely	Most Likely	Next Most Likely	Most Likely	Next Most Likely	Most Likely	Next Most Likely
Season	4	31.4%	27.5%	30.2%	27.3%	31.2%	27.5%	31.1%	26.9%	32.2%	27.0%
Material Type	2	60.5%		59.3%		59.9%		58.7%		58.6%	
Container Type	2	75.5%		76.0%		75.7%		75.0%		75.4%	
Shift	3	37.3%	34.6%	37.6%	33.9%	36.3%	34.8%	36.9%	34.7%	36.3%	35.0%
Location	2	54.4%		54.3%		54.6%		53.3%		54.6%	
Contributing Action	3	53.4%	33.6%	54.7%	32.5%	54.0%	33.5%	53.8%	33.2%	54.8%	32.2%
Causing Object	3	63.3%	20.8%	63.9%	20.8%	62.7%	20.8%	63.6%	20.4%	63.6%	20.6%
Failure Mode	2	69.5%		70.2%		69.7%		70.1%		70.8%	
Failure Item-Area	3	51.2%	32.5%	51.6%	32.5%	51.1%	32.8%	50.6%	33.2%	51.1%	33.0%
Release Quantity	3	87.6%	9.5%	88.1%	9.1%	87.7%	9.8%	88.0%	9.3%	88.3%	9.1%
Note: 'Next Most Likely' percentage not listed for binary variables.											

Another means of testing the quality of a network is through a *MAP*, or maximum aposteriori probability.(224) A *MAP* is the probability of the most likely joint state of the explanatory variables given the outcome variable and is a feature available in *GeNIe*. A *MAP* can be compared to the conditional probability as determined using record counts from a database. The proximity of these probabilities is an indication of the quality of the network. The *MAP* of the variables in stages two through four, which were highly influential to medium dollar loss, was compared to the conditional probability based on record counts from the database. This comparison was made for each of the five networks and shows good agreement between these two probabilities, as shown in Table 58.

Table 58: MAP for Medium Dollar Loss vs. Database Probability Calculation.

	Network				
	T1	T2	T3	T4	T5
MAP - Medium Dollar Loss	16.1%	16.0%	16.5%	16.4%	16.6%
Database Probability Calculation	16.5%	16.5%	16.9%	17.0%	17.0%

4.1.4.3 Release Quantity Outcome – Testing and Quality

The prediction accuracy for the other outcome variable, release quantity, was determined in the same fashion as that for dollar loss. The accuracy for release quantity was actually better than the accuracy for dollar loss and is given in Table 59. The most likely category matched the actual value approximately 87% of the time, versus approximately 70% in the case of dollar loss. When there was not a match on the most likely category, the next most likely category matched approximately 10% of the time, ranging from 10.0% to 10.9% in the test sets. For each of the two types of accuracies, the percentages were very close in value across the five networks, likely due to the size of the training and test sets.

Table 59: Prediction Accuracies for Release Quantity.

	Network				
	T1	T2	T3	T4	T5
Most Likely = Actual	86.7%	86.8%	86.4%	86.9%	87.1%
Next Most Likely = Actual	10.3%	10.3%	10.9%	10.1%	10.0%

A test was likewise performed in the reverse direction to determine the ability to predict the parent variables of release quantity based on the category of this outcome variable. Percentages for the most likely and next most likely category matches for each explanatory variable are given in Table 60. Again, the accuracy was always better than random and notably better for some of the container-related variables, which included container type, contributing action, causing object and failure mode.

Table 60: Prediction Accuracies for Release Quantity Explanatory Variables.

		Matches for Release Quantity by Network									
		T1		T2		T3		T4		T5	
Explanatory Variable	Cat.	Most Likely	Next Most Likely	Most Likely	Next Most Likely	Most Likely	Next Most Likely	Most Likely	Next Most Likely	Most Likely	Next Most Likely
Season	4	31.2%	27.4%	30.2%	27.1%	30.8%	27.5%	31.6%	26.5%	31.9%	26.9%
Material Type	2	60.5%		59.3%		59.9%		58.7%		58.6%	
Container Type	2	75.5%		76.0%		75.7%		75.0%		75.4%	
Shift	3	36.8%	35.1%	35.0%	36.4%	36.0%	35.2%	36.1%	35.6%	36.2%	35.1%
Location	2	51.4%		51.4%		51.7%		51.2%		51.9%	
Contributing Action	3	52.6%	34.3%	53.8%	33.5%	53.9%	33.6%	54.0%	33.0%	54.4%	32.6%
Causing Object	3	60.6%	22.1%	60.9%	22.1%	59.8%	22.8%	60.3%	22.1%	59.9%	23.2%
Failure Mode	2	69.5%		70.2%		69.7%		70.1%		70.8%	
Failure Item-Area	3	43.1%	40.6%	43.2%	41.0%	43.3%	41.2%	42.5%	41.3%	43.4%	40.7%

As another test of the quality of the network, a *MAP* of the variables in stages two and three, which were most influential on medium release quantity, was compared to the conditional probability as determined using record counts from the database. This comparison for each of the five networks is given in Table 61 and indicates good agreement.

Table 61: MAP for Medium Release Quantity vs. Database Probability Calculation.

	Network				
	T1	T2	T3	T4	T5
MAP - Medium Release Quantity	20.3%	20.4%	20.5%	19.2%	20.4%
Database Probability Calculation	20.4%	20.3%	20.5%	19.2%	20.5%

4.2 RESULTS AND INFERENCES OF THE BAYESIAN NETWORK FOR THE WORKED EXAMPLE

Bayesian networks can be used in both a strategic and tactical manner. Thus, they can be used to make long-range plans as well as to answer queries of a more reactive, short-term nature. An overarching goal of this research was identification of the key variables in a large categorical database, which represented a strategic use of the Bayesian network. The next two sub-sections discuss the strategic results related to dollar loss and release quantity and focus on the key variables and desirable changes identified for them. The final sub-section discusses potential tactical uses of the Bayesian network by the Office of Hazardous Materials, such as "what-if" analyses surrounding exemption approvals and occurrence spikes.

4.2.1 Dollar Loss Outcome – Strategic Results and Inferences

One of the most valuable types of information that can be determined from a Bayesian network developed using *GeNIe* is a ranking of the explanatory variables based on their information value relative to the outcome variable. This can also be viewed as a measure of their influence on or the degree to which they reduce the uncertainty in the outcome variable.[225, 226] The concept of value of information is one method of handling or studying decision models. The explanatory variables were ranked according to their degree of influence on, or information value relative to, each category of dollar loss for each of the five networks. Thus, five sets of inferences are presented and summarized in the following sections. The five sets of ranking results for zero, small, and medium dollar loss are presented in Table 62 through Table 64, respectively. In summary, causing object was the leading diagnostic variable for dollar loss, regardless of the loss category, followed by the failure item-area. Thus, these two variables were found to be the most influential or informative variables relative to dollar loss. In Table 62, which presents the

ranking results specific to zero dollar loss, causing object was the leading variable in 3/5 networks and nearly tied as the leading variable in a fourth network (T5). For zero loss, failure item-area was clearly the second leading variable, based on the results from 3/5 networks.

Table 62: Ranking of 'Zero' Dollar Loss Parent Variables.

	Zero Dollar Loss Category									
Rank	T1	Ranking Value	T2	Ranking Value	T3	Ranking Value	T4	Ranking Value	T5	Ranking Value
1	CO	0.0972	CO	0.0957	CO	0.0968	FIA	0.0966	FIA	0.0936
2	FIA	0.0939	FIA	0.0927	FIA	0.0956	CO	0.0951	CO	0.0935
3	L	0.0084	L	0.0080	L	0.0083	L	0.0096	L	0.0083
4	RQ	0.0062	RQ	0.0072	RQ	0.0064	RQ	0.0073	RQ	0.0067
5	M	0.0035	CA	0.0033	CA	0.0031	M	0.0036	CA	0.0030
6	CA	0.0029	M	0.0030	M	0.0030	CA	0.0031	M	0.0029
7	FM	0.0015	FM	0.0017	FM	0.0017	FM	0.0017	FM	0.0010
8	C	0.0006	C	0.0008	C	0.0009	C	0.0010	C	0.0008
9	SH	0.0003	SH	0.0001	SH	0.0003	SE	0.0004	SH	0.0003
10	SE	0.0002	SE	0.0001	SE	0.0002	SH	0.0002	SE	0.0003

For both small and medium dollar loss, which are shown in Table 63 and Table 64, causing object was clearly the leading variable, followed by failure item-area, both based on 5/5 networks.

Table 63: Ranking of 'Small' Dollar Loss Parent Variables.

	Small Dollar Loss Category									
Rank	T1	Ranking Value	T2	Ranking Value	T3	Ranking Value	T4	Ranking Value	T5	Ranking Value
1	CO	0.0186	CO	0.0183	CO	0.0192	CO	0.0173	CO	0.0179
2	FIA	0.0161	FIA	0.0158	FIA	0.0167	FIA	0.0164	FIA	0.0158
3	RQ	0.0051	RQ	0.0057	RQ	0.0054	RQ	0.0065	RQ	0.0057
4	CA	0.0030	CA	0.0030	CA	0.0027	CA	0.0028	CA	0.0031
5	FM	0.0021	FM	0.0020	FM	0.0020	FM	0.0023	FM	0.0026
6	SE	0.0021	SE	0.0014	SE	0.0016	SE	0.0020	SE	0.0017
7	SH	0.0007	SH	0.0005	SH	0.0005	SH	0.0008	SH	0.0005
8	M	0.0005	M	0.0003	M	0.0004	M	0.0004	M	0.0002
9	C	0.0001	C	<0.0001	C	0.0001	C	<0.0001	C	0.0001
10	L	<0.0001	L	<0.0001	L	<0.0001	L	<0.0001	L	<0.0001

Table 64: Ranking of 'Medium' Dollar Loss Parent Variables.

	Medium Dollar Loss Category									
Rank	T1	Ranking Value	T2	Ranking Value	T3	Ranking Value	T4	Ranking Value	T5	Ranking Value
1	CO	0.0393	CO	0.0385	CO	0.0413	CO	0.0390	CO	0.0375
2	FIA	0.0307	FIA	0.0306	FIA	0.0330	FIA	0.0306	FIA	0.0314
3	CA	0.0186	CA	0.0175	CA	0.0195	CA	0.0190	CA	0.0186
4	FM	0.0172	FM	0.0172	FM	0.0176	FM	0.0182	FM	0.0171
5	L	0.0153	L	0.0151	L	0.0152	L	0.0142	L	0.0155
6	RQ	0.0049	RQ	0.0061	RQ	0.0050	RQ	0.0061	RQ	0.0062
7	SE	0.0045	SE	0.0050	SE	0.0046	SE	0.0053	SE	0.0043
8	SH	0.0029	SH	0.0028	C	0.0037	SH	0.0039	SH	0.0032
9	C	0.0028	C	0.0027	SH	0.0033	C	0.0028	C	0.0029
10	M	0.0013	M	0.0017	M	0.0014	M	0.0020	M	0.0018

Therefore, based on these results, causing object and item-area should be the top focuses of policy or operational change initiatives. In general, the container failure variables occupied upper positions in the rankings for dollar loss, indicating the value of this type of information.

In addition, *GeNIe* can be used to determine desirable changes in an explanatory variable in order to best impact an outcome variable. For example, using *GeNIe*, it was determined that a reduction in incidents involving the floor and water/liquid as the causing object should be targeted, based on all five networks. This was determined based on changes in the probability distribution of dollar loss given the particular category of causing object. For example, using the first network (T1) to illustrate this method, setting the floor and water category of causing object as evidence led to an increase in the occurrence of both small and medium dollar loss and a decrease in zero loss relative to no evidence, as shown in Table 65. Thus, the floor and water causing object had an undesirable effect on dollar loss and should be the target of reduction efforts to best impact dollar loss. The "none" and "other" categories of causing object generally produced the opposite effect on dollar loss.

Table 65: Effects of Causing Object on Dollar Loss. (T1 network)

Causing Object	Dollar Loss		
	ZERO	SMALL	MEDIUM
no evidence	0.167	0.718	0.116
-			
Floor and water/ liquid	0.083	0.770	0.147
None	0.199	0.682	0.119
Other	0.364	0.607	0.029

This method of examining the probabilities to select the preferred alternative is based on the concept of *stochastic dominance*. Using stochastic dominance, the dominating alternative is the one more likely to lead to a particular outcome or consequence. It is in effect the better gamble. Stochastic dominance is a means to formally screen various alternatives and eliminate choices based on the pattern of the probabilities.[227] Based on the T1 network as shown in Table 65, the best option or policy for a reduction in dollar loss was the floor and water/liquid combination because it increased the probability for a medium loss the most relative to no evidence (from 0.116 to 0.147). In addition, it also increased the probability for a small loss from 0.718 to 0.770, while the other categories decreased this probability. Although this analysis for causing object was fairly straightforward, more than one policy change initiative may be desirable for a variable, depending on the category of dollar loss targeted for reduction. For, although the medium category has a higher dollar amount associated with it, it occurs less frequently than the small category (12% versus 72%, respectively). Thus, targeting the small category may impact more incidents.

A summary of recommended policy changes for each variable based on all five networks using this same type of analysis is given in Table 66. A few of the recommendations are based on "best of five" network outcomes where necessary. For example, this table shows that

incidents involving the floor and water category should be the focus of reduction efforts in order to best impact dollar loss. For failure item-area, there are two courses of action, depending on the category of dollar loss targeted. Specifically, if small loss is to be reduced, the best category to target is basic packaging material on the bottom of the container. However, to best impact medium dollar loss, either basic packaging material on the bottom or closures on top should be targeted. When evaluated from an overall perspective, the best category of failure item-area to target for reduction is basic packaging material on the bottom since it increases the probability of both a small and medium dollar loss and (undesirably) decreases the probability of zero loss the most. The five sets of probability distributions for dollar loss on which these recommended policy changes are based are provided in Table 81 through Table 85 in Appendix C.

Table 66: Recommended Policy Changes for Impacting Dollar Loss.

Explanatory Variable	Targeted Categories
Causing Object	To best impact small and medium loss, target floor and water/liquid. Overall best choice is floor and water/liquid since it also decreases the probability of zero loss.
Failure Item-Area	To best impact small loss, target basic package material on bottom. To best impact medium loss, target closure on top or basic package material on bottom. Overall best choice is basic package material on bottom since it increases probability of small and medium loss and decreases probability of zero loss the most.
Contributing Action	To best impact small loss, target loose fitting or valve or other. To best impact medium loss, target improper loading and dropped. Overall best choice is loose fitting or valve since it increases probability of small and medium loss and decreases probability of zero loss the most.
Failure Mode	To best impact small loss, target other. To best impact medium loss, target punctured and crushed.
Location	To best impact medium loss, target suburban/ commercial/ eastern. There is limited information for small loss.
Release Quantity	To best impact small loss, target small release quantity. To best impact medium loss, target medium release quantity.
Season	To best impact small loss, target summer. To best impact medium loss, target fall.

Table 66 (continued).

Container Type	To best impact medium loss, target bottle. There is limited information for small loss. There is limited information provided by fiber box. Overall best choice is bottle since it also decreases the probability of zero loss.
Shift	To best impact small loss, target twilight. To best impact medium loss, target day.
Material Type	To best impact small and medium loss, target corrosives. Overall best choice is corrosives since it also decreases the probability of zero loss.

4.2.2 Release Quantity Outcome – Strategic Results and Inferences

As with the dollar loss outcome, the container failure variables were the most influential variables relative to release quantity based on the entropy-based ranking procedure in *GeNIe*. However, for release quantity, the leading container failure variable varied by category of release quantity. For example, as shown in Table 67, causing object (*CO*) was the most influential variable relative to zero release quantity, followed by failure item-area (*FI*A), both based on 5/5 networks.

Table 67: Ranking of 'Zero' Release Quantity Parent Variables.

	Zero Release Quantity Category									
Rank	**T1**	**Ranking Value**	**T2**	**Ranking Value**	**T3**	**Ranking Value**	**T4**	**Ranking Value**	**T5**	**Ranking Value**
1	CO	0.0250	CO	0.0256	CO	0.0277	CO	0.0259	CO	0.0276
2	FIA	0.0112	FIA	0.0120	FIA	0.0132	FIA	0.0110	FIA	0.0123
3	CA	0.0042	CA	0.0065	CA	0.0050	CA	0.0053	CA	0.0071
4	FM	0.0030	FM	0.0039	FM	0.0039	FM	0.0043	FM	0.0044
5	SE	0.0020	SE	0.0024	SE	0.0027	L	0.0015	SE	0.0021
6	L	0.0016	L	0.0013	L	0.0013	SE	0.0009	L	0.0017
7	C	0.0009	C	0.0006	C	0.0005	M	0.0001	SH	0.0003
8	SH	<0.0001	SH	<0.0001	SH	0.0005	SH	<0.0001	C	0.0002
9	M	<0.0001	M	<0.0001	M	0.0001	C	<0.0001	M	<0.0001

However, as shown in Table 68, failure mode (*FM*) was the leading variable in 3/5 networks and tied as the leading variable in another network (T1) for small release quantity. Contributing action (*CA*) was a close second to failure mode based on 3/5 networks.

Table 68: Ranking of 'Small' Release Quantity Parent Variables.

	Small Release Quantity Category									
Rank	T1	Ranking Value	T2	Ranking Value	T3	Ranking Value	T4	Ranking Value	T5	Ranking Value
1	CA	0.0093	FM	0.0092	FM	0.0090	FIA	0.0083	FM	0.0085
2	FM	0.0093	CA	0.0089	CA	0.0083	FM	0.0081	CA	0.0080
3	FIA	0.0077	FIA	0.0084	FIA	0.0078	CA	0.0081	FIA	0.0070
4	CO	0.0015	CO	0.0016	CO	0.0021	CO	0.0017	CO	0.0016
5	C	0.0010	C	0.0010	C	0.0008	C	0.0013	C	0.0010
6	L	0.0009	L	0.0009	L	0.0007	L	0.0008	L	0.0006
7	M	0.0002	SH	0.0003	SH	0.0002	M	0.0003	SE	0.0003
8	SH	0.0002	SE	0.0002	M	0.0002	SH	0.0002	M	0.0002
9	SE	0.0002	M	0.0002	SE	0.0002	SE	<0.0001	SH	0.0002

Finally, for medium release quantity, contributing action (*CA*) was the most influential, informative variable, based on 4 of the 5 networks. Failure mode (*FM*) was the second most influential variable for medium release quantity based on 4/5 networks, as shown in Table 69. Thus, based on the results in Table 68 and Table 69, contributing action and failure mode took on more influential roles for small and medium release quantity versus small and medium dollar loss. In contrast, causing object and failure item-area were more influential to dollar loss than to release quantity.

Table 69: Ranking of 'Medium' Release Quantity Parent Variables.

	Medium Release Quantity Category									
Rank	T1	Ranking Value	T2	Ranking Value	T3	Ranking Value	T4	Ranking Value	T5	Ranking Value
1	CA	0.0194	CA	0.0202	FM	0.0188	CA	0.0176	CA	0.0188
2	FM	0.0182	FM	0.0185	CA	0.0186	FM	0.0171	FM	0.0179
3	FIA	0.0068	FIA	0.0075	FIA	0.0064	FIA	0.0068	FIA	0.0060

Table 69 (continued).

4	CO	0.0051	CO	0.0058	CO	0.0058	CO	0.0050	CO	0.0059
5	L	0.0027	L	0.0025	L	0.0021	L	0.0024	L	0.0021
6	C	0.0025	C	0.0024	C	0.0020	C	0.0022	C	0.0018
7	SE	0.0002	SH	0.0003	SE	0.0002	M	0.0003	M	0.0002
8	M	0.0002	M	0.0002	M	0.0002	SH	0.0002	SE	0.0002
9	SH	0.0001	SE	0.0002	SH	0.0001	SE	0.0001	SH	0.0001

In order to demonstrate the analysis of the recommended policy changes for release quantity, contributing action will be used as an example. Based on the ranking procedure, this variable should be targeted in order to best impact medium release quantity. However, the category of contributing action to pursue depends on the particular category of release quantity targeted for reduction. For, although a medium release involves a greater quantity of material, it occurs less frequently than a small release (9.4% versus 87.9%, respectively). Since the improper loading and dropped category increases the probability of a medium release, as shown in Table 70, it should be the target of reduction efforts. However, since a loose fitting or valve increases the probability of a small release the most (0.911 versus 0.895 for "other"), it should be pursued if the goal is to best impact a small release quantity.

Table 70: Effects of Contributing Action on Release Quantity. (T1 network)

Contributing Action	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.879	0.094
-			
Improper Loading and Dropped	0.020	0.841	0.139
Other	0.031	0.895	0.075
Loose Fitting or Valve	0.033	0.911	0.056

The recommended policy changes for each parent variable of release quantity based on this same type of analysis and all five networks are summarized in Table 72. Note that for material type,

shift, and season, there are no recommended policy changes due to the limited information these variables provide. This is due to the small changes in the probability distribution of release quantity when evidence is introduced on these variables. Refer to Table 71 for an illustration of this using the midnight shift as an example. The distribution of release quantity given the midnight shift is nearly the same as the distribution given "no evidence," as seen by comparing these two rows in Table 71. The same is true given the day and twilight shifts.

Table 71: Effects of Shift on Release Quantity. (T1 network)

.Shift	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.879	0.094
-			
Midnight	0.028	0.879	0.093
Day	0.027	0.883	0.091
Twilight	0.028	0.874	0.098

The supporting probability distributions for the recommended policy changes in Table 72 are located in Table 86 through Table 90 in Appendix C.

Table 72: Recommended Policy Changes for Impacting Release Quantity.

Explanatory Variable	Targeted Categories
Contributing Action	To best impact small quantity, target loose fitting or valve. To best impact medium quantity, target improper loading and dropped.
Failure Item-Area	To best impact small quantity, target closure on top. To best impact medium quantity, target basic package material on top.
Failure Mode	To best impact small quantity, target other. To best impact medium quantity, target punctured and crushed.
Causing Object	To best impact small quantity, target none. To best impact medium quantity, target floor and water/liquid.
Container Type	To best impact small quantity, target bottle. There is limited information provided by fiber box.
Location	There is limited information for small quantity. To best impact medium quantity, target suburban/ commercial/ eastern.

Table 72 (continued).

Material Type	There is limited information provided by material type.
Shift	There is limited information provided by shift.
Season	There is limited information provided by season.

4.2.3 Potential Tactical Uses of the Bayesian Network for Decision Making

An important use of a Bayesian network is tactical, or operational, decision making based on either predictive or diagnostic inference. Predictive inference can take the form of various "what-if" analyses, in which the effects of one or more explanatory variables on the conditional probability distributions of the outcome variables are determined. A potential use of predictive "what-if" analysis at the Office of Hazardous Materials (OHM) is for exemption, or special permit, approvals. In this decision process, the OHM must evaluate individual requests for exemptions from the hazardous materials transportation regulations. Conversely, diagnostic inference may involve "what-if" analysis in the opposite direction from effects to explanations, for potential use in risk reduction analysis or investigation of occurrence spikes. In addition, *MAP*'s, or maximum aposteriori probabilities, may be used diagnostically to gain a basic understanding of an occurrence spike or accident scenario. A *MAP* characterizes an accident scenario by identifying the most likely joint state or combination of the explanatory variables given the outcome variable. These tactical uses of the Bayesian network by the OHM will be discussed further in the next sections.

4.2.3.1 What If Analysis

As a first step in starting a tactical analysis program, the Office of Hazardous Materials can use the Bayesian network developed in this research to obtain an understanding of the relationships between the variables in a hazardous materials unloading release, based on their top categories. Based on the face validation study, the OHM does not have a complete understanding of the

relationships between the dependent and independent variables in an unloading release. A basic understanding of the relationships can be obtained by performing various "what-if" analyses in both a predictive and diagnostic manner. Predictive analysis involves setting the value of one or more known, or observed, explanatory variables as evidence in the network. The effects on the conditional probability distributions of the remaining, or unknown, downstream variables are determined by updating, or evaluating, the network. For example, the OHM may wish to investigate the impact of time of day on hazmat releases during unloading. Specifically, if the time of day for unloading flammable liquids is changed from the midnight to the daytime shift, a predictive "what-if" analysis can identify the impact on the quantity released and the monetary damages.

Using diagnostic "what if" analysis, a Bayesian network can be used to determine the distributions of the explanatory variables to identify the categories that are most or least likely given the outcome. For example, as part of a risk reduction study, the OHM may wish to analyze the ideal situation of zero quantity released and compare it to a medium-quantity release to understand the factors involved. Specifically, the OHM may wish to evaluate a zero-release versus medium-release situation in which the bottom of the basic packaging material failed, in order to identify possible risk reduction alternatives. To do this, the known variables, namely zero release quantity and the bottom of basic packaging material, are set as evidence, and the values of the remaining variables are updated as the explanations. This can be compared to a medium-quantity release to understand the differences. To make this comparison, evidence is set of a medium-quantity release to determine any changes in the distribution of the explanatory factors.

Exemption/Special Permit Analysis The use of the Bayesian network model to analyze exemptions, or special permits, at the OHM is a potential application of predictive "what-if" analysis. For example, suppose an exemption is being considered for the relocation of a lab with infectious substances from Rockville, MD to Fairfax, VA, which are suburban, commercial locations in the greater Washington, DC area. An exemption is likely necessary in this case due to the nature of the material being transported. Within an exemption, certain variables may be fixed, such as material type or location. In this scenario, the material type and locations are fixed. However, there are certain variables that may be set or prescribed by the OHM as part of the exemption granting process in order to minimize the possibility of an undesirable outcome. For example, in the previous scenario, the time of day or the season of the year for the relocation could be prescribed by the OHM in the exemption that is granted. Also, the type of container or packaging may be a point of negotiation or additional control for the OHM.[228] Given these non-fixed variables, the OHM can use the Bayesian network and "what if" analysis to prescribe such variables to minimize the probability of the consequences. For example, if the lab were relocated during the midnight shift in the summer months versus the daytime shift in the winter months, how would the consequences be impacted? Also, what is the impact on the consequences of various containers for transporting infectious substances? These questions can be answered by setting the fixed as well as the prescribed variables as evidence and updating the probabilities of the unknown variables, including the consequences.

However, in its present form, the Bayesian network model is specific to unloading incidents and the top material and container types. These include corrosives and flammable liquids, and fiber boxes and bottles, respectively. In addition, people-related outcomes are not considered by the present model, although this is the biggest concern when granting exemptions. Thus,

considering the lab relocation scenario above, the present model could not be made specific to infectious substances or the types of containers often used to transport infectious substances, such as plastic bags. Exemptions can get very specific in terms of the factors involved. Given this constraint, the existing Bayesian network model could still be used during the exemption approval process to gain very general insight into the occurrence of hazmat incidents. For example, how do daytime, suburban incidents impact the monetary damages in unloading accidents and possibly other scenarios? Additionally, assume the OHM has a concern about potential inadequate handling of containers. Using the existing model, the impact of dropping an improperly-loaded container on monetary loss can be determined and applied in a general sense to other situations. However, taking a different approach, the present model could be expanded, or new models developed, in order to provide a decision tool that specifically considers certain variables or categories, such as the infectious material type, infectious-material containers, or human-related consequences.

Analysis of Occurrence Spikes In addition to performing risk reduction analysis, diagnostic "what-if" analysis of a Bayesian network can be used to investigate occurrence spikes reported to the OHM. Increases in the occurrence of particular hazmat releases are often reported to the OHM from the field. An example of an occurrence spike might be many loose closures on bottles during flight. The OHM currently investigates such reports by using the HMIRS to determine similar reported incidents and the associated shippers for direct communication with them.[229] To illustrate the use of "what-if" analysis for occurrence spikes, assume there is a sharp increase in the occurrence of medium-quantity releases in which the bottom of the basic package material failed. The OHM can begin its investigation by getting a basic understanding of the distributions of the unknown factors associated with the occurrence spike using "what-if"

analysis. Specifically, the known variables, namely medium release quantity and bottom of the packaging material, are set as evidence to obtain the distributions of the unknown variables. The degree of prevalence of various categories of the unknown variables provides insight into the accident situation. These results may be compared to those for a zero-quantity release for possible additional insight. If additional information or explanation becomes known about the occurrence spike, it can be set as evidence to determine the impact on the remaining variables. It's possible that additional information may make certain other explanatory variables more or less likely if the common outcome variable is known. Thus, known explanatory variables may "explain away" certain unknown explanatory variables, thereby making them more or less likely, given knowledge on a common outcome. In the example of loose closures during flight, the OHM may find it useful to know most likely categories for variables such as time of day, season of the year, or material type, which can be obtained using diagnostic style "what-if" analysis.

The OHM may wish to obtain an overall characterization of an occurrence spike by running a *MAP*. Using a *MAP*, the OHM can determine the most likely combination of variables surrounding the occurrence spike. For example, which combination of unknown variables is most likely given failure of the bottom of the basic packaging material and a medium release of material? A *MAP* serves to tell a story of the events surrounding an accident. It characterizes an accident and answers the question, "What does the accident look like?" As a demonstrated example of a *MAP*, a *MAP* of the container failure variables in stages two through four, which were most influential to medium dollar loss, was run. A *MAP* can be run using a subset of the explanatory variables, thereby increasing the combination likelihood and potential usefulness relative to using many explanatory variables.[230] In the case of a release in which medium dollar loss occurred, it's most likely that the container and its contents were improperly loaded and

dropped, thereby impacting the ground and encountering water or other liquid. In addition, the container was consequently punctured and crushed, leading to failure of the basic package material on the bottom of the container and the release of a small amount of material to the environment. This most likely scenario seems plausible. In the same way, *MAPs* can be used to gain insight into occurrence spikes. The *MAPs* for the variables in stages two through four given medium dollar loss across the five networks are summarized in Table 73. Each *MAP* had an occurrence probability of approximately 16.5%, and the categories of the *MAP*s were the same across the five networks.

Table 73: *MAP* of Most Influential Variables on Medium Dollar Loss.

	***MAP* by Network**				
Explanatory Variable	**T1**	**T2**	**T3**	**T4**	**T5**
CO	Floor and water/liquid	Floor and water/liquid	Floor and water/liquid	Floor and water/liquid	Floor and water/liquid
CA	Improper loading and dropped	Improper loading and dropped	Improper loading and dropped	Improper loading and dropped	Improper loading and dropped
FIA	Basic package material on bottom	Basic package material on bottom	Basic package material on bottom	Basic package material on bottom	Basic package material on bottom
FM	Punctured and crushed	Punctured and crushed	Punctured and crushed	Punctured and crushed	Punctured and crushed
RQ	Small	Small	Small	Small	Small

*MAP*s for medium release quantity were also run to characterize the worst case in terms of quantity released. The results are similar to those for dollar loss and are given in Table 74. Based on 4/5 of the networks, a medium-quantity release is most likely characterized by an improperly-loaded and dropped container and contents that impact the ground and encounter water. However, while the container is most likely punctured and crushed, it's the basic package

material on *top* of the container that fails in the case of medium release quantity. The Office of Hazardous Materials may find it useful to compare *MAP*s for insight into various events or problems, as was done here for dollar loss versus release quantity.

Table 74: *MAP* of Most Influential Variables on Medium Release Quantity.

	***MAP* by Network**				
Explanatory Variable	**T1**	**T2**	**T3**	**T4**	**T5**
CO	Floor and water/liquid	Floor and water/liquid	Floor and water/liquid	Floor and water/liquid	Floor and water/liquid
CA	Improper loading and dropped	Improper loading and dropped	Improper loading and dropped	other	Improper loading and dropped
FIA	Basic package material on top	Basic package material on top	Basic package material on top	Basic package material on bottom	Basic package material on top
FM	Punctured and crushed	Punctured and crushed	Punctured and crushed	other	Punctured and crushed

Another decision support query useful for occurrence spikes is the determination of the most influential variables given the occurrence of another variable. For example, suppose a risk assessment engineer is focusing on incidents involving the "floor and water/liquid" combination, perhaps because there has been a recent increase in this particular causing object. Given that an incident has the "floor and water" combination as its causing object, what variable should the engineer next pursue to best impact medium dollar loss? This can be determined in *GeNIe* by setting the "floor and water" combination as evidence in the Diagnostics module and then updating the rankings of the remaining variables. In this example, the engineer should focus on the contributing action that led to the failure, based on its being the next most influential variable in 5/5 networks. This query involved a conditional analysis between two variables. Using *GeNIe*, it is also possible to determine the optimal joint combination of two or more parent

variables for the occurrence of a particular category of dollar loss. This can be done in *GeNIe* by converting the Bayesian network to an influence diagram and assigning utilities to the categories of dollar loss.[231]

4.3 VALIDATION

Models can be validated in various ways to assess their accuracy and quality. Two forms of validation were performed to assess the hazardous materials release models developed in this research. First, five-fold cross validation was performed on the Bayesian network models, as discussed previously in section 4.1.4.1, yielding very similar accuracies across the five networks for predicting dollar loss and release quantity. In cross validating, a portion of the available data was retained for testing, and results predicted by the model were compared to actual values on the test records. Tests were conducted in both a forward and reverse direction, meaning that accuracy was assessed relative to both the outcome and explanatory variables. For the network in which dollar loss was the outcome variable, the accuracies associated with predicting the most likely and next most likely categories of dollar loss were approximately 70% and 23%, respectively, across the five networks, as discussed in 4.1.4.2. For the network in which release quantity was the outcome variable, the respective accuracies were approximately 87% and 10% across the five networks. Although there is not a standard for assessing prediction accuracies, these results appear reasonable given the possibility that other important variables were not part of the model.

Face validation is another method for assessing a model. With face validation, the reasonableness of the model is determined, along with the presence of any obvious flaws. In addition, inputs and outputs are also evaluated. A face validation study of the base structure of

the Bayesian network and its results was conducted with five engineers and scientists at the DoT's Office of Hazardous Materials in August 2005. The base structure included the temporal layout of the simplified variables and the five stages of a hazmat release. Each of the five members of the panel, who were familiar with hazmat transportation and the HMIRS, assessed the base structure as "reasonable." The entire association structure of the hazmat release network could not be assessed by this group due to its size. As a possible caution concerning the face validation, the panelists did not have an in depth knowledge of the HMIRS consistent with performing exploratory analyses of the data. Rather, the panelists' experiences with the data have been limited to reactive use in support of day-to-day operations and investigation of specific problems and issues.

A series of questions was posed to this panel to compare their knowledge with various results of the model. Although not all associations could be evaluated, the panelists were asked to identify the direct associations in stage one. Their predictions versus the model results for stage one were encouraging. All five panelists identified the material type to container type direct association, based on regulations surrounding usage of certain containers for certain materials. Three of the five panelists identified the season to material type association, based on their belief of seasonal usage of various materials. None of the panelists identified the shift to location association but were subsequently able to offer plausible explanations for this association.

The panelists were also asked to rank the explanatory variables based on their influence on both release quantity and dollar loss. Their predictions were compared to the results of the entropy-based ranking procedure in *GeNIe* as a means of face validating the results. A grid of the results showing the amount of agreement between the panelists' ranking and the model's ranking of each explanatory variable was compiled, as shown in Table 75 and Table 76. The

model's ranking was determined by considering the ranking results of each of the five networks, using "best of five" where necessary. In Table 75, the explanatory variables for release quantity are listed on the left side, in descending order of information value as ranked by *GeNIe.* The numbers across the top are used to record the panelists' ranking of a variable. Each cell contains the number of panelists having the particular response. For example, if two panelists ranked *CA* as third in terms of influence, a "2" would be entered in column 3 in the row for *CA*. Agreement between the panelists and the model is indicated by a large number of responses in the shaded area. This area was arbitrarily defined as each cell on the diagonal, which represents an exact match between panelist and model, plus one cell to the left and to the right. For medium release quantity, the number of total responses that fell within the shaded area was 19/45, or 42.2%, as illustrated in Table 75.

Table 75: Model vs. DoT Panelist Ranking of Variables for Medium Release Quantity.

Number of Responses											
			DoT Ranking								
Model Ranking			1	2	3	4	5	6	7	8	9
1	CA			1	1	2		1			
2	FM		2		2		1				
3	FIA			1	2			2			
4	CO					2	2	1			
5	L								5		
6	C		3	1		1					
7	SE						1		1	1	2
8	M			2			1	1		1	
9	SH								1	2	2

For medium dollar loss, the agreement between the model and the panelists was less, with 13 out of 50 (26%) responses in the shaded area, as shown in Table 76. In examining this table, the panelists' predictions for dollar loss tended to be opposite the model's predictions, as evidenced

in the large number of responses in the top right and bottom left corners of the grid in Table 76. Thus, the validation results for release quantity were more encouraging than those for dollar loss.

Table 76: Model vs. DoT Panelist Ranking of Variables for Medium Dollar Loss.

Number of Responses												
			DoT Ranking									
Model Ranking			1	2	3	4	5	6	7	8	9	10
1	CO							3		1	1	
2	FIA					1			3	1		
3	CA				1	1	1				1	1
4	FM				4		1					
5	L							1	2	2		
6	RQ		3	2								
7	SE							1		1	1	2
8	SH					1					3	1
9	C			1		2	2					
10	M		2	2			1					

In the face validation study, the panelists predicted different leading variables for dollar loss versus release quantity. For example, although the model ranked the container failure variables as most influential to both release quantity and dollar loss, the panelists ranked the container failure variables as more influential to release quantity than to dollar loss. This is shown by the greater number of entries in the upper left portion of Table 75 versus Table 76. Based on this, the panelists thought the events associated with container's failure were more important to the release quantity than to the monetary loss. In addition, *GeNIe* ranked material type *M* as one of the least influential variables to both release quantity and dollar loss. However, the panelists ranked material type as more influential to dollar loss than to release quantity. This is shown by the additional number of #1 and #2 panelist rankings for material type in Table 76 versus Table 75. In terms of the most influential variables, the panelists thought that release quantity and material type were most influential to medium dollar loss, as shown by the number of entries in

columns 1 and 2 in Table 76. This appears reasonable given that their frame of reference likely begins with the particular hazardous "material." However, the panelists thought that container type was most influential to medium release quantity, as given by the four entries in columns 1 and 2 in Table 75.

In addition to face validating the results of the ranking procedure, the panelists were asked to face validate *MAP* results for the two most influential explanatory variables for release quantity and dollar loss, as shown in Table 77. For medium dollar loss, the panelists were asked to identify the most likely combination of causing object *CO* and failure item-area *FIA*. Three of the five panelists predicted the combination identified by 5/5 networks of floor and water and basic package material on the bottom of the container. Likewise, for medium release quantity, four of the five panelists predicted 1) improperly loaded and dropped and 2) punctured and crushed as the most likely combination of its top variables contributing action *CA* and failure mode *FM*, as identified by 5/5 networks. Thus, the majority of panelists predicted the most likely combinations of the top variables for dollar loss and release quantity. The survey questionnaire used for the face validation is shown in Figure 23 in Appendix D.

Table 77: Model vs. DoT Panelist MAP Results.

	Most Influential		
Outcome	**Explanatory Variables**	**Categories**	**Panelists Predicted**
Medium Dollar Loss	1. CO	Floor and water	3/5
	2. FIA	Basic package material on bottom	
Medium Release Quantity	1. CA	Improperly loaded and dropped	4/5
	2. FM	Punctured and crushed	

As an overall indication of face validity, the panelists felt this research should be considered for application at the DoT. They felt the graphical aspect of the model was helpful in problem visualization.

5.0 CONCLUSIONS

Using a methodology for categorical variables involving simplification, measurement of associations, and construction of a Bayesian network, a large database was analyzed to build a data-congruent decision model of an engineering policy problem. The methodology employs a combination of existing categorical data analysis techniques to develop the qualitative structure of the decision model. Specifically, new, simplifying variables were developed using latent class analysis, and measurement of associations was accomplished through loglinear modeling, together forming a three step modeling approach.

5.1 CONTRIBUTION - METHODOLOGY

This methodology for analyzing a large categorical database was developed as part of an initial data modeling effort using a database within an unexplored area (hazardous materials releases). It is a methodology that can be used to "get one's hands around" a complex database for which few or no modeling efforts have taken place in the past. With this methodology, data-driven analysis techniques can be combined with subject matter knowledge to enhance the usual decision modeling process. The first stage of the methodology focuses the modeler on the top categories and variables as well as generalized versions of the variables, which is necessary for developing an initial, or first-generation, data model of an event or system. In addition, after choosing top categories and variables by way of Pareto-style analysis and generalizing variables

where possible, new simplifying variables are created using the data modeling technique of latent class analysis. Thus, the simplification consists of a combination of category elimination, which may be somewhat-subjective and benefit from subject matter expertise, and data modeling, which is very conducive to a first time analysis of the variables.

Similarly, the second stage of the methodology is also very conducive to an initial modeling effort within a subject area. It represents an enhancement to more-traditional methods for building the relationship structure of a decision model. It allows relationships to be determined based on data-driven associations as well as expert opinion where available. It supports the modeling effort in the absence of prior theory or empirical analysis of the database or subject area. The third stage of the methodology entails the construction of a traditional decision model consisting of only random variables. The decision model, or Bayesian network, allows the combination of probability theory and information theory in identifying the most influential variables and desirable changes for them relative to a chosen outcome variable. In addition to such strategic diagnostic analysis, more tactical-style analyses can also be made using the decision model, including "what-if" and sensitivity analysis. Sensitivity analysis is useful in cases where the conditional probabilities are based on expert opinion or perhaps a smaller amount of data. The tactical analysis can be either predictive or diagnostic, including "what-if" analysis for decisions such as exemption approvals. Tactical analysis may also include the creation of *MAPs*, or maximum aposteriori probabilities, for a basic understanding of accident scenarios and situations, such as spikes in the occurrence of certain events.

This methodology may be a good candidate for application within an area such as Homeland Security given its relative newness. It's possible to envision that certain categorical variables, such as gender or country of origin, may influence a suspicion or threat level, which eventually

may impact human health or life and agriculture. In addition, another area conducive to the development of system or network models based on a large amount of categorical data is the health arena. For example, the National Center for Health Statistics (NCHS) collects interrelated data such as gender, parts of the body, health conditions, and reasons for avoiding medical care or testing, which influence a person's health. Although the NCHS conducts a large amount of empirical research, network or systems-style models are not commonly or formally used there currently. Data-driven Bayesian networks have not penetrated their approach to proactive decision analysis, although there is interest in them.

In conclusion, the methodology developed as part of this dissertation is a general, flexible approach that can be applied to areas having large amounts of categorical data. It is focused on data-driven development and therefore may be particularly useful for less-explored areas. It can be supported and enhanced by the amount of subject matter knowledge desired.

5.2 CONTRIBUTION - HAZMAT RELEASE LITERATURE

Using the decision model, the most influential variables relative to dollar loss and release quantity were determined for a hazmat unloading accident. For both of these outcome variables, the most influential variables were the container failure variables. Specifically, for a small and medium release quantity, the top three influential variables were the action contributing to the failure of the container *(CA)*, the item-area that failed *(FIA)*, and the mode of failure *(FM)*. For a medium dollar loss, the leading variables in order of influence were as follows: object causing the failure *(CO)*, the item-area *(FIA)*, and the action contributing to the failure *(CA)*. For small dollar loss, release quantity *(RQ)* was third in terms of influence, versus contributing action (*CA)* for medium dollar loss. In addition, the recommended operational or policy changes for each

explanatory variable to decrease the probability of release quantity and dollar loss were determined. They were determined based on the effects of each category of the explanatory variable on the probability distributions of the outcome variable using the decision model. For example, for causing object, the best change to pursue to decrease the probability for dollar loss is a reduction in incidents involving a combination of the floor and water/liquid as the causing object. For contributing action, a top variable for release quantity, the best changes are reductions in incidents involving a loose fitting or valve as well as improperly loaded and dropped containers, depending on whether a small or medium release quantity is targeted, respectively. Five Bayesian networks were built so that five-fold cross validation could be done. The five sets of cross validation results closely agreed and were reasonable, with approximately 70% accuracy in predicting dollar loss and 87% accuracy in predicting release quantity. Thus, the results regarding the influential variables were based on five separate networks. The results of the face validation study indicate an opportunity for use of a Bayesian network model at the OHM for providing insight to both strategic and tactical decisions.

The hazmat release database had not previously been analyzed in this manner by the DoT, due in part to the lack of penetration of categorical analysis methods into the engineering arena. In addition, the hazardous materials transportation literature has been focused on minimum risk routing and calculation of risk in quantitative risk assessment studies. There has been a lack of focus on post-accident, exploratory use of the incident data, in particular to identify critical variables and policy changes for them. There has also been a lack of focus in the literature on transportation support activities, such as container unloading, despite the fact that the majority of incidents occur at this point. Thus, this research also contributes to the literature in terms of exploring hazmat unloading activity.

5.3 FUTURE RESEARCH

Opportunities for future study exist both in terms of the hazardous materials release problem as well as other areas in which the methodology can be applied. Within the hazardous materials problem, additional types of consequences should be studied, especially the non-material consequences, such as injuries, deaths, and evacuations. The Office of Hazardous Materials tends to focus more on human-related versus material consequences. In addition to studying other types of consequences, additional phases of the transportation process should be studied, including loading, transport, and storage. As mentioned previously, the container failure variable "other" should be eliminated from future analysis. Although "other" is associated with many incidents and consequences, it does not provide definitive information for problem analysis. On the new incident form instituted by the DoT in 2004, the "other" category has been eliminated. Therefore, the database based on the new incident form could be the subject of future research.

As a different approach to developing the decision model for the hazmat release problem, the learning module within *GeNIe* could be applied to the data in order to build the association structure. It would be useful to compare the association structure as determined from the three step modeling approach with the structure as determined using the learning module. The learning algorithm provides an alternative means of establishing the model's structure.

Due to modeling constraints, some decisions were made concerning categories or variables to eliminate during the initial phase of the simplification. The Pareto principle could not be followed and applied for all variables or sets of variables. Unfortunately, the sensitivity of the model and results to this constraint is unknown. This is an area for future research and may lead to later-generation models of a hazardous materials release during unloading.

APPENDIX A

ADDITIONAL ANALYSES - STAGE ONE

Table 78: Stage One Residual and Component L^2 Results.

		RESIDUAL			COMPONENT			
Run	Model	L^2	df	p	L^2	df	p	Significant
1	[SE][M][C][SH][L] (mutual independence)	125.2625	87	0.0045				
	[M C]	107.5191	86	0.0581	17.7434	1	0.0001	Y
	[SE M]	111.9227	84	0.0226	13.3398	3	0.0040	Y
	[SH L]	117.808	85	0.0107	7.4545	2	0.0241	Y
	[L C]	124.3989	86	0.0043	0.8636	1	0.3527	
	[L M]	124.9873	86	0.0039	0.2752	1	0.5999	
	[L SE]	123.2545	84	0.0034	2.0080	3	0.5707	
	[SE C]	121.9592	84	0.0043	3.3033	3	0.3472	
	[SH C]	125.0906	85	0.0031	0.1719	2	0.9176	
	[SH M]	120.9113	85	0.0064	4.3512	2	0.1135	
	[SH SE]	116.2602	81	0.0062	9.0023	6	0.1734	
2	[SE][M][C][SH][L] (mutual independence)	114.1234	87	0.0272				
	[M C]	94.4533	86	0.2498	19.6701	1	0.0001	Y
	[SE M]	104.4828	84	0.0645	9.6406	3	0.0219	Y
	[SH L]	104.0196	85	0.0789	10.1038	2	0.0064	Y
	[L C]	113.703	86	0.0244	0.4204	1	0.5167	
	[L M]	113.6544	86	0.0246	0.4690	1	0.4934	
	[L SE]	111.531	84	0.0239	2.5924	3	0.4588	
	[SE C]	110.4271	84	0.0282	3.6963	3	0.2962	
	[SH C]	113.9375	85	0.0198	0.1859	2	0.9112	
	[SH M]	113.0544	85	0.0226	1.0690	2	0.5860	
	[SH SE]	106.7612	81	0.0292	7.3622	6	0.2886	
3	[SE][M][C][SH][L] (mutual independence)	128.0326	87	0.0028				
	[M C]	113.3556	86	0.0257	14.6770	1	0.0001	Y
	[SE M]	114.7256	84	0.0146	13.3070	3	0.0040	Y
	[SH L]	120.4478	85	0.0069	7.5848	2	0.0225	Y

Table 78 (continued).

	[L C]	127.6727	86	0.0024	0.3599	1	0.5486	
	[L M]	122.4895	86	0.0060	5.5431	1	0.0186	Y
	[L SE]	126.0251	84	0.0021	2.0075	3	0.5709	
	[SE C]	127.3769	84	0.0016	0.6557	3	0.8836	
	[SH C]	127.9781	85	0.0018	0.0545	2	0.9731	
	[SH M]	125.9098	85	0.0026	2.1228	2	0.3460	
	[SH SE]	113.5523	81	0.0099	14.4803	6	0.0247	Y
4	[SE][M][C][SH][L] (mutual independence)	121.5583	87	0.0085				
	[M C]	110.5451	86	0.0385	11.0132	1	0.0009	Y
	[SE M]	104.083	84	0.0679	17.4753	3	0.0006	Y
	[SH L]	116.4828	85	0.0133	5.0755	2	0.0790	
	[L C]	121.2689	86	0.0074	0.2894	1	0.5906	
	[L M]	120.604	86	0.0082	0.9543	1	0.3286	
	[L SE]	119.7449	84	0.0064	1.8134	3	0.6120	
	[SE C]	117.9295	84	0.0086	3.6288	3	0.3044	
	[SH C]	121.2985	85	0.0060	0.2598	2	0.8782	
	[SH M]	121.302	85	0.0060	0.2563	2	0.8797	
	[SH SE]	110.1527	81	0.0173	11.4056	6	0.0766	
5	[SE][M][C][SH][L] (mutual independence)	141.63	87	0.0002				
	[M C]	115.7958	86	0.0178	25.8342	1	0.0001	Y
	[SE M]	131.5206	84	0.0007	10.1094	3	0.0177	Y
	[SH L]	125.0913	85	0.0031	16.5387	2	0.0003	Y
	[L C]	141.5651	86	0.0002	0.0649	1	0.7989	
	[L M]	133.7182	86	0.0008	7.9118	1	0.0049	Y
	[L SE]	135.4763	84	0.0003	6.1537	3	0.1044	
	[SE C]	137.6344	84	0.0002	3.9956	3	0.2619	
	[SH C]	141.5397	85	0.0001	0.0903	2	0.9559	
	[SH M]	137.4482	85	0.0003	4.1818	2	0.1236	
	[SH SE]	138.9105	81	0.00007	2.7195	6	0.8431	
6	[SE][M][C][SH][L] (mutual independence)	118.2752	87	0.0145				
	[L M]	115.9667	86	0.0173	2.3085	1	0.1287	
	[SH SE]	113.0043	81	0.0109	5.2709	6	0.5096	
	[SH M]	113.6011	85	0.0208	4.6741	2	0.0966	
7	[SE][M][C][SH][L] (mutual independence)	125.8119	87	0.0041				
	[L M]	125.5721	86	0.0035	0.2398	1	0.6244	
	[SH SE]	122.9788	81	0.0018	2.8331	6	0.8295	
	[SH M]	124.5677	85	0.0034	1.2442	2	0.5368	

Table 79: Stage One Correction Procedure Matrices.

LOCATION

Transition Matrix	
0.998213	0.001890
0.001930	0.998174

Inverse Transition Matrix	
1.001794	-0.001894
-0.001933	1.001833

UNCORRECTED			CORRECTED			CORRECTED (ROUNDED)	
33	43		33.02382193	42.98378703		33	43
11	5		10.98996722	5.01163461		11	5
52	54		52.0110443	53.99956812		52	54
20	16		19.99506358	16.0085406		20	16
33	44		33.02574766	43.98196143		33	44
8	9		8.003032316	8.99866968		8	9
68	66		68.00555458	66.00786108		68	66
25	29		25.011161	28.99424535		25	29
33	27		32.99301035	27.01299663		33	27
8	8		8.001106592	8.00049528		8	8
31	40		31.02161956	39.98548881		31	40
11	13		11.00537301	12.99702981		11	13
23	28		23.01281007	27.99229593		23	28
6	5		5.99890422	5.00219706		6	5
36	49		36.03001408	48.97849596		36	49
13	16		13.00757538	15.99532803		13	16
46	36		45.98710566	36.02110386		46	36
10	11		10.00330896	10.9987935		10	11
55	57		55.01145927	56.99975385		55	57
23	17		22.99162711	17.01237753		23	17
34	34		34.00470302	34.00210494		34	34
9	3		8.989690572	3.01151079		9	3
73	53		72.97158317	53.04103143		73	53
22	24		22.00689458	23.99771082		22	24
27	27		27.00373475	27.00167157		27	27
4	15		4.02173626	14.98016604		4	15
50	26		49.96069882	26.0469099		50	26
15	18		15.00785203	17.99545185		15	18
34	28		33.99314867	28.01305854		34	28
4	9		4.010181916	8.99111964		4	9
38	44		38.01681066	43.99139898		38	44
16	11		15.99258456	11.01011856		16	11
41	29		40.9825626	29.02444551		41	29
4	13		4.017884812	12.98381724		4	13
49	42		48.99329781	42.01581279		49	42
10	14		10.00908614	13.9933167		10	14
29	29		29.0040114	29.00179539		29	29
13	10		12.99602104	10.00628163		13	10
51	40		50.98587156	40.02323901		51	40

Table 79 (continued).

23	22		23.00125573	22.00324953		23	22
44	32		43.98297757	32.02463124		44	32
8	1		7.987626524	1.01327448		8	1
43	35		42.99054214	35.01726693		43	35
14	11		13.99615936	11.00634354		14	11
25	27		25.00730955	26.99789655		25	27
8	9		8.003032316	8.99866968		8	9
37	25		36.9820093	25.02419787		37	25
13	13		13.00179821	13.00080483		13	13

APPENDIX B

ADDITIONAL ANALYSES - STAGES FOUR AND FIVE

Table 80: Latent Variable Transition Matrices.

LOCATION

Transition Matrix	
0.998213	0.001890
0.001930	0.998174

Inverse Transition Matrix	
1.001794	-0.001894
-0.001933	1.001833

CONTRIBUTING ACTION

Transition Matrix		
0.996606	0.002320	0.001070
0.003110	0.996462	0.000422
0.002660	0.000781	0.996557

Inverse Transition Matrix		
1.003416	-0.002337	-0.001079
-0.003135	1.003558	-0.000422
-0.002679	-0.000780	1.003459

CAUSING OBJECT

Transition Matrix		
0.998978	0.000743	0.000330
0.001410	0.998125	0.000467
0.000617	0.000519	0.998823

Inverse Transition Matrix		
1.001024	-0.000745	-0.000331
-0.001412	1.001880	-0.000468
-0.000618	-0.000520	1.001179

FAILURE MODE

Transition Matrix	
0.998811	0.001190
0.001810	0.998190

Inverse Transition Matrix	
1.001193	-0.001193
-0.001816	1.001816

FAILURE ITEM-AREA

Transition Matrix		
0.998055	0.001210	0.000737
0.002220	0.997540	0.000245
0.001500	0.000278	0.998220

Inverse Transition Matrix		
1.001953	-0.001214	-0.000739
-0.002225	1.002469	-0.000245
-0.001506	-0.000278	1.001784

APPENDIX C

ADDITIONAL ANALYSES - BAYESIAN NETWORKS

Table 81: Dollar Loss Distribution by Explanatory Variable. (T1 Network)

Causing Object	Dollar Loss		
	ZERO	SMALL	MEDIUM
no evidence	0.167	0.718	0.116
Floor and water/ liquid	0.083	0.770	0.147
None	0.199	0.682	0.119
Other	0.364	0.607	0.029
Contributing Action	Dollar Loss		
no evidence	0.167	0.718	0.116
Improper Loading and Dropped	0.161	0.680	0.158
Other	0.180	0.738	0.081
Loose Fitting or Valve	0.124	0.731	0.146
Location	Dollar Loss		
no evidence	0.167	0.718	0.116
Suburban/ Commercial/ Eastern	0.135	0.717	0.148
Urban/ Industrial or Commercial/ Eastern or Western	0.200	0.719	0.082
Season	Dollar Loss		
no evidence	0.167	0.718	0.116
Spring	0.173	0.711	0.116
Summer	0.161	0.744	0.095
Fall	0.171	0.683	0.146
Winter	0.163	0.724	0.113
Container Type	Dollar Loss		
no evidence	0.167	0.718	0.116
Fiber Box	0.172	0.721	0.107
Bottle	0.151	0.707	0.141

Failure Item-Area	Dollar Loss		
	ZERO	SMALL	MEDIUM
no evidence	0.167	0.718	0.116
Basic Package Material on Top	0.289	0.648	0.063
Basic Package Material on Bottom	0.064	0.783	0.154
Closure on Top	0.103	0.737	0.160
Release Quantity	Dollar Loss		
no evidence	0.167	0.718	0.116
Zero	0.347	0.559	0.094
Small	0.160	0.730	0.110
Medium	0.175	0.647	0.178
Failure Mode	Dollar Loss		
no evidence	0.167	0.718	0.116
Other	0.176	0.733	0.092
Punctured and Crushed	0.146	0.683	0.172
Shift	Dollar Loss		
no evidence	0.167	0.718	0.116
Midnight	0.174	0.720	0.106
Day	0.163	0.702	0.135
Twilight	0.161	0.735	0.104

Material Type	Dollar Loss		
no evidence	0.167	0.718	0.116
Flammable Liquids	0.192	0.704	0.104
Corrosives	0.149	0.727	0.124

Table 82: Dollar Loss Distribution by Explanatory Variable. (T2 Network)

Causing Object	Dollar Loss		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.168	0.716	0.116
Floor and water/ liquid	0.085	0.768	0.148
None	0.199	0.685	0.116
Other	0.364	0.605	0.031
Contributing Action	**Dollar Loss**		
no evidence	0.168	0.716	0.116
Improper Loading and Dropped	0.164	0.679	0.157
Other	0.182	0.736	0.082
Loose Fitting or Valve	0.121	0.736	0.143
Location	**Dollar Loss**		
no evidence	0.168	0.716	0.116
Suburban/ Commercial/ Eastern	0.137	0.715	0.148
Urban/ Industrial or Commercial/ Eastern or Western	0.200	0.718	0.082
Season	**Dollar Loss**		
no evidence	0.168	0.716	0.116
Spring	0.174	0.710	0.116
Summer	0.167	0.739	0.094
Fall	0.162	0.689	0.149
Winter	0.167	0.720	0.113
Container Type	**Dollar Loss**		
no evidence	0.168	0.716	0.116
Fiber Box	0.174	0.719	0.108
Bottle	0.150	0.709	0.141

Failure Item-Area	Dollar Loss		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.168	0.716	0.116
Basic Package Material on Top	0.290	0.647	0.063
Basic Package Material on Bottom	0.065	0.780	0.155
Closure on Top	0.105	0.738	0.156
Release Quantity	**Dollar Loss**		
no evidence	0.168	0.716	0.116
Zero	0.365	0.551	0.084
Small	0.161	0.730	0.109
Medium	0.176	0.640	0.184
Failure Mode	**Dollar Loss**		
no evidence	0.168	0.716	0.116
Other	0.177	0.731	0.092
Punctured and Crushed	0.146	0.682	0.172
Shift	**Dollar Loss**		
no evidence	0.168	0.716	0.116
Midnight	0.173	0.722	0.105
Day	0.163	0.701	0.135
Twilight	0.167	0.728	0.105

Material Type	Dollar Loss		
no evidence	0.168	0.716	0.116
Flammable Liquids	0.192	0.706	0.102
Corrosives	0.151	0.724	0.125

Table 83: Dollar Loss Distribution by Explanatory Variable. (T3 Network)

Causing Object	Dollar Loss		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.169	0.717	0.114
Floor and water/ liquid	0.085	0.770	0.145
None	0.203	0.676	0.121
Other	0.368	0.606	0.026
Contributing Action	**Dollar Loss**		
no evidence	0.169	0.717	0.114
Improper Loading and Dropped	0.163	0.681	0.155
Other	0.184	0.737	0.079
Loose Fitting or Valve	0.124	0.726	0.150
Location	**Dollar Loss**		
no evidence	0.169	0.717	0.114
Suburban/ Commercial/ Eastern	0.137	0.717	0.146
Urban/ Industrial or Commercial/ Eastern or Western	0.202	0.717	0.081
Season	**Dollar Loss**		
no evidence	0.169	0.717	0.114
Spring	0.176	0.710	0.114
Summer	0.167	0.739	0.093
Fall	0.170	0.686	0.144
Winter	0.161	0.726	0.113
Container Type	**Dollar Loss**		
no evidence	0.169	0.717	0.114
Fiber Box	0.175	0.720	0.105
Bottle	0.150	0.707	0.143

Failure Item-Area	Dollar Loss		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.169	0.717	0.114
Basic Package Material on Top	0.294	0.646	0.060
Basic Package Material on Bottom	0.064	0.784	0.152
Closure on Top	0.104	0.735	0.162
Release Quantity	**Dollar Loss**		
no evidence	0.169	0.717	0.114
Zero	0.348	0.560	0.092
Small	0.162	0.730	0.108
Medium	0.183	0.641	0.176
Failure Mode	**Dollar Loss**		
no evidence	0.169	0.717	0.114
Other	0.179	0.731	0.090
Punctured and Crushed	0.147	0.683	0.170
Shift	**Dollar Loss**		
no evidence	0.169	0.717	0.114
Midnight	0.177	0.721	0.102
Day	0.162	0.703	0.135
Twilight	0.168	0.730	0.103

Material Type	Dollar Loss		
no evidence	0.169	0.717	0.114
Flammable Liquids	0.193	0.705	0.102
Corrosives	0.153	0.725	0.122

Table 84: Dollar Loss Distribution by Explanatory Variable. (T4 Network)

Causing Object	Dollar Loss		
	ZERO	SMALL	MEDIUM
no evidence	0.167	0.714	0.118
Floor and water/ liquid	0.085	0.764	0.151
None	0.200	0.681	0.119
Other	0.362	0.607	0.031
Contributing Action	**Dollar Loss**		
no evidence	0.167	0.714	0.118
Improper Loading and Dropped	0.160	0.678	0.162
Other	0.183	0.734	0.083
Loose Fitting or Valve	0.125	0.727	0.148
Location	**Dollar Loss**		
no evidence	0.167	0.714	0.118
Suburban/ Commercial/ Eastern	0.134	0.716	0.150
Urban/ Industrial or Commercial/ Eastern or Western	0.203	0.712	0.085
Season	**Dollar Loss**		
no evidence	0.167	0.714	0.118
Spring	0.179	0.705	0.116
Summer	0.161	0.741	0.098
Fall	0.165	0.681	0.154
Winter	0.164	0.723	0.113
Container Type	**Dollar Loss**		
no evidence	0.167	0.714	0.118
Fiber Box	0.174	0.716	0.110
Bottle	0.148	0.707	0.145

Failure Item-Area	Dollar Loss		
	ZERO	SMALL	MEDIUM
no evidence	0.167	0.714	0.118
Basic Package Material on Top	0.292	0.643	0.064
Basic Package Material on Bottom	0.062	0.780	0.158
Closure on Top	0.104	0.735	0.161
Release Quantity	**Dollar Loss**		
no evidence	0.167	0.714	0.118
Zero	0.360	0.547	0.092
Small	0.159	0.729	0.112
Medium	0.186	0.626	0.188
Failure Mode	**Dollar Loss**		
no evidence	0.167	0.714	0.118
Other	0.177	0.730	0.093
Punctured and Crushed	0.145	0.678	0.177
Shift	**Dollar Loss**		
no evidence	0.167	0.714	0.118
Midnight	0.174	0.719	0.107
Day	0.162	0.696	0.141
Twilight	0.166	0.730	0.104

Material Type	Dollar Loss		
no evidence	0.167	0.714	0.118
Flammable Liquids	0.193	0.703	0.104
Corrosives	0.150	0.722	0.128

Table 85: Dollar Loss Distribution by Explanatory Variable. (T5 Network)

Causing Object	Dollar Loss		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.168	0.715	0.117
Floor and water/ liquid	0.086	0.766	0.148
None	0.201	0.679	0.120
Other	0.363	0.605	0.031
Contributing Action	**Dollar Loss**		
no evidence	0.168	0.715	0.117
Improper Loading and Dropped	0.165	0.677	0.158
Other	0.181	0.737	0.082
Loose Fitting or Valve	0.123	0.727	0.150
Location	**Dollar Loss**		
no evidence	0.168	0.715	0.117
Suburban/ Commercial/ Eastern	0.136	0.714	0.150
Urban/ Industrial or Commercial/ Eastern or Western	0.201	0.716	0.083
Season	**Dollar Loss**		
no evidence	0.168	0.715	0.117
Spring	0.176	0.709	0.115
Summer	0.165	0.738	0.097
Fall	0.169	0.683	0.148
Winter	0.161	0.722	0.117
Container Type	**Dollar Loss**		
no evidence	0.168	0.715	0.117
Fiber Box	0.174	0.718	0.108
Bottle	0.151	0.706	0.144

Failure Item-Area	Dollar Loss		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.168	0.715	0.117
Basic Package Material on Top	0.291	0.646	0.063
Basic Package Material on Bottom	0.064	0.780	0.156
Closure on Top	0.106	0.732	0.162
Release Quantity	**Dollar Loss**		
no evidence	0.168	0.715	0.117
Zero	0.357	0.551	0.092
Small	0.161	0.728	0.110
Medium	0.176	0.637	0.187
Failure Mode	**Dollar Loss**		
no evidence	0.168	0.715	0.117
Other	0.176	0.732	0.093
Punctured and Crushed	0.151	0.676	0.173
Shift	**Dollar Loss**		
no evidence	0.168	0.715	0.117
Midnight	0.176	0.720	0.104
Day	0.161	0.701	0.138
Twilight	0.166	0.726	0.108

Material Type	Dollar Loss		
no evidence	0.168	0.715	0.117
Flammable Liquids	0.192	0.705	0.103
Corrosives	0.152	0.721	0.127

Table 86: Release Quantity Distribution by Explanatory Variable. (T1 Network)

Causing Object	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.879	0.094
Floor and water/ liquid	0.017	0.876	0.106
None	0.033	0.901	0.067
Other	0.050	0.868	0.081

Contributing Action	Release Quantity		
no evidence	0.028	0.879	0.094
Improper Loading and Dropped	0.020	0.841	0.139
Other	0.031	0.895	0.075
Loose Fitting or Valve	0.033	0.911	0.056

Location	Release Quantity		
no evidence	0.028	0.879	0.094
Suburban/ Commercial/ Eastern	0.024	0.870	0.105
Urban/ Industrial or Commercial/ Eastern or Western	0.031	0.887	0.082

Season	Release Quantity		
no evidence	0.028	0.879	0.094
Spring	0.032	0.876	0.092
Summer	0.023	0.880	0.097
Fall	0.027	0.884	0.089
Winter	0.029	0.874	0.097

Container Type	Release Quantity		
no evidence	0.028	0.879	0.094
Fiber Box	0.026	0.874	0.100
Bottle	0.032	0.894	0.074

Failure Item-Area	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.879	0.094
Basic Package Material on Top	0.037	0.852	0.111
Basic Package Material on Bottom	0.018	0.892	0.090
Closure on Top	0.026	0.916	0.058

Failure Mode	Release Quantity		
no evidence	0.028	0.879	0.094
Other	0.030	0.897	0.073
Punctured and Crushed	0.021	0.836	0.143

Shift	Release Quantity		
no evidence	0.028	0.879	0.094
Midnight	0.028	0.879	0.093
Day	0.027	0.883	0.091
Twilight	0.028	0.874	0.098

Material Type	Release Quantity		
no evidence	0.028	0.879	0.094
Flammable Liquids	0.028	0.874	0.098
Corrosives	0.027	0.882	0.091

Table 87: Release Quantity Distribution by Explanatory Variable. (T2 Network)

Causing Object	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.878	0.095
Floor and water/ liquid	0.017	0.875	0.108
None	0.034	0.901	0.065
Other	0.050	0.868	0.082

Contributing Action	Release Quantity		
no evidence	0.028	0.878	0.095
Improper Loading and Dropped	0.019	0.841	0.140
Other	0.032	0.893	0.076
Loose Fitting or Valve	0.035	0.913	0.053

Location	Release Quantity		
no evidence	0.028	0.878	0.095
Suburban/ Commercial/ Eastern	0.025	0.870	0.106
Urban/ Industrial or Commercial/ Eastern or Western	0.031	0.887	0.083

Season	Release Quantity		
no evidence	0.028	0.878	0.095
Spring	0.033	0.873	0.094
Summer	0.023	0.880	0.097
Fall	0.026	0.884	0.090
Winter	0.029	0.875	0.096

Container Type	Release Quantity		
no evidence	0.028	0.878	0.095
Fiber Box	0.026	0.873	0.101
Bottle	0.031	0.894	0.075

Failure Item-Area	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.878	0.095
Basic Package Material on Top	0.037	0.850	0.113
Basic Package Material on Bottom	0.018	0.891	0.091
Closure on Top	0.026	0.917	0.057

Failure Mode	Release Quantity		
no evidence	0.028	0.878	0.095
Other	0.031	0.896	0.073
Punctured and Crushed	0.020	0.836	0.144

Shift	Release Quantity		
no evidence	0.028	0.878	0.095
Midnight	0.028	0.875	0.098
Day	0.027	0.884	0.089
Twilight	0.028	0.874	0.098

Material Type	Release Quantity		
no evidence	0.028	0.878	0.095
Flammable Liquids	0.028	0.873	0.099
Corrosives	0.027	0.881	0.092

Table 88: Release Quantity Distribution by Explanatory Variable. (T3 Network)

Causing Object	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.879	0.093
Floor and water/ liquid	0.018	0.877	0.106
None	0.034	0.904	0.062
Other	0.053	0.864	0.082

Contributing Action	Release Quantity		
no evidence	0.028	0.879	0.093
Improper Loading and Dropped	0.020	0.844	0.136
Other	0.032	0.893	0.075
Loose Fitting or Valve	0.034	0.912	0.054

Location	Release Quantity		
no evidence	0.028	0.879	0.093
Suburban/ Commercial/ Eastern	0.025	0.872	0.103
Urban/ Industrial or Commercial/ Eastern or Western	0.032	0.886	0.082

Season	Release Quantity		
no evidence	0.028	0.879	0.093
Spring	0.034	0.873	0.093
Summer	0.023	0.880	0.097
Fall	0.029	0.883	0.088
Winter	0.029	0.878	0.092

Container Type	Release Quantity		
no evidence	0.028	0.879	0.093
Fiber Box	0.027	0.874	0.099
Bottle	0.032	0.893	0.076

Failure Item-Area	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.879	0.093
Basic Package Material on Top	0.039	0.852	0.109
Basic Package Material on Bottom	0.018	0.892	0.090
Closure on Top	0.026	0.916	0.058

Failure Mode	Release Quantity		
no evidence	0.028	0.879	0.093
Other	0.032	0.896	0.072
Punctured and Crushed	0.021	0.837	0.143

Shift	Release Quantity		
no evidence	0.028	0.879	0.093
Midnight	0.029	0.878	0.093
Day	0.026	0.884	0.090
Twilight	0.030	0.873	0.096

Material Type	Release Quantity		
no evidence	0.028	0.879	0.093
Flammable Liquids	0.029	0.874	0.096
Corrosives	0.028	0.882	0.090

Table 89: Release Quantity Distribution by Explanatory Variable. (T4 Network)

Causing Object	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.878	0.094
Floor and water/ liquid	0.017	0.876	0.106
None	0.034	0.901	0.065
Other	0.051	0.865	0.084

Contributing Action	Release Quantity		
no evidence	0.028	0.878	0.094
Improper Loading and Dropped	0.020	0.844	0.136
Other	0.032	0.891	0.076
Loose Fitting or Valve	0.031	0.913	0.056

Location	Release Quantity		
no evidence	0.028	0.878	0.094
Suburban/ Commercial/ Eastern	0.025	0.870	0.105
Urban/ Industrial or Commercial/ Eastern or Western	0.031	0.886	0.083

Season	Release Quantity		
no evidence	0.028	0.878	0.094
Spring	0.031	0.877	0.092
Summer	0.025	0.881	0.095
Fall	0.028	0.880	0.092
Winter	0.028	0.873	0.099

Container Type	Release Quantity		
no evidence	0.028	0.878	0.094
Fiber Box	0.027	0.872	0.100
Bottle	0.029	0.895	0.076

Failure Item-Area	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.878	0.094
Basic Package Material on Top	0.038	0.850	0.112
Basic Package Material on Bottom	0.019	0.892	0.089
Closure on Top	0.024	0.916	0.059

Failure Mode	Release Quantity		
no evidence	0.028	0.878	0.094
Other	0.031	0.895	0.074
Punctured and Crushed	0.020	0.838	0.142

Shift	Release Quantity		
no evidence	0.028	0.878	0.094
Midnight	0.028	0.877	0.095
Day	0.027	0.883	0.090
Twilight	0.029	0.873	0.098

Material Type	Release Quantity		
no evidence	0.028	0.878	0.094
Flammable Liquids	0.029	0.872	0.099
Corrosives	0.027	0.882	0.091

Table 90: Release Quantity Distribution by Explanatory Variable. (T5 Network)

Causing Object	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.877	0.095
Floor and water/ liquid	0.017	0.874	0.108
None	0.034	0.901	0.065
Other	0.052	0.867	0.081

Contributing Action	Release Quantity		
no evidence	0.028	0.877	0.095
Improper Loading and Dropped	0.019	0.843	0.139
Other	0.033	0.892	0.076
Loose Fitting or Valve	0.033	0.910	0.057

Location	Release Quantity		
no evidence	0.028	0.877	0.095
Suburban/ Commercial/ Eastern	0.025	0.870	0.105
Urban/ Industrial or Commercial/ Eastern or Western	0.031	0.885	0.084

Season	Release Quantity		
no evidence	0.028	0.877	0.095
Spring	0.033	0.871	0.097
Summer	0.023	0.881	0.096
Fall	0.027	0.884	0.089
Winter	0.030	0.874	0.096

Container Type	Release Quantity		
no evidence	0.028	0.877	0.095
Fiber Box	0.027	0.872	0.100
Bottle	0.030	0.893	0.078

Failure Item-Area	Release Quantity		
	ZERO	**SMALL**	**MEDIUM**
no evidence	0.028	0.877	0.095
Basic Package Material on Top	0.038	0.852	0.110
Basic Package Material on Bottom	0.018	0.890	0.092
Closure on Top	0.027	0.913	0.060

Failure Mode	Release Quantity		
no evidence	0.028	0.877	0.095
Other	0.031	0.895	0.074
Punctured and Crushed	0.020	0.837	0.144

Shift	Release Quantity		
no evidence	0.028	0.877	0.095
Midnight	0.029	0.876	0.095
Day	0.026	0.882	0.092
Twilight	0.029	0.873	0.098

Material Type	Release Quantity		
no evidence	0.028	0.877	0.095
Flammable Liquids	0.029	0.872	0.099
Corrosives	0.027	0.881	0.092

APPENDIX D

FACE VALIDATION

Validation Questionaire

Base all answers on Unloading incidents associated with highway transport from 1993-2002 and on the categories used in this research. Refer to "Variables, Categories and Incident Types Considered."

Name and job function: _____
Years of hazmat transportation experience: _____

1) Is the hazmat release Bayesian network (slide #6) reasonable in terms of the temporal layout of the variables and the 5 stages? Please comment if you wish to.

2) Direct associations among 3 pairs of variables in Stage 1 were found using loglinear analysis. Which 3 pairs of variables are directly associated (related) and interpret "why" they are related.

Interpretation

a. _____and _____ _____

b. _____and _____ _____

c. _____and _____ _____

Eg. Go Hiking and Encounter Bear *More likely to see a bear in a wooded area.*

3) For the 3 associations in Stage 1 uncovered using loglinear modeling, do they make sense? Why?

4) Rank the following variables according to their degree of influence on releases involving *medium dollar loss*. Medium dollar loss is >$500 but <= $25,000.
Use 1 (most influential on dollar loss) through 10 (least influential), and use each number just once.

		No direct association
Container Type	_____	__
Failure Item-Area	_____	__
Location	_____	__
Material Type	_____	__
Causing Object	_____	__
Release Quantity	_____	__
Season	_____	__
Contributing Action	_____	__
Shift	_____	__
Failure Mode	_____	__

5) For the top 3 variables, why do they strongly influence medium dollar loss?

1._____
2. _____
3. _____

6) Rank the following variables according to their degree of influence on releases involving *medium release quantity*. Medium release quantity is >1 gal but <= 100 gal.
Use 1 (most influential) through 9 (least influential), and use each number just once.

		No association
Container Type	_____	__
Failure Item-Area	_____	__
Location	_____	__
Material Type	_____	__
Causing Object	_____	__
Season	_____	__
Contributing Action	_____	__
Shift	_____	__
Failure Mode	_____	__

7) For the top 3 variables, why do they strongly influence medium release quantity?
1._____
2. _____
3. _____

8) Assume you want to reduce occurrences of *medium dollar loss* releases. For your "top" variable in question #4, which category of this variable would you attempt to reduce first based on its level of impact?

_____

a. What operational changes could be made to reduce occurrences of this top category?
_____
_____

9) Assume you want to reduce occurrences of *medium release quantity* releases. For your "top" variable in question #6, which category of this variable would you attempt to reduce first based on its level of impact?

_____

a. What operational changes could be made to reduce occurrences of this top category?
_____
_____

10) Which combination of causing object and failure item-area occurs most of the time during a release involving medium dollar loss? (>$500 but <= $25,000)
"What tells the story of a medium dollar loss?"

	Group 1	Group 2	Group 3	Group 4
Causing Object	floor/water	other	floor/water	floor/water
Failure Item-Area	basic material on top of container	closure on top of container	basic material on bottom of container	closure on top of container

11) Which combination of causing object and failure item-area occurs most of the time during a release involving zero dollar loss? ($0)
"What tells the story of a zero dollar loss?"

	Group 1	Group 2	Group 3	Group 4
Causing Object	floor/water	other	floor/water	floor/water
Failure Item-Area	basic material on top of container	closure on top of container	basic material on bottom of container	closure on top of container

12) Which combination of contributing action and failure mode occurs most of the time during a release involving medium release quantity? (>1 gal but <= 100 gal)

	Group 1	Group 2	Group 3	Group 4
Contributing Action	Loose fitting/closure	improperly loaded and dropped	Loose fitting/closure	other
Failure Mode	punctured and crushed	punctured and crushed	other	other

13) Which combination of contributing action and failure mode occurs most of the time during a release involving zero release quantity? (0 gal)

	Group 1	Group 2	Group 3	Group 4
Contributing Action	Loose fitting/closure	improperly loaded and dropped	Loose fitting/closure	other
Failure Mode	punctured and crushed	punctured and crushed	other	other

Figure 23: Face Validation Questionaire.

APPENDIX E

CORRECTION PROCEDURE SOURCE CODE – LOCATION TRANSITION MATRIX

```
'CREATES THE TRANSITION MATRIX FOR LOCATION

Set Excel1 = CreateObject("Excel.Application")
Set Excel1 = GetObject("c:\Location_Classification.xls")
Set Sheet1 = Excel1.WorkSheets.Item("Class")
Open "c:\ Location_TransMatrix" For Output As #1


Dim EndRow, EndCol
Dim r, c
Dim CondProb() As Single, ClassProb() As Single
Dim SumMatrix() As Single
Dim TotalProb, SumProb
Dim NumClasses
Dim Dim AreaTypeCol, LandUseCol, GeoDivCol
Dim NumIndVar, NumPatterns
Dim Prob
Dim category
Dim ModalCatCol
Dim clProb

NumClasses = 2
NumIndVar = 3
ModalCatCol = NumIndVar + NumClasses + 1

'-----------------------------------
'Set Classification Probabilities
'Sixth position is modal category
EndRow = 36
EndCol = 6
NumPatterns = EndRow
ReDim ClassProb(EndRow, EndCol)

r = 1
c = 1
While r <= EndRow
  While c <= EndCol
    ClassProb(r, c) = Sheet1.Cells(r, c)
    c = c + 1
  Wend
  r = r + 1
  c = 1
Wend

'-----------------------------------
'Set Conditional Probabilities
Set Sheet2 = Excel1.WorkSheets.Item("CondProb")
EndRow = 13
EndCol = 3
NumCategories = EndRow
ReDim CondProb(EndRow, EndCol)
```

```
r = 1
c = 1
While r <= EndRow
  While c <= EndCol
    CondProb(r, c) = Sheet2.Cells(r, c)
    c = c + 1
  Wend
  r = r + 1
  c = 1
Wend

'-----------------------------------
'Calculate transition matrix using classification and conditional probabilities
TotalProb = 1
SumProb = 0
AreaTypeCol = 1
LandUseCol = 2
GeoDivCol = 3
ReDim SumMatrix(NumClasses)
For c2 = 1 To NumClasses   'true latent variable
 For c = 1 To NumClasses  'predicted latent variable
  CondProbCol = c2 + 1
  ClassCol = c + NumIndVar
  For r = 1 To NumPatterns
    category = ClassProb(r, AreaTypeCol)
    For i = 1 To NumCategories
      If category = CondProb(i, 1) Then
        Prob = CondProb(i, CondProbCol)
      End If
    Next i
    TotalProb = TotalProb * Prob
    category = ClassProb(r, LandUseCol)
    For i = 1 To NumCategories
      If category = CondProb(i, 1) Then
        Prob = CondProb(i, CondProbCol)
      End If
    Next i
    TotalProb = TotalProb * Prob
    category = ClassProb(r, GeoDivCol)
    For i = 1 To NumCategories
      If category = CondProb(i, 1) Then
        Prob = CondProb(i, CondProbCol)
      End If
    Next i
    TotalProb = TotalProb * Prob
    If c = ClassProb(r, ModalCatCol) Then
      clProb = 1
    Else
      clProb = 0
    End If
    TotalProb = TotalProb * clProb
    SumProb = SumProb + TotalProb
    TotalProb = 1
  Next r
  SumMatrix(c) = SumProb
  SumProb = 0
 Next c  'predicted
 Print #1, SumMatrix(1) & " " & SumMatrix(2)
Next c2  'true
```

Figure 24: Source Code for Location Transition Matrix.

Table 91: "Class" Worksheet for Source Code.

281	271	1	0.0015	0.9985	2
281	271	2	0.0012	0.9988	2
281	271	3	0.001	0.999	2
281	271	4	0.0006	0.9994	2
281	271	5	0.0007	0.9993	2
281	271	6	0.0004	0.9996	2
281	271	7	0.0005	0.9995	2
281	271	8	0.0005	0.9995	2
281	271	9	0.0005	0.9995	2
281	272	1	0.0029	0.9971	2
281	272	2	0.0023	0.9977	2
281	272	3	0.0019	0.9981	2
281	272	4	0.0012	0.9988	2
281	272	5	0.0014	0.9986	2
281	272	6	0.0007	0.9993	2
281	272	7	0.0009	0.9991	2
281	272	8	0.001	0.999	2
281	272	9	0.001	0.999	2
282	271	1	0.9994	0.0006	1
282	271	2	0.9992	0.0008	1
282	271	3	0.9991	0.0009	1
282	271	4	0.9985	0.0015	1
282	271	5	0.9987	0.0013	1
282	271	6	0.9976	0.0024	1
282	271	7	0.9981	0.0019	1
282	271	8	0.9983	0.0017	1
282	271	9	0.9982	0.0018	1
282	272	1	0.9997	0.0003	1
282	272	2	0.9996	0.0004	1
282	272	3	0.9995	0.0005	1
282	272	4	0.9992	0.0008	1
282	272	5	0.9993	0.0007	1
282	272	6	0.9988	0.0012	1
282	272	7	0.999	0.001	1
282	272	8	0.9991	0.0009	1
282	272	9	0.9991	0.0009	1
AREATYPE	LANDUSE	DIVISION	Class 1	Class 2	Modal

Table 92: "CondProb" Worksheet for Source Code.

281	0.001	0.9991
282	0.999	0.0009
271	0.3754	0.5404
272	0.6246	0.4596

Table 92 (continued).

1	0.059	0.028
2	0.1574	0.0941
3	0.2691	0.1937
4	0.0795	0.096
5	0.1462	0.1488
6	0.0487	0.0924
7	0.0847	0.1277
8	0.0592	0.0818
9	0.0963	0.1376

BIBLIOGRAPHY

[1] Agresti, Alan, Categorical Data Analysis (New York: Wiley-Interscience, 2002), pp. xiii, xv.

[2] Agresti, Alan, Categorical Data Analysis (New York: John Wiley & Sons, 1990), pp. vii, 1.

[3] Ref. 1, Op. Cit., pp. 1, 619.

[4] Personal Communication with and review of course offerings of industrial engineering departments at Georgia Tech, University of Michigan, Penn State, Northwestern, Purdue, Stanford, University of California at Berkeley, Virginia Tech, Cornell, and Texas A&M, August 22, 2005 – September 19, 2005. Based on U.S. News and World Report, America's Best Graduate Schools, 2006 Edition.

[5] "Functions of the Office of Hazardous Materials Safety," <http://hazmat.dot.gov/contact/ohhms_fn.htm> (accessed July 19, 2005).

[6] Ref. 2, Op. Cit., pp. 1.

[7] Kennedy, John J., Analyzing Qualitative Data Log-Linear Analysis for Behavioral Research (New York: Praeger Publishers, 1992), pp. xvii.

[8] Vermunt, Jeroen K. and Jay Magidson, Latent Class Analysis (Belmont, MA: Statistical Innovations Inc., 2002), pp. 1, unpublished.

[9] Personal Communication with Doug Reeves, Risk Assessment Engineer, Office of Hazardous Materials Safety, Research and Special Programs Administration, Department of Transportation, Washington D.C., September 24, 2002.

[10] Bolck, Annabel, Marcel Croon, and Jacques Hagenaars, "Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators," Political Analysis, Vol. 12, No. 1 (2004), pp. 4-5.

[11] Personal Communication with Demetra Collia, Mathematical Statistician and Project Manager of Safety Data Initiative, Bureau of Transportation Statistics, Department of Transportation, Washington D.C., August 1, 2005.

[12] Ibid., August 1, 2005.

[13] Safety in Numbers, Using Statistics to Make the Transportation System Safer – Safety Data Action Plan (Washington D.C.: Bureau of Transportation Statistics/ Department of Transportation, 2000), pp. 12, unpublished.

[14] Ref. 11, Op. Cit., August 5, 2005.

[15] Ref. 13, Op. Cit., pp. 15.

[16] "Project 10 Overview – Expand, Improve, and Coordinate Safety Data Analysis," <http://www.bts.gov/publications/safety_in_numbers_conference_2002/project10/project10_overview.html> (accessed July 22, 2004 and July 18, 2005).

[17] Ref. 11, Op. Cit., August 1, 2005.

[18] Guidelines for Applying Criteria to Designate Routes for Transporting Hazardous Materials (Washington, DC: Office of Highway Safety/Federal Highway Administration/ Department of Transportation, 1994), pp.14-15, unpublished.

[19] Personal Communication with Iris Shimizu, Mathematical Statistician, Office of Research and Methodology, National Center for Health Statistics, Center for Disease Control, Department of Health and Human Services, Hyattsville, MD., September 9, 2005 – September 20, 2005.

[20] Personal Communication with Doug Williams, Chief, Statistical Research and Survey Design Staff, Office of Research and Methodology, National Center for Health Statistics, Center for Disease Control, Department of Health and Human Services, Hyattsville, MD., September 19, 2005.

[21] "Office of Research and Development, Federal Stewardship in Service to Homeland Security" <http://www.dhs.gov/dhspublic/index.jsp> (accessed August 29, 2005).

[22] Burns, William J., and Robert T. Clemen, "Covariance Structure Models and Influence Diagrams," Management Science, Vol. 39, No. 7 (1993), pp. 816-834.

[23] Harwood, Douglas W., John G. Viner, and Eugene R. Russell, "Procedure for Developing Truck Accident and Release Rates for Hazmat Routing," Journal of Transportation Engineering (1993), pp. 191, 196.

[24] Center for Chemical Process Safety, 19th Annual International Conference, Emergency Planning: Preparedness, Prevention and Response, "Fuzzy Logic Methodology for Accident Frequency Assessment in Hazardous Materials Transportation by Yuanhua Qiao, Michela Gentile, and M. Sam Mannan" (unpublished), pp. 215-224.

[25] ReVelle, Charles, Jared Cohon, and Donald Shobrys, "Simultaneous Siting and Routing in the Disposal of Hazardous Wastes," Transportation Science, Vol. 25, No. 2 (1991), pp. 141,144.

[26] Current, John and Samuel Ratick, "A Model to Assess Risk, Equity and Efficiency in Facility Location and Transportation of Hazardous Materials," Location Science, Vol. 3, No. 3 (1995), pp. 198.

[27] Leonelli, Paolo, Sarah Bonvicini, and Gigliola Spadoni, "New Detailed Numerical Procedures for Calculating Risk Measures in Hazardous Materials Transportation," Journal of Loss Prevention in the Process Industries, Vol. 12 (1999), pp. 509.

[28] Ibid., pp. 511.

[29] Erkut, Erhan and Vedat Verter, "Modeling of Transport Risk for Hazardous Materials," Operations Research, Vol. 46, No. 5 (1998), pp. 625-642.

[30] Harwood, Douglas W., Eugene R. Russell, and John G. Viner, "Characteristics of Accidents and Incidents in Highway Transportation of Hazardous Materials," Transportation Research Record 1245 (1989), pp. 24.

[31] Harwood, Douglas W., John G. Viner, and Eugene R. Russell, "Truck Accident Rate Model for Hazardous Materials Routing," Transportation Research Record 1264 (1990), pp. 12-13.

[32] A National Risk Assessment for Selected Hazardous Materials in Transportation (Argonne, IL: Argonne National Lab, 2000), pp. 56, unpublished.

[33] Moses, Leon N. and Dan Lindstrom, eds., Transportation of Hazardous Materials Issues in Law, Social Science, and Engineering, "Databases and Needs for Risk Assessment of Hazardous Materials Shipments by Trucks, by Antoine G. Hobeika and Sigon Kim" (Boston: Kluwer Academic Publishers, 1993), pp. 146.

[34] Ref. 9, Op. Cit., February 6, 2002.

[35] Maher, Michael J. and Ian Summersgill, "A Comprehensive Methodology for the Fitting of Predictive Accident Models," Accident Analysis and Prevention, Vol. 28, No. 3 (1996), pp. 281-296.

[36] Miaou, Shaw-Pin and Harry Lum, "Modeling Vehicle Accidents and Highway Geometric Design Relationships," Accident Analysis and Prevention, Vol. 25, No. 6 (1993), pp. 689-709.

[37] Milton, John and Fred Mannering, "The Relationship Among Highway Geometrics, Traffic-Related Elements and Motor-Vehicle Accident Frequencies," Transportation, Vol. 25 (1998), p. 395-413.

[38] Shankar, Venkataraman, Fred Mannering, and Woodrow Barfield, "Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies," Accident Analysis and Prevention, Vol. 27, No. 3 (1995), pp. 371-389.

[39] Shankar, V., J. Milton, and F. Mannering, "Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry," Accident Analysis and Prevention, Vol. 29, No. 6 (1997), pp. 829-837.

[40] Ref. 27, Op. Cit., pp. 507-515.

[41] Bubbico, Roberto, Cinzia Ferrari, and Barbara Mazzarotta, "Risk Analysis of LPG Transport by Road and Rail," Journal of Loss Prevention in the Process Industries, Vol. 13 (1999), pp. 27-31.

[42] Jorissen, R.E. and P.J.M. Stallen, eds., Quantified Societal Risk and Policy Making. Technology, Risk, and Society: An International Series in Risk Analysis, Volume 12, "Risk Criteria for the Transportation of Hazardous Materials, by Henk G. Roodbol" (Boston: Kluwer Academic Publishers, 1998), pp. 41-47.

[43] Scanlon, Raymond D. and Edmund J. Cantilli, "Assessing the Risk and Safety in the Transportation of Hazardous Materials," Transportation Research Record 1020 (1985), pp. 6-11.

[44] Rao, K. Rajeshwar, S. Venkateswar Rao, and V. Chary, "Estimation of Risk Indices of Chemicals During Transportation," Process Safety Progress, Vol. 23, No. 2 (2004), pp. 149-154.

[45] Erkut, Erhan and Vedat Verter, "A Framework for Hazardous Materials Transport Risk Assessment," Risk Analysis Vol. 15, No. 5 (1995), pp. 590.

[46] Iakovou, Eleftherios, Christos Douligeris, Huan Li, Chi Ip, and Lalit Yudhbir, "A Maritime Global Route Planning Model for Hazardous Materials Transportation," Transportation Science,Vol. 33, No. 1 (1999), pp. 36.

[47] Gheorghe, Adrian V., Jurg Birchmeier, Dan Vamanu, Ioannis Papazoglou, and Wolfgang Kroger, "Comprehensive Risk Assessment for Rail Transportation of Dangerous Goods: a Validated Platform for Decision Support," Reliability Engineering and System Safety, Vol. 88 (2005), pp. 247-272.

[48] Ref. 25, Op. Cit., pp. 138-145.

[49] Erkut, Erhan and Armann Ingolfsson, "Catastrophe Avoidance Models for Hazardous Materials Route Planning," Transportation Science, Vol. 34, No. 2 (2000), pp. 166.

[50] Sivakumar, Raj A., Rajan Batta, and Mark H. Karwan, "A Multiple Route Conditional Risk Model for Transporting Hazardous Materials," INFOR, Vol. 33, No. 1 (1995), pp. 20-33.

[51] Jin, Honghua, Rajan Batta, and Mark H. Karwan, "On the Analysis of Two New Models for Transporting Hazardous Materials," Operations Research, Vol. 44, No. 5 (1996), pp. 710-723.

[52] Glickman, Theodore S. and Mary Anne Sontag, "The Tradeoffs Associated with Rerouting Highway Shipments of Hazardous Materials to Minimize Risk," Risk Analysis, Vol. 15, No. 1 (1995), pp. 62-63.

[53] Patel, Minnie H. and Alan J. Horowitz, "Optimal Routing of Hazardous Materials Considering Risk of Spill," Transportation Research-A, Vol. 28A, No. 2 (1994), pp. 119-132.

[54] Abkowitz, Mark, Mark Lepofsky, and Paul Cheng, "Selecting Criteria for Designating Hazardous Materials Highway Routes," Transportation Research Record 1333, (1992), pp. 30-35.

[55] Ref. 29, Op. Cit., pp. 630.

[56] Li, Ching Chun, Path Analysis: A Primer (Pacific Grove, CA: Boxwood Press, 1975), pp. 137.

[57] Loehlin, John C., Latent Variable Models An Introduction to Factor, Path, and Structural Analysis (Hillsdale, NJ: Lawrence Erlbaum Associates, 1987), pp. 9-15.

[58] Kendall, M.G. and C.A. O'Muircheartaigh, Path Analysis and Model Building, No. 2 / TECH. 414 (The Hague: International Statistical Institute, 1977), pp. 1-17.

[59] Asher, Herbert B., Causal Modeling (Newbury Park, CA.: Sage Publications, 1983), pp. 29-35.

[60] Romney, David M. and John M. Bynner, The Structure of Personal Characteristics (Westport, CT.: Praeger, 1992), pp. 8-11.

[61] Hagenaars, Jacques A., Loglinear Models with Latent Variables (Newbury Park, CA: Sage Publications, 1993), pp. 17, 49.

[62] Ref. 7, Op. Cit., pp. 239-240.

[63] Ritschard, Gilbert, Jean Kellerhals, Michael Olszak, and Massimo Sardi, "Path Analysis with Partial Association Measures," Quality & Quantity, Vol. 30 (1996), pp. 47-49, 55.

[64] Fienberg, Stephen E., The Analysis of Cross Classified Categorical Data (Cambridge: MIT Press, 1980), pp. 120.

[65] Leitner, Helga and H. Wohlschlagl, "Incorporating Polytomous Nominal Variables in Path Analysis," Bremer Beitrage zur Geographie und Raumplanung, Vol. 8 (1987), pp. 244.

[66] Glisson, Charles A. and Henry Man-Kwong Mok, "Methodological Observations on Applied Behavioral Science. Incorporating Nominal Variables in Path Analysis: A Cross Cultural Example with Human Service Organizations," The Journal of Applied Behavioral Science, Vol. 19, No. 1 (1983), pp.95-100.

[67] Heise, David R., "Employing Nominal Variables, Induced Variables, and Block Variables in Path Analysis," Sociological Methods & Research, Vol. 1, No. 2 (1972), pp. 147-151.

[68] Ref. 65, Op. Cit., pp. 242.

[69] Agresti, Alan and Barbara Finlay, Statistical Methods for the Social Sciences (Upper Saddle River, N.J.: Prentice Hall, 1997), pp. 634-635, 638.

[70] Kelloway, E. Kevin, Using LISREL for Structural Equation Modeling: A Researcher's Guide (Thousand Oaks, CA: Sage Publications, 1998), pp. 103.

[71] Saris, Willen E. and Irmtraud N. Gallhofer, eds., Sociometric Research, Volume 2 Data Analysis, "LCAG – Loglinear Modelling with Latent Variables: a Modified LISREL Approach, by J. A. Hagenaars," (New York: St. Martin's Press, 1988), pp. 111.

[72] Ref. 10, Op. Cit., pp. 4.

[73] Vermunt, Jeroen K., LEM: A General Program for the Analysis of Categorical Data (Tilburg, The Netherlands: Tilburg University, September 25, 1997), pp. 41, unpublished.

[74] Personal Communication with Michael Denisenko, Sales and Marketing Manager, Statistical Innovations Inc., Belmont, MA., March 31, 2004.

[75] Ibid., April 5, 2004.

[76] Personal Communication with Jeroen Vermunt Ph.D., Professor, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands, May 18, 2004.

[77] Ref. 10, Op. Cit., pp. 4-5.

[78] Marcoulides, George A. and Irini Moustaki, eds., Latent Variable and Latent Structure Models, "Using Predicted Latent Scores in General Latent Structure Models by Marcel Croon" (Mahwah, NJ: Lawrence Erlbaum Associates, 2002), pp. 199.

[79] Ref. 76, Op. Cit., May 18, 2004.

[80] Ref. 78, Op. Cit., pp. 198.

[81] Ref. 78, Op. Cit., pp. 198.

[82] Ref. 76, Op. Cit., May 11, 2004.

[83] Anderson, James C. and David W. Gerbing, "Assumptions and Comparative Strengths of the Two-Step Approach," Sociological Methods & Research, Volume 20, No. 3 (1992), pp. 321-331.

[84] Ref. 69, Op. Cit., pp. 589.

[85] Garson Ph.D., G. David, "Log-Linear, Logit, and Probit Models," <http://www2.chass.ncsu.edu/garson/pa765/logit.htm> (accessed March 27, 2004).

[86] Knoke, David and Peter J. Burke, Log-Linear Models (Newbury Park, CA.: Sage Publications, 1980), pp. 11-12, 22-24.

[87] Ibid., pp. 30.

[88] Ref. 7, Op. Cit., pp. 7, 14, 16-17.

[89] Ref. 2, Op. Cit., pp. 130-131, 152-153.

[90] Ref. 2, Op. Cit., pp. 217, 228-229.

[91] Marshall, Kneale T. and Robert M. Oliver, Decision Making and Forecasting (New York: McGraw-Hill, Inc., 1995), pp. 137.

[92] Ref. 7, Op. Cit., pp. 71, 93.

[93] Ref. 7, Op. Cit., pp. 93, 127.

[94] Ref. 7, Op. Cit., pp. 192.

[95] Ref. 7, Op. Cit., pp. 192.

[96] Ref. 7, Op. Cit., pp. 39.

[97] Hagenaars, Jacques A. and Allan L. McCutcheon, eds., Applied Latent Class Analysis, "Directed Loglinear Modeling with Latent Variables. Causal Models for Categorical Data with Nonsystematic and Systematic Measurement Errors by Jacques A. Hagenaars" (New York: Cambridge University Press, 2002), pp. 255.

[98] Dallal Ph.D., Gerard E., "Contingency Tables," (updated October 19, 2003), <http://www.tufts.edu/~gdallal/ctab.htm> (accessed October 11, 2004).

[99] Personal Communication with Jacques Hagenaars Ph.D., Professor, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands, December 21, 2004.

[100] Hagenaars, Jacques A., Categorical Longitudinal Data (Newbury Park, CA: Sage Publications, 1990), pp. 60.

[101] Ref. 2, Op. Cit., pp. 131-132.

[102] Garson Ph.D., G. David, "Log-Linear, Logit, and Probit Models," <http://www2.chass.ncsu.edu/garson/pa765/logit.htm> (accessed October 6, 2004).

[103] Ref. 7, Op. Cit., pp. 205.

[104] Ref. 99, Op. Cit., January 11, 2005.

[105] Ref. 99, Op. Cit., December 21, 2004 to January 26, 2005.

[106] Ref. 97, Op. Cit., pp. 259-260, 264-265.

[107] Ref. 78, Op. Cit., pp. 199.

[108] Ref. 10, Op. Cit., pp. 3-5.

[109] Ref. 76, Op. Cit., May 11, 2004.

[110] Ref. 76, Op. Cit., May 11, 2004.

[111] Ref. 10, Op. Cit., pp. 3, 6, 13-17.

[112] Ref. 10, Op. Cit., pp. 11.

[113] Ref. 78, Op. Cit., pp. 207-208.

[114] Croon, M., "A General Procedure for Correcting Joint Distributions of Predicted Latent Class Scores" (Tilburg, The Netherlands: Tilburg University, January, 2005), pp. 1-5, unpublished.

[115] "The Kronecker Tensor Product," <http://www.mathworks.com/access/helpdesk/help/techdoc/math/mat_linalg9.html> (accessed January 7, 2005).

[116] Ref. 114, Op. Cit., pp. 4.

[117] Vermunt, Jeroen K. and Jay Magidson, Latent Variable (Belmont, MA: Statistical Innovations Inc., 2002), pp.1, unpublished.

[118] Vermunt, Jeroen K. and Jay Magidson, Local Independence (Belmont, MA: Statistical Innovations Inc., 2002), pp.1, unpublished.

[119] McCutcheon, Allan L., Latent Class Analysis (Newbury Park, CA.: Sage Publications, 1987), pp.4, 8.

[120] Ref. 10, Op. Cit., pp. 4-5.

[121] Uebersax Ph.D., John, "LCA Frequently Asked Questions (FAQ)," (updated October 5, 2001), <http://ourworld.compuserve.com/homepages/jsuebersax/faq.htm> (accessed September 16, 2002).

[122] Collins, Linda M. and Larry A. Seitz, eds., Advances in Data Analysis for Prevention Intervention Research, NIDA Research Monograph No. 142, "Latent Class Analysis of Substance Abuse Patterns, by John S. Uebersax" (Rockville, MD.: U.S. Department of Health and Human Services/Public Health Service/NIH/NIDA, 1994), pp. 64-80.

[123] Ref. 61, Op. Cit., pp. 20.

[124] Ref. 119, Op. Cit., pp. 4.

[125] Arminger, Gerhard, Clifford C. Clogg, and Michael E. Sobel, eds., Handbook of Statistical Modeling for the Social and Behavioral Sciences, "Latent Class Models, by Clifford C. Clogg" (New York: Plenum Press, 1995), pp. 334.

[126] Personal Communication with Jay Magidson Ph.D., President, Statistical Innovations Inc., Belmont, MA, October 2003 to November 2003.

[127] Ref. 119, Op. Cit., pp. 32.

[128] Personal Communication with Scott Eliason, Ph.D., Associate Professor, Department of Sociology, University of Minnesota, Minneapolis, MN., January 2003.

[129] Ref. 126, Op. Cit., October 2003 to November 2003.

[130] Ref. 119, Op. Cit., pp. 18-21, 33.

[131] Ref. 8, Op. Cit., pp.4.

[132] Dayton, C. Mitchell., Latent Class Scaling Analysis (Thousand Oaks, CA: Sage Publications, 1998), pp. 12.

[133] Uebersax Ph.D., John, "A Brief Study of Local Maximum Solutions in Latent Class Analysis," (updated August 10, 2000), <http://ourworld.compuserve.com/homepages/jsuebersax/local.htm> (accessed February 15, 2003).

[134] Ref.132, Op. Cit., pp. 12-13.

[135] Vermunt, Jeroen K. and Jay Magidson, Latent Class Cluster Analysis (Belmont, MA: Statistical Innovations Inc., 2002), pp.6, unpublished.

[136] Ref. 119, Op. Cit., pp. 23-25.

[137] Ref. 8, Op. Cit., pp. 5.

[138] Ref. 119, Op. Cit., pp. 32.

[139] Ref.132, Op. Cit., pp. 20.

[140] Ref. 122, Op. Cit., pp. 70.

[141] Ref. 7, Op. Cit., pp. 251.

[142] Ref. 100, Op. Cit., pp. 66.

[143] Garson Ph.D., G. David, "Latent Class Analysis," <http://www2.chass.ncsu.edu/garson/pa765/latclass.htm> (accessed November 8, 2002).

[144] Ref.132, Op. Cit., pp.18-19.

[145] Ref.132, Op. Cit., pp. 19-20.

[146] Ref. 121, Op. Cit.

[147] Ref. 119, Op. Cit., pp. 35-37.

[148] Ref. 8, Op. Cit., pp. 3.

[149] Ref. 119, Op. Cit., pp. 35-36.

[150] Ref. 119, Op. Cit., pp. 37.

[151] Ref. 8, Op. Cit., pp. 7.

[152] Ref. 143, Op. Cit.

[153] Ref. 125, Op. Cit., pp. 335.

[154] Ref. 126, Op. Cit., October 2003 to November 2003.

[155] Ref. 8, Op. Cit., pp. 5.

[156] Ref. 61, Op. Cit., pp. 25.

[157] Ref. 8, Op. Cit., pp. 5.

[158] Ref. 133, Op. Cit.

[159] Ref. 8, Op. Cit., pp. 5.

[160] Jensen, Finn V., An Introduction to Bayesian Networks (New York: Springer, 1996), pp. 18.

[161] Charniak, Eugene, "Bayesian Networks without Tears," AI Magazine, (Winter 1991), pp. 50-54.

[162] Murphy Ph.D., Kevin, "A Brief Introduction to Graphical Models and Bayesian Networks," (created 1998), <http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html> (accessed May 15, 2005).

[163] Personal Communication with Marek J. Druzdzel Ph.D., Associate Professor, Decision Systems Laboratory, School of Information Sciences, University of Pittsburgh, Pittsburgh, PA., April 13, 2005 and April 27, 2005.

[164] Jagt, Randy M, "Support for Multiple Cause Diagnosis with Bayesian Networks" (unpublished M.S. Thesis, Department of Mediamatics, Information Technology and Systems, Delft University of Technology, the Netherlands and Information Sciences Department, University of Pittsburgh, Pitsburgh, PA., 2002).

[165] Hagenaars, Jacques A., "Categorical Causal Modeling. Latent Class Analysis and Directed Log-Linear Models with Latent Variables," Sociological Methods & Research, Volume 26, No. 4 (1998), pp. 452-453.

[166] Cooper, Gregory F. and Edward Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data," Machine Learning, Vol. 9 (1992), pp. 309-347.

[167] Ref. 163, Op. Cit., September 14, 2005.

[168] Ref. 160, Op. Cit., pp. 10, 19-20.

[169] Ref. 160, Op. Cit., pp. 16, 56.

[170] Personal Communication with Finn V. Jensen Ph.D., Professor, Department of Computer Science, Aalborg University, Aalborg, Denmark, February 17, 2005.

[171] "Decision Theoretic Modeling, Probability,"<http://genie.sis.pitt.edu/GeNIeHelp/index.html> (accessed November 3, 2005).

[172] Ref. 164, Op. Cit., pp. 15.

[173] "Elements of Genie, Inference Algorithms, Bayesian Network Algorithms, Exact Algorithms, Clustering Algorithm,"<http://genie.sis.pitt.edu/GeNIeHelp/index.html> (accessed November 3, 2005).

[174] Witten, Ian H. and Eibe Frank, Data Mining (San Francisco, CA: Morgan Kaufmann Publishers, 2000), pp. 89-94.

[175] Ref. 64, Op. Cit., pp. 120.

[176] Goodman, Leo A., "Causal Analysis of Data from Panel Studies and Other Kinds of Surveys," The American Journal of Sociology, Vol. 78, No. 5 (1973), pp. 1135.

[177] Personal Communication with Carlisle Smith, Ohio Public Utilities Commission, Columbus, OH., October 31, 2002.

[178] Ref. 9, Op. Cit., December 2, 2002.

[179] Ref. 102, Op. Cit.

[180] Koch, Richard, The 80/20 Principle The Secret of Achieving More with Less (New York: Doubleday, 1998), pp. 34.

[181] "Census Regions and Divisions of the United States," (created September 12, 2001, updated July 7, 2005), <http://www.census.gov/geo/www/us_regdiv.pdf> (accessed August 2004).

[182] Ref. 9, Op. Cit., January 24, 2005 – January 26, 2005.

[183] Guide for Preparing Hazardous Materials Incidents Reports (Washington D.C.: Research and Special Programs Administration/Department of Transportation, January 2004), pp. 7, unpublished.

[184] Ref. 9, Op. Cit., January 24, 2005 – January 26, 2005.

[185] Ref. 78, Op. Cit., pp. 198.

[186] Ref. 61, Op. Cit., pp. 41.

[187] Personal Communication with C. Mitchell Dayton Ph.D., Professor and Chair, Department of Measurement, Statistics and Evaluation, College of Education, University of Maryland, College Park, MD., December 11, 2002.

[188] Ref. 126, Op. Cit., November 17, 2003 and April 28, 2004.

[189] Ref. 2, Op. Cit., pp. 222-223.

[190] Ref. 86, Op. Cit., pp. 44.

[191] Pate-Cornell, M. Elisabeth and Dean M. Murphy, "Human and Management Factors in Probabilistic Risk Analysis: the SAM Approach and Observations from Recent Applications," Reliability Engineering and System Safety, Vol. 53 (1996), pp. 115-126.

[192] Personal Communication with Ron Duych (Engineer), Steve Hwang (Chemist), Doug Reeves (Supervisory General Engineer), Tanya Schreiber (Scientist), Kin Wong, Ph.D (General Engineer), Department of Transportation, Washington D.C., August 17, 2005.

[193] Ref. 86, Op. Cit., pp. 44.

[194] Ref. 61, Op. Cit., pp. 41.

[195] Personal Communication with Kin Wong, Ph.D., General Engineer, Office of Hazardous Materials Safety, Research and Special Programs Administration, Department of Transportation, Washington D.C., December 1, 2004.

[196] Ref. 2, Op. Cit., pp. 222-223.

[197] Ref. 2, Op. Cit., pp. 223.

[198] Ref. 86, Op. Cit., pp. 44-45.

[199] Ref. 76, Op. Cit., May 11, 2004.

[200] Personal Communication with Marcel Croon Ph.D., Senior Lecturer, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands, January 13, 2005.

[201] Ref. 2, Op. Cit., pp. 217.

[202] Ref. 7, Op. Cit., pp. 245.

[203] Ref. 100, Op. Cit., pp. 68-69.

[204] Ref. 7, Op. Cit., pp. 243.

[205] Ref. 61, Op. Cit., pp. 48-49.

[206] Ref. 99, Op. Cit., December 21, 2004 to January 11, 2005.

[207] Ref. 99, Op. Cit., January 26, 2005.

[208] Ref. 10, Op. Cit., pp. 17.

[209] Ref. 97, Op. Cit., pp. 257.

[210] Ref. 165, Op. Cit., pp. 457, 479.

[211] Ref. 99, Op. Cit., November 5, 2004 and February 15, 2005.

[212] Ref. 10, Op. Cit., pp. 17.

[213] Ref. 97, Op. Cit., pp. 257.

[214] Ref. 165, Op. Cit., pp. 457, 479.

[215] Ref. 99, Op. Cit., February 15, 2005.

[216] Ref. 99, Op. Cit., February 27, 2005.

[217] Intelligent Information Systems VII Workshop Proceedings, Malbork, Poland, June 15-19, 1998, "A Probabilistic Causal Model for Diagnosis of Liver Disorders, by Agnieszka Onisko, Marek J. Druzdzel, and Hanna Wasyluk," pp. 383.

[218] Ref. 163, Op. Cit., March 9, 2005 to March 11, 2005.

[219] Ref. 163, Op. Cit., March 9, 2005 to March 11, 2005.

[220] Ref. 217, Op. Cit., pp. 383.

[221] Ref. 163, Op. Cit., March 7 and April 13, 2005.

[222] Ref. 217, Op. Cit., pp. 383-384.

[223] Ref. 163, Op. Cit., June 8, 2005.

[224] Ref. 163, Op. Cit., June 8, 2005.

[225] Ref. 164, Op. Cit., pp. 23, 29.

[226] Ref. 163, Op. Cit., April 13, 2005 and April 27, 2005.

[227] Clemen, Robert T. and Terence Reilly, Making Hard Decisions with Decision Tools (Pacific Grove, CA: Duxbury Thomson Learning, 2001), pp. 133-137.

[228] Ref. 9, Op. Cit., February 6, 2006.

[229] Ref. 9, Op. Cit., February 2, 2006.

[230] Ref. 163, Op. Cit., June 8, 2005.

[231] Ref. 163, Op Cit., April 27, 2005 and June 8, 2005.

REFERENCES NOT CITED

Boffey, T.B., and J. Karkazis, "Linear Versus Nonlinear Models for Hazmat Routing," INFOR, Vol. 33, No. 2 (1995), pp. 114-117.

Egidi, Demetrio, Franco P. Foraboschi, Gigliola Spadoni , and Aniello Amendola, "The ARIPAR Project: Analysis of the Major Accident Risks Connected with Industrial and Transportation Activities in the Ravenna Area," Reliability Engineering and System Safety, Vol. 49 (1995), pp. 75-89.

Erkut, Erhan and Armann Ingolfsson, "Transport Risk Models for Hazardous Materials: Revisited," Operations Research Letters, Vol. 33 (2005), pp. 81-89.

Erkut, Erhan and Vedat Verter, "Modeling of Transport Risk for Hazardous Materials," Operations Research, Vol. 46, No. 5 (1998), pp. 625-642.

Kara, B. Y., E. Erkut, and V. Veter, "Accurate Calculation of Hazardous Materials Transport Risks," Operations Research Letters, Vol. 31, No. 6 (2003), pp. 285-292.

Kara, Bahar Y. and Vedat Veter, "Designing a Road Network for Hazardous Materials Transportation," Transportation Science, Vol. 38, No. 2 (2004), pp. 188-196.

Saccomanno, F. Frank and A. Y. W. Chan, "Economic Evaluation of Routing Strategies for Hazardous Road Shipments," Transportation Research Record 1020, (1985), pp. 12-18.

Saccomanno, F. F. and J. H. Shortreed, "Hazmat Transport Risks: Societal and Individual Perspectives," Journal of Transportation Engineering, Vol. 119, No. 2 (1993), pp. 177-188.

Sherali, Hanif D., Laora D. Brizendine, Theodore S. Glickman, and Shivaram Subramanian, "Low Probability-High Consequence Considerations in Routing Hazardous Material Shipments," Transportation Science,Vol. 31, No. 3 (1997), pp. 237-251.

Sivakumar, Raj A., Rajan Batta, and Mark H. Karwan, "A Network-Based Model for Transporting Extremely Hazardous Materials," Operations Research Letters, Vol. 13 (1993), pp. 85-93.

Verter, Vedat and Bahar Y. Kara, "A GIS-Based Framework for Hazardous Materials Transport Risk Assessment," Risk Analysis, Vol. 21, No. 6 (2001), pp. 1109-1120.