ANALYSIS OF THE HSEES CHEMICAL INCIDENT DATABASE USING DATA AND TEXT MINING METHODOLOGIES

A Thesis

by

MAHDIYATI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2011

Major Subject: Safety Engineering

ANALYSIS OF THE HSEES CHEMICAL INCIDENT DATABASE USING DATA AND TEXT MINING METHODOLOGIES

A Thesis

by

MAHDIYATI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee, M. Sam Mannan
Committee Members, Marietta Tretter
Mahmoud El-Halwagi
Head of Department Michael Pishko

May 2011

Major Subject: Safety Engineering

ABSTRACT

Analysis of the HSEES Chemical Incident Database Using Data and Text Mining Methodologies. (May 2011)

Mahdiyati, B.S., Institut Teknologi Bandung

Chair of Advisory Committee: Dr. M. Sam Mannan

Chemical incidents can be prevented or mitigated by improving safety performance and implementing the lessons learned from past incidents. Despite some limitations in the range of information they provide, chemical incident databases can be utilized as sources of lessons learned from incidents by evaluating patterns and relationships that exist between the data variables. Much of the previous research focused on studying the causal factors of incidents; hence, this research analyzes the chemical incidents from both the causal and consequence elements of the incidents.

A subset of incidents data reported to the Hazardous Substance Emergency Events Surveillance (HSEES) chemical incident database from 2002-2006 was analyzed using data mining and text mining methodologies. Both methodologies were performed with the aid of STATISTICA™ software. The analysis studied 12,737 chemical process related incidents and extracted descriptions of incidents in free-text data format from 3,316 incident reports. The structured data was analyzed using data mining tools such as classification and regression trees, association rules, and cluster analysis. The unstructured data (textual data) was transformed into structured data using text mining, and subsequently analyzed further using data mining tools such as, feature selections and cluster analysis.

The data mining analysis demonstrated that this technique can be used in estimating the incident severity based on input variables of release quantity and distance between victims and source of release. Using the subset data of ammonia release, the classification and regression tree produced 23 final nodes. Each of the final nodes corresponded to a range of release quantity and, of distance between victims and

source of release. For each node, the severity of injury was estimated from the observed severity scores' average. The association rule identified the conditional probability for incidents involving piping, chlorine, ammonia, and benzene in the value of 0.19, 0.04, 0.12, and 0.04 respectively. The text mining was utilized successfully to generate elements of incidents that can be used in developing incident scenarios. Also, the research has identified information gaps in the HSEES database that can be improved to enhance future data analysis. The findings from data mining and text mining should then be used to modify or revise design, operation, emergency response planning or other management strategies.

DEDICATION

To: *Bara Chatun,* my dearest mother for her unconditional love, patience, and belief in myself, who has driven me to become a wholesome person

Syukri Abdullah, my dearest father for his love, encouragement, exemplary hard work, and advice, who has inspired me to achieve further in life

Aviscenna, Fina Mutqina and Taufik Ridha, my siblings for their love, support and companionship, who have colored my life with joy and honesty

ACKNOWLEDGEMENTS

I would like to express a profound appreciation to Dr. Sam Mannan, for his continuous encouragement throughout the years of my master study. He is an individual who demonstrates integrity, intelligence, and passion toward his works and life, and cares for his students; he will always be an inspiration to me to achieve further in life.

I would like to thank Dr. Marietta Tretter for sharing her knowledge and enthusiasm in data mining. She has provided valuable ideas, suggestions and comments that guided me throughout the research. I am very grateful for Dr. Mahmoud El-Halwagi, his kindness and encouragement has kept me going further with this research. I would like to honor Mr. Mike O'Connor, for his ideas and passion on the area of chemical incident data analysis and also his support and vision to the Mary Kay O'Connor Process Safety Center in becoming the leading institution in process safety.

I would also like to thank Dr. Hans Pasman, Dr. Trevor Kletz, Dr. Jack Chosnek, Dr. Maria Molnarne, Dr. Maria Papadaki, Dr. Adam Markowski, and Dr. Simon Waldram, for their wonderful suggestions to my research. I would like to acknowledge Dr. Xiaodan Gao, Dr. Dedy Ng and Dr. William Rogers, for the brainstorm sessions and critiques on my research. I would like to honor the late Dr. Harry West. The fellowship that bears his name has financially supported my study. I want to thank Valerie Green, Towanna Arnold, Mary Cass, and Donna Startz, for their help on the non-technical work during my master's program.

I thank my friends and colleagues, Anisa Safitri, Jaffee Suardin, Sara Khan, Katherine Prem, Suhani Patel, Carmen Osorio, Jiejia Wang and others for helping me throughout my master years, and making my experience at the Center valuable.

Finally, I would like to express my gratitude to my mother, father and siblings for their love, support and patience throughout my life.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1 Introduction

The chemical process industries want to improve their safety performance due to an increase in safety awareness and the better understanding of the risks that pertain to the process industry activities. One of the ways to improve safety performance in the industry is to monitor the lagging and leading indicators in order to evaluate the present state of chemical incidents and predict their tendencies in the near future. Examples of lagging indicators include number of incidents, number of victims, number and type of equipment failure, number of hazardous substance spill, lost work days, etc.

The U.S. has several chemical incident databases that have been established to monitor lagging indicators, maintain incident records and reduce the effect of incidents, both onsite and offsite. The data collected by these chemical incident databases are then analyzed to obtain useful information. The most typical analysis is trend analysis using statistics where a single variable is usually plotted against a period of time. While trend analysis can provide a good visualization of incident tendencies and a guide for prioritizing the focus of improvements, a more comprehensive analysis that includes multiple variables can be performed in order to get more benefit from the data.

The lessons learned from an incident should be used to modify or revise design, operation, maintenance, emergency response planning, and other management strategies. Through thorough investigation, the root causes and lessons learned from incidents can be extracted, and implemented to improve the safety of the industrial processes. However, it is neither efficient nor effective to investigate all incidents. Therefore, a comprehensive analysis on the chemical incident needs to be performed.

This thesis follows the style of *Journal of Loss Prevention in the Process Industries*.

This analysis should provide general information about chemical incident trends as well as specific information that indicate types of incidents that need immediate attention and follow up.

1.2 Motivation

As shown in Figure 1 (MKOPSC, 2009), the US has numerous chemical incident databases such as the Hazardous Substances Emergency Events Surveillance (HSEES), Risk Management Plan (RMP), Occupational Safety and Health Administration (OSHA), National Response Center (NRC), Department of Transportation (DOT) - Hazardous Materials Information and Resource System (HMIRS) and others. Each of the databases was established to serve different purposes and cover different areas; therefore information contained in each database varies from one to another.

The HSEES database collected incidents which occurred in fixed facilities, during transportation activities, and in areas other than the industrial facilities. This database gathered many details about the adverse effects of incidents on human health. The Environmental Protection Agency (EPA) administers the RMP database which compiles incidents involving regulated chemicals above a certain threshold quantity in a number of covered processes in 5 years time period. The RMP data covers a wide range of information including the onsite and offsite impacts of the incidents.

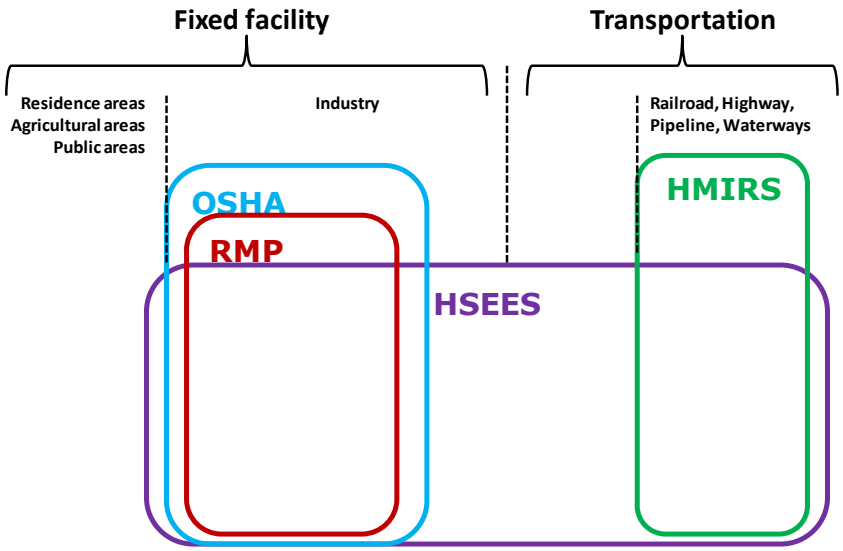


Figure 1: Coverage of several chemical incident databases in the U.S.

The OSHA chemical incidents database collects work-related incidents which include a high number of hard hat safety type incidents. The HMIRS chemical incident database was established to meet the federal hazardous material transportation regulation. All modes of transportation except for pipeline and bulk marine transportation are covered by the HMIRS database. The process industry should take advantage of these chemical incident databases because the incident data contains useful information on incident prevention and mitigation.

The effort to collect information on incidents and learn from the mistakes made by others has been considered a good industrial practice. After the Flixborough catastrophe incident in the UK occurred in 1970, the Institution of Chemical Engineers (IChemE) published the Loss Prevention Bulletin to share incident case studies and lessons learned from each of them. There were many other similar efforts established by safety professionals around the globe. Since then, a steady continuous improvement in industrial safety has been observed, and this has been complemented by the implementation of a proper safety management system (Jones et al., 1999).

Based on the availability of chemical incident database and their evident benefits, this research has been tailored to explore the HSEES database by analyzing relationships between its variables. These variables will be selected based on their potential for producing information that can be used as a reference for improving safety performance in general and, specifically in risk reduction efforts. The perimeters of the analysis obtained from the chemical incident databases are subject to the shortcomings that exist in the databases (Mannan et al., 1999). Therefore, the analyses should be generated as prudent work to ensure that these variables are translated into meaningful information and furthermore, knowledge.

1.3 Background: Chemical Incident Analysis

There has been various analysis and research conducted on chemical incident databases by either individuals or organizations. The most common methodologies were statistical analysis. Much of the research performed statistical analysis to chemical incident data (Uth, 1999; Wakakura & Iiduka, 1999; ATSDR, 2006; Welles et al., 2009) and more recent research has used data mining analysis.

The ATSDR published an annual report on its website on the chemical incidents collected by HSEES. The report contained elaborate statistical analysis on incidents that occurred in different facilities and produced analysis describing the number of chemicals released, the number of victims, the distribution of victims based on their age, sex, occupation, etc, the number of injuries and fatalities based on the severity and type of injuries, and other information regarding response and decontamination activities (ATSDR, 2004). The findings on the annual report were disseminated to health and safety professionals, emergency responders, and other parties.

The New York HSEES state agency conducted a state-specific statistical analysis on incident data. Several patterns were found such as carbon monoxide poisoning due to underground utility cable fires and significant mercury spills in schools and other public areas. Partnering with the Local Emergency Planning Committees (LEPCs), the New York HSEES system has provided the state incident data, lessons learned and case

studies as inputs for the committees. Through analysis, the New York HSEES discerned an increasing number of clandestine drug labs and supplied law enforcement, fire fighters, and other related parties with information about the hazards posed by these labs, which helped them in preparing their response plans and actions (Welles et al., 2009).

The Mary Kay O'Connor Process Safety Center has published several papers and produced a number of theses and dissertations on chemical incident data analysis using various databases. A subset of HSEES data related to the chemical process from 2002 to 2004 was analyzed using data mining tools such as cluster analysis, decision tree and logistic regression, and text mining tools. In this research, the cluster analysis produced 3 clusters each of which had 4 discernable characteristics of the type of industry, contributing factors, the state where the release occurred, and the number of chemicals released.

Text mining was used to analyze the variable which describes a brief summary of each incident. The text mining results were used as input for cluster analysis and produced clusters that provided a better description of the incidents compared to the previous cluster analysis using non textual data. A decision tree model was used to predict the outcome of the incidents and the generated model was able to correctly predict incidents that resulted in injury with 16% accuracy. The decision tree analysis was performed using text inputs. The results showed that the model had the ability to predict incidents with injury up to 57% accuracy. Logistic regression using data and text input was performed to predict the likelihood of an injury occurring given certain variables were present in the data. The prediction model showed and quantified that particular contributing factors, chemicals and industries significantly increased the likelihood of injuries occurring (Veltman, 2008).

A subset of the NRC incident data was analyzed using two data mining techniques; decision tree analysis and association rules. The decision tree technique was applied to describe and classify incident data that led to fires or explosions and injuries as consequences of releases. The association rule technique was applied to produce lift values for the variables type of equipment and type of chemicals involved in the

incidents. The lift values were proposed to be used as a factor to update the equipment failure probability values so they are chemical-specific (Anand, 2005).

Another research project used a subset of HSEES data from 2000 to 2004 to generate trend analysis using many different variables. The analysis was performed by classifying the incidents into two categories: system interruption events and system comparison events. This research also used a scaling ratio to estimate the national incident statistics based on the amount of HSEES data (Obidullah, 2006).

A subset of the RMP data was studied with the objective of generating trend analysis for frequently released chemicals and evaluating the limitations of the database. The study focused on potential improvements to the database by relating the failure rate obtained from the RMP database to an existing failure rate database, such as Offshore Reliability Data (OREDA), and including factors such as hazard information for the chemicals released and analytical information or lessons learned from the incidents (Al-Qurashi, 2000).

Text mining was performed using the Major Accidents Reporting System (MARS), a chemical incident database created by the European Union member countries to produce clusters of incidents. The analysis produced importance plots of variables that can predict incidents with particular elements (the dependent variable) using independent variables of other elements of the incidents. The importance plots showed the F-values of each predictor to evaluate their significance in predicting the dependent variable. The results from the text mining were also used to cluster the incident data to observe their natural clusters. Furthermore, the clusters observed were applied to develop the chemical incident taxonomy that would later be used in the active and knowledge-based incident retrieval system (Khan, 2010).

An analysis of the narrative text analysis was performed with the Kentucky tractor fatality reports, producing likelihood values for incidents with certain outcomes such as death at the scene or fatally crushed. The likelihood of an incident's outcome was modeled using logistic regression and predictor variables such as tractor equipment (front-end loaders, counterweight, roll-over protective structure), environmental conditions (muddy terrains), victims' conditions (thrown away, overturn), tractors'

mechanical factors (brakes, seat belt), and incident location (slope, flat terrain). This research found that the likelihood of being crushed by a tractor as a consequence of a tractor incident increased by a factor of 8.8 and 6.2 respectively for incidents where the tractor rolled over and the tractor was operating on sloped areas. While the likelihood of an incident resulting in death at the scene increased by a factor of 9.1for tractor operations equipped with front-end loaders (Bunn et al., 2008).

The HSEES and RMP were also used to generate annual exceedance frequencies using data on the number of fatalities or injuries. Linear regression was performed to evaluate the relationship between predictor variables such as the number of major injuries, minor injuries, and evacuations, and the number of fatalities that occurred. The resulting regression equation showed that the number of fatalities was strongly influenced by the number of lower consequence events; the existence of incidents with fatalities indicated the existence of a high number of lower consequence events, such as injuries and evacuations (Prem et al., 2010).

From the literature review, it can be observed that previous studies on chemical incident data have mainly focused on trend analyses of a single variable of the chemical incident database. Much of the previous research performed multivariate analysis using data mining methodology and variables that described the cause of the incidents. Many however, did not use variables that describe the consequence of the incidents such as severity. Therefore, this research tried to improve the multivariate analysis by including variables that describe the severity experienced by the victims. This research also studied the relationship between the consequence and the causal factors of incidents such as the quantity of release and the severity of the incident in terms of injury or fatality.

1.4 Objectives

The research objectives are to provide information from past incidents which can be used as a basis for developing recommendations to improve safety performance in the industry. In order to generate such information, this research has been tailored:

- To obtain the trend analysis of chemical incidents reported to the HSEES database from 2002 to 2006 by evaluating the following variables:
  - Number of incidents and types of incidental releases
  - Causal factors such as: contributing causes, type and amount of chemical released
  - Consequence factors such as: number and type of injuries, number of fatalities
- To investigate the relationship between the quantity of the chemicals released and the severity of the consequences, in respect to adverse health effects.
- To produce pattern analysis that identifies and quantifies the association between two or more variables that describes the type of releases, equipment and chemicals.
- To perform text mining on the HSEES incident comment variable
- To propose recommendations to improve the current incident collection system in order to produce better analysis that benefits process safety.

2. INCIDENT DATABASE

This research selected the Hazardous Substance Emergency Events Surveillance (HSEES) database for analysis because it was considered the most comprehensive and reliable database due to its active reporting system and relatively wide range of data collected which covers up to 16 states in the US.

2.1 HSEES Chemical Incident Database

The HSEES database was created and managed by the Agency for Toxic Substances and Disease Registry (ATSDR) of the Centers for Disease Control and Prevention (CDC) with the objective of collecting incident data, particularly data that describes the adverse health effects of incidents due to chemical releases, in terms of the morbidity and mortality experienced by the workers, emergency responders and general public (ATSDR, 2004). This surveillance system was established in 1993 and was ended in 2010 due to funding related issues. HSEES was replaced by another chemical incident database called the National Toxic Substances Incident Program (NTSIP) in 2010.

Fifteen states participated in HSEES annually from 2002 to 2005 and 14 states participated in 2006. The states which consistently participated throughout the active years of HSEES were Colorado, Iowa, Louisiana, Minnesota, New Jersey, New York, North Carolina, Oregon, Texas, Utah, Washington, and Wisconsin. In 2010, before HSEES data collection ended, the state participation went down to 7.

The flow of information on incidents reported to the HSEES chemical incident data collection system is shown in Figure 2 (MKOPSC, 2009). The health department of each participating state was expected to report the incidents through a web-based collection system within 48 hours after the incident had occurred.

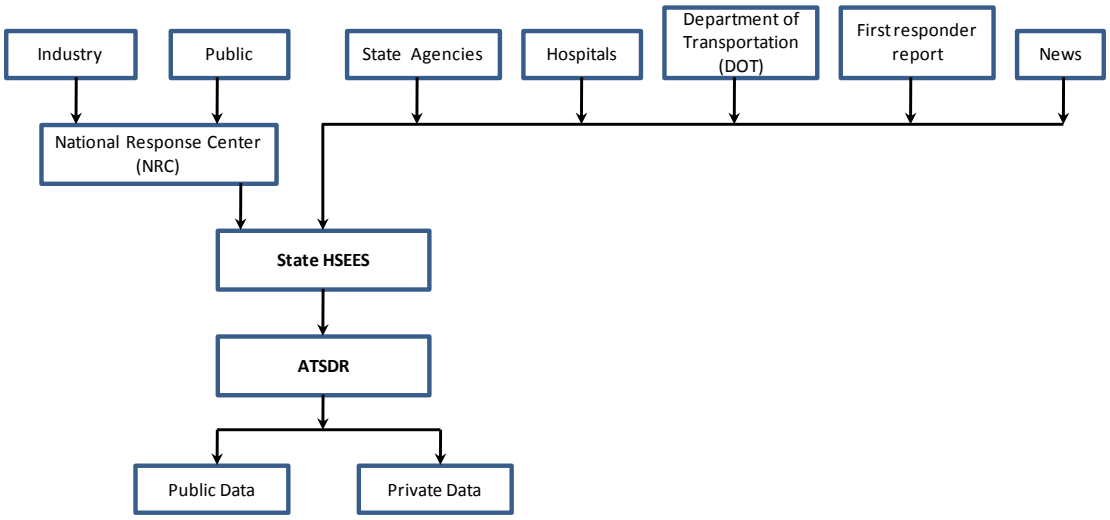


Figure 2: Information flow of the incidents reported to HSEES system

An incident was reported to HSEES if there was an uncontrolled or illegal hazardous substance release or if there was the threat of a hazardous substance release other than petroleum, where more than 10 lbs or 1 gallon was released. If no hazardous substance was released, the incident must still be reported to HSEES in the event that an evacuation, order for sheltering in place or other public health precaution was put into place.

The HSEES is a massive chemical incident database system that has more than 100 variables and 120,145 incident reports collected from 1993-2006. Several examples of information that is described in the HSEES variables are as follows:

- Event identification number and notification information.
- Description of the incident: date, time, location, type of industry, equipment, contributing factors, chemicals, physical state of the release, quantity of the chemical released, etc.
- Description of the victims: number of injuries and fatalities, type of injuries, severity of injuries, location where victims were found, personal protection equipment worn by the victims, number of people evacuated, etc.

- Potential community exposure: population within ¼, ½, 1 mile from the incident area, land use information.
- Response to and termination of the incident.

Due to the broad range of incident collected by HSEES, as shown in Figure 3, this study focuses on a particular segment of the incident data which is pertinent to chemical process incidents. This segmented data limits the scope of the analysis to 12,737 incident reports.

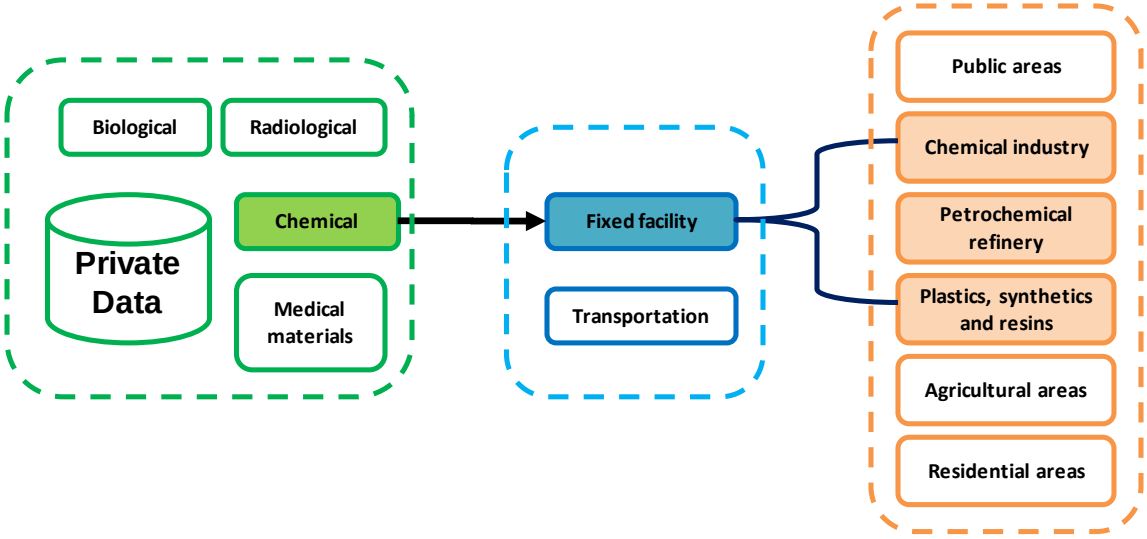


Figure 3: Subset data of HSEES used in the research

The HSEES data used in this analysis was extracted from a Microsoft Access database. The chemical incident database was organized in a relational table system, where tables containing event or incident information, chemicals, and victims' descriptions can be linked using a unique identification key. Table 1 shows a snapshot of the chemical incident database with several examples of variables used in this research.

Table 1: HSEES chemical incident database

Event ID	Date	Substance name	Release quantity (lbs)	Quantity category	Victims	Equipment/ Facility	Release Type	Industry
A	01/18/2002	Diphenyl	257	100 - 999	11	Ancillary process equipment	Volatilization/ aerosolized	Petroleum refining
B	01/18/2002	Hydrogen sulfide	100	100 - 999	16	Dump/waste area	Volatilization/ aerosolized	Pulp, paper, & paperboard mills
C	02/11/2002	Ammonia	180	100 - 999	0	Process vessel	Volatilization/ aerosolized	Industrial & misc chemicals
D	02/05/2002	Sulfuric Acid	5000	1000 - 9,999	0	Process vessel		Industrial & misc chemicals
E	03/02/2002	Benzenesulfonyl hydrazide	245	100 - 999	0	Process vessel	Volatilization/ aerosolized	Plastics, synthetics, & resins
F	03/12/2002	Methyl Mercaptan	213	100 - 999	0	Process vessel		Pulp, paper, & paperboard mills
G	03/29/2002	Hydrogen sulfide	113	100 - 999	0	Process vessel	Volatilization/ aerosolized	Pulp, paper, & paperboard mills
H	04/11/2002	Ammonia	30,000	10,000 - 99,999	0	Ancillary process equipment		Industrial & misc chemicals
I	04/06/2002	Methyl Mercaptan	126	100 - 999	0	Process vessel		Pulp, paper, & paperboard mills
J	05/09/2002	Ethyl acrylate	775	100 - 999	0	Process vessel	Volatilization/ aerosolized	Miscellaneous fabricated metal products
K	08/04/2002	tert-Butyl alcohol	920	100 - 999	0	Storage above ground	Volatilization/ aerosolized	Industrial & misc chemicals
L	08/07/2002	Ammonia	339	100 - 999	0	Process vessel	Volatilization/ aerosolized	Industrial & misc chemicals
M	08/19/2002	Sodium Hydroxide	1200	1000 - 9,999	0	Storage above ground	Spill	Miscellaneous plastics products
N	08/30/2002	XYLENOL	5800	1000 - 9,999	0	Storage above ground	Spill	Industrial & misc chemicals
O	08/30/2002	BENZENE	26	10 - 99	0	Process vessel	Volatilization/ aerosolized	Industrial & misc chemicals
P	09/09/2002	ACETONE	175	100 - 999	0	Process vessel	Explosion	Industrial & misc chemicals
Q	09/19/2002	Methylene Chloride	2000	1000 - 9,999	0	Piping		Miscellaneous plastics products

3. METHODOLOGY

3.1 Research Methodology

This research analyzed the HSEES database using data and text mining algorithms as shown in Figure 4. The first step of the data analysis is data and variable selection. As mentioned in the previous section, the analysis focuses on a subset of HSEES data pertinent to incidents which occurred inside industrial facilities. The selections of the variables were made to include the causal factors and consequence of the incidents and other relevant information. The incident reports containing sparse data were omitted.

Once the data has been selected, an exploratory analysis was performed using statistical means. Selected variables were plotted and observed to get the feel of the data. Then, findings from data exploration step were used as a reference in conducting the data and text mining process.

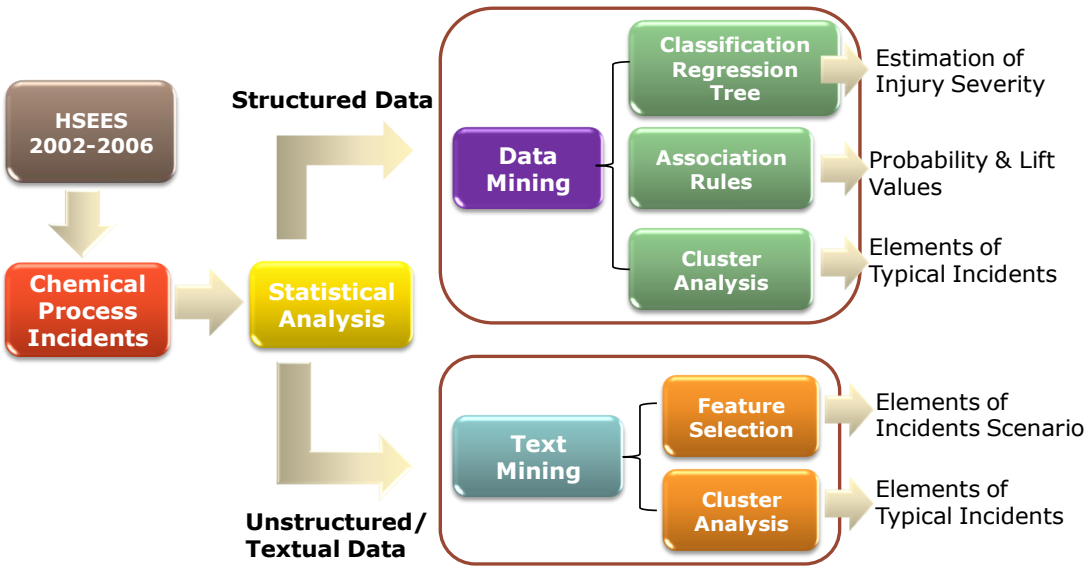


Figure 4: Research methodology

Prior to conducting data mining, the selected data must undergo some preparation which involved cleaning the data of duplicate entries and formatting the data attributes. Finally, the data was analyzed using data mining tools such as cluster analysis, association rule, and classification and regression tree (CRT). The clustering process focuses on determining if the computed groupings have meaningful values and explicit characteristics (Cerrito, 2006). The association rule identifies patterns of events which attributes are associated. The pattern values were then compared with the patterns found in previous research. The decision tree was performed to evaluate the relationship between a numbers of predictor variables that determine the severity of an incident.

Text mining was performed to analyze the textual data (unstructured data). HSEES has a variable called “comment” that contains brief description of the incidents and often provides information that the other variables (structured data) do not. Therefore, there is an opportunity to explore the textual data in order to obtain valuable information. A detailed explanation on data and text mining is given in the following sections.

3.2 Data Mining

Data mining is used as a tool to analyze the HSEES database due to its ability to compute large data, perform multivariate analysis and produce predictive models. Data mining is commonly used to identify patterns, associations or relationships in the data variables (Edelstein, 1999). One of the benefits of using data mining is that relationships or patterns that are neither obvious nor noticeable can be identified.

3.2.1 Cluster Analysis

Cluster analysis is an unsupervised data mining algorithm where the dependent variable is not specified (Cerrito, 2006). The clustering focuses on grouping members with similar attributes and finding meaningful clusters that are distinctive of each other.

The interpretation of the clusters is a subjective process, depending on the interest and perspective of the analyst. The clustering of HSEES data was performed to produce meaningful groups of incidents that share similar characteristics. The incident groups can then be used to describe typical incidents reported to HSEES.

The idea behind cluster analysis is to separate data points into groups or clusters so the total variation among the incident reports is minimized within each group and maximized between groups. The degree of variation is evaluated using the Euclidian distance shown in Equation 1. Using the geometric distance of each case in the multi variable space as the objective function, the clusters are optimized to minimize the distance between members in each cluster and to maximize the distance between clusters (Hand et al., 2001).

$$d_E(i, j) = \left(\sum_{k=1}^{p} \left(x_k(i) - x_k(j) \right)^2 \right)^{1/2}$$

k-means clustering was selected for the analysis because it can handle data larger than 250 points. In *k*-means clustering, the user can assign a *k*-number of clusters and observe the clusters profile to determine whether the clusters have meaning. Another way to perform *k*-means clustering is to let the software optimize the number of clusters, given the range of cluster set by the users.

k-means clustering analysis is typically used for continuous variables, however it can also process categorical variables in which all distances are binary (0 or 1). The variable is assigned 0 when the attribute of the data point is the same as the attribute with the highest frequency in a cluster, otherwise it will be assigned 1 (Nisbet et al., 2009). Once the number of clusters is optimized, the profile of each cluster should be evaluated to identify their characteristics and to determine whether the clusters are reasonable (Cerrito, 2006).

3.2.2 Association Rule

The association rule is a data mining tool used for extracting patterns and associations between the variables. The association rule, first known as the market basket analysis, was used to identify items that customers bought together frequently. These purchasing patterns were then used as considerations for marketing strategies such as creating specials or bundling option for multiple items, etc.

Association rules obtained the patterns by identifying the occurrences of two or more variable attributes and their relative frequency of occurrence. The set of the association rules consists of a left-hand side proposition (the antecedent) and a right-hand side proposition (the consequent), which are presented using the following form:

"*If* event A occurs, *then* B occurs with a probability of *x,* and this pair of events occurs with a probability of *y* in all of the events."

The parameters that quantify the rule set are *x* and *y*, which are the confidence and support, respectively. Further expression and explanations of the parameters, are given as follows:

- The support value is computed as the joint probability or relative frequency of events A and B occurring simultaneously.

$$\textbf{Support} = \frac{\textbf{Number of events both A and B occurred simultaneously}}{\textbf{Total number of events}} = \mathbf{P(A \cap B)}$$

- The confidence value is the conditional probability of event B occurring, given event A has occurred. It indicates the ratio of the probability of the antecedent (A) and the consequence (B) occurring simultaneously with the probability of the antecedent (A).

$$\textbf{Confidence} = \frac{\textbf{Support}}{\textbf{Probability of A occurring}} = \frac{\mathbf{P\ (A \cap B)}}{\mathbf{P(A)}}$$

- The lift value is a measure of the likelihood of B occurring given A has occurred relative to the likelihood of B occurring independently.

$$\textbf{Lift value} = \frac{\textbf{Confidence}}{\textbf{Probability of B occurring}} = \frac{\frac{\mathbf{P}(\mathbf{A} \cap \mathbf{B})}{\mathbf{P}(\mathbf{A})}}{\mathbf{P}(\mathbf{B})}$$

3.2.3 Classification and Regression Trees

A classification and regression tree (CRT) is a data mining algorithm that is used to classify or estimate a dependent variable based on predictor variables, which either can be categorical or continuous. The relationship between the variables is organized into a tree-like structure where the root node is split into two or more branches. Each branch represents classes or ranges of the root node. The splitting process continues until certain stopping rules are satisfied (Nisbet et al., 2009). The final nodes of the CRT are called the terminal nodes.

CRT is used to study the relationship between the predictor variables and the dependent variable because of its ability to perform piecewise regressions and produces easy to understand results. The piecewise regression accommodates the presumption that not all incidents are the same but at the same time it attempts to estimate the outcome using predictor variables. The objective of the CRT is to partition the data at a point (node), so it produces subsequent branches that fit the piece regressions with minimum error or that give a maximum R-squared value.

The splitting criterion evaluates the reduction in the distribution of the dependent variable between subsequent node and the root node (Matignon, 2007). There are two steps in the splitting process, which are determining the best split for each input variable and choosing the best split that considers the multiple input variables. The node where the splitting begins is selected based on its improvement of predictive accuracy, which is measured by node impurity (Statsoft, 2008). For regression cases, a least squared deviation criterion is applied to measure node impurity. The least squared deviation is computed as:

$$R(t) = \frac{1}{N_w(t)} \sum_{iet} w_i f_i (y_i - \bar{y}(t))^2$$

where, $N_w(t)$ is the weighted number of cases in node t, w_i is the value of weighting variable for case i, f_i is the value of frequency variable, y_i is the value of response variable, and $y(t)$ is the weighted mean for node t.

The stopping rule is an important parameter in constructing the tree because the size of the tree affects the interpretation greatly. A tree that is too small results in an unreliable estimate, while a tree that is too large results in overfitting the data (Matignon, 2007). The stopping rules used in this research is prune on variance, which means that the tree will stop splitting when the variance of the current tree is better than the tree with further splits.

Incident severity is a function of many different factors, such as quantity of release, weather conditions, properties of the chemicals released, distance of victims in respect to the source of release, susceptibility of the people, and other factors. Hence, it cannot be expressed using a simple model. However, it is still of interest to study how significant several factors, such as release quantity and distance of victims in respect to the source of release, affect the severity of incidents. Complex factors, such as interaction between variables or the domino effect, may not be presented well in the tree model.

3.3 Text Mining

Text mining is a process of analyzing textual data (unstructured data), by extracting meaningful numeric indices from the text. The transformation from text data to numeric indices allows the data to be processed further using data mining algorithms (Statsoft, 2008). The information from text data can be used to derive summaries for the words contained in the documents or the summaries of the documents (Nisbet et al., 2009).

The steps of text mining are shown in Figure 5 and further explanations of each step are provided in the following sections.

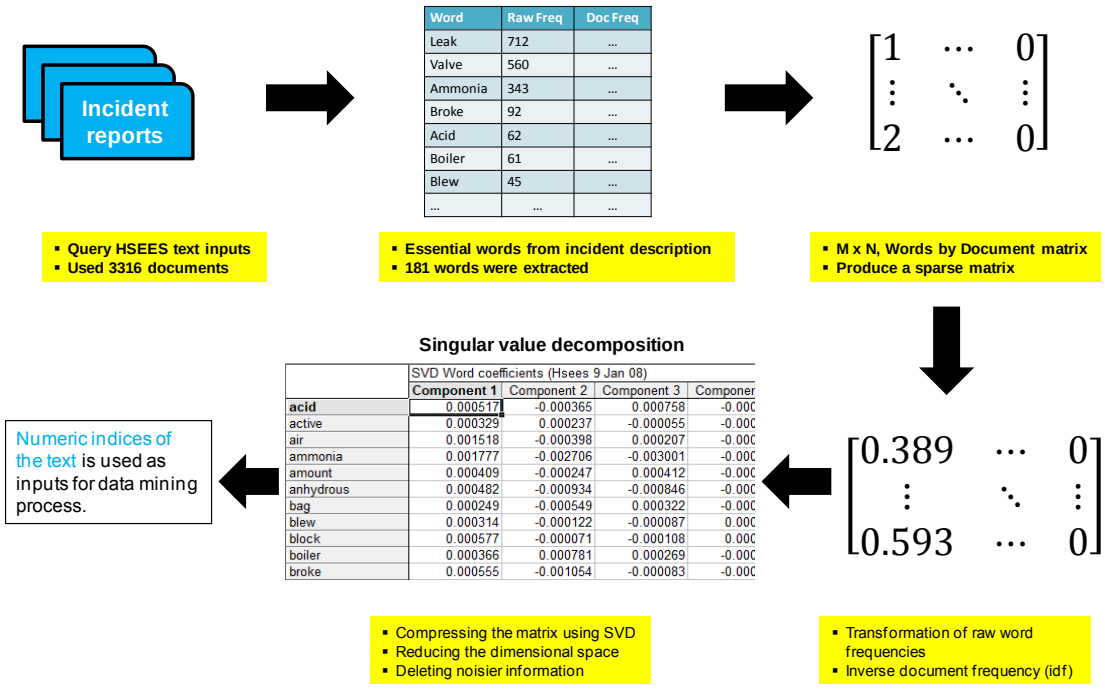


Figure 5: Text mining algorithm

3.3.1 Preparation of Text Data

Before applying text mining, there are several steps to be performed in order to prepare the text data. The following steps reduce and filter the existing words in the incident reports:

- ***Creating a stop words list***

  Stop words are considered non-essential words that do not help in distinguishing the information that pertains to this work. Therefore, these words need to be excluded from the indexing process. Example of English stop words are words such as conjunctives, articles, auxiliary verbs, prepositions, etc. This research

edited the Statistica built-in English stop words list to include additional non-essential words observed in the incident descriptions. The examples of the stop words are incident, event, notification, release, response, caller, etc shown in Appendix.

- ***Creating a start words list***
  Start words lists are lists of words that give significant meaning to the analysis, and in this case are the opposite of the stop words lists. Start words list can be created to ensure that the essential or particular words of interest are included in the indexing process. This research did not use a start words list and used stop words list to extract more words from the document.

- ***Stemming and identifying phrases and synonyms***
  Stemming is a data preparation step for reducing words to their roots so that different grammatical forms of a word are identified and treated as the same word. Phrases and synonyms of the stop words can be specified so that they are excluded from the indexing process as well. The phrases used in this text mining are shown in Appendix.

3.3.2 Indexing and Transforming Word Frequencies

The words from the documents (incident reports) were extracted and each of the selected word frequencies was computed, as shown in Table 2. The full list of extracted words is shown in Appendix. In general, the raw word frequencies can be used as a parameter that reflects how salient a word is in every document. However, the importance of the word itself cannot be determined based on its frequency alone, therefore the word frequencies need to be transformed into a form that accounts for relative importance of the word in all of the documents.

Table 2: Words extracted using inverse document frequencies

Word-*i*	Word	Frequency
1	Leak	712
2	Valve	560
3	Line	392
4	Ammonia	343
5	Failure	338
6	Flare	329
7	Process	285
…	…	…
…	…	…
180	Broken	35
181	Smell	34

This research used inverse document frequencies because it covers both the relative frequencies of word occurrence and the word's semantic specificities in the documents. The inverse document frequencies (*idf*) transformation takes into account the relative document frequencies (*df*) of different words. This word frequencies transformation accounts both the specificity of the words (document frequencies) and the overall of the words (word frequencies) for the word *i* and document *j* respectively (Statsoft, 2008):

$$idf\ (i,j) = \begin{cases} 0 & if\ wf_{i,j} = 0 \\ \left(1 + log\left(wf_{i,j}\right)\right) \log \frac{N}{dfi} & if\ wf_{i,j} \geq 1 \end{cases}$$

Where *N* is the total number of documents, *dfi* is the document frequency for the word *i*. The formula includes dampening of words frequencies and a weighting factor to evaluate a word's relative occurrence. The weighting factor *(log (N/dfi))* is valued at 0 (minimum) if the word occurs in all documents, and valued at 1 (maximum) if the word only occurs in one document (Statsoft, 2008). The transformed word frequencies then are used in further text mining computation such as SVD calculation.

3.3.3 Latent Semantic Indexing

A vector space model can be used to represent the words and documents extracted from the text data. Table 3 shows examples of words and documents used in this research. The words and documents are supposedly represented in matrix A, where the matrix columns correspond to the documents and the matrix rows correspond to the words of the text, as shown in Figure 6. Matrix A would then be normalized to produce matrix values shown in Figure 7.

Table 3: Words and documents of HSEES incident descriptions

Word	Document
W1: Leak	D1: Flange leak occurred in a section of 4 inch pipe on the feed system heat exchangers. Material released was a liquid with some atomized droplets, which caused the material to vaporize and disperse.
W2: Tank	D2: When aligning a pipe to a well, a drain valve was not closed completely causing release.
W3: Pipe	D3: Pressure in vessel was too high, causing ammonia to be released from relief valve.

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

Figure 6: The 3x3 words by document matrix

$$A = \begin{bmatrix} 0.389 & 0 & 0 \\ 0 & 0 & 0 \\ 0.593 & 0.774 & 0 \end{bmatrix}$$

Figure 7: The normalized matrix A

The extraction of words and documents can be performed using the exact word and document match but also using other relevant documents that fit the context. Therefore, latent semantic indexing (LSI) is used to understand the semantic space of the words, beyond literal words matching. LSI employs vector space representation of both words and documents to find other documents relevant to a word (Berry & Browne, 1999).

In the case of representing large incident text data in word and document matrix, the resulting matrices would be large and sparse, with many zeros showing that many words only appear in a few documents. Therefore, the matrix size can be compressed such that it still retains useful information and is more efficient in representing the words and documents. A matrix decomposition algorithm such as singular value decomposition (SVD) can be used to reduce the matrix size.

SVD reduces the overall dimensions of the input matrix A by extracting the common semantic space of the data (Han & Kamber, 2006). The reduced dimensional space should represent the largest degree of variability between words and documents so that the latent semantic space that organizes the words and documents can be identified. Hence, singular value decomposition can determine the few underlying dimensions that account for most of the contents or meaning of the document and words that were extracted (Statsoft, 2008).

The singular value decomposition theorem states that matrix A can be decomposed as follows:

$$A = USV^T$$

U and *V* are orthogonal matrices, where $U^TU = I$ *and* $V^TV = I$. *S* is the diagonal matrix with singular values. The SVD calculation consists of evaluating the eigen value and eigen vectors, AA^T and A^TA in respective. The eigen value AA^T is presented as column *V*, the eigen vectors A^TA is represented as column *U* and the singular values in *S* are calculated as the square roots of eigen values from AA^T or A^TA (MIT, 2002).

Once the SVD computation has been performed, the results can be visualized using a scree plot. The scree plot indicated the number of components that are useful by locating the elbow of the plot, the point where the plot decreases smoothly and the singular values becomes steady. After the text data has been processed numerically, the data can be analyzed using common data mining tools such as cluster analysis, feature selections, CRT, etc.

3.3.4 Feature Selection Tools

The feature selection tool can be used to identify important predictors that have strong relationships to the dependent variables. This relationship is based on the presence of the predictors in respect to the dependent variables in all of the documents. The feature selection tool produces bar plots where the importance of the predictors are ranked based on their F-values.

3.3.5 Cluster Analysis Using Text Inputs

Cluster analysis was performed using the component scores of the words and documents produced by the SVD. The same cluster analysis concept explained in section 3.2.1 applies to the component scores. The component score can be used to perform cluster analysis on words that relate to each other (Raja & Tretter, 2010), and furthermore to evaluate the typical incidents that are reported to HSEES. The results of this cluster analysis will be evaluated and compared to cluster analysis without text.

4. RESULTS AND DISCUSSIONS

The following analysis used HSEES data pertinent to chemical process incidents that occurred from 2002 to 2006. This subset of HSEES data was selected because it was relatively recent, the number of states participated were comparable and the structure of the reporting system during this HSEES active period was uniform. The HSEES data prior to this period had a slightly different format and terms used as attributes of the variables.

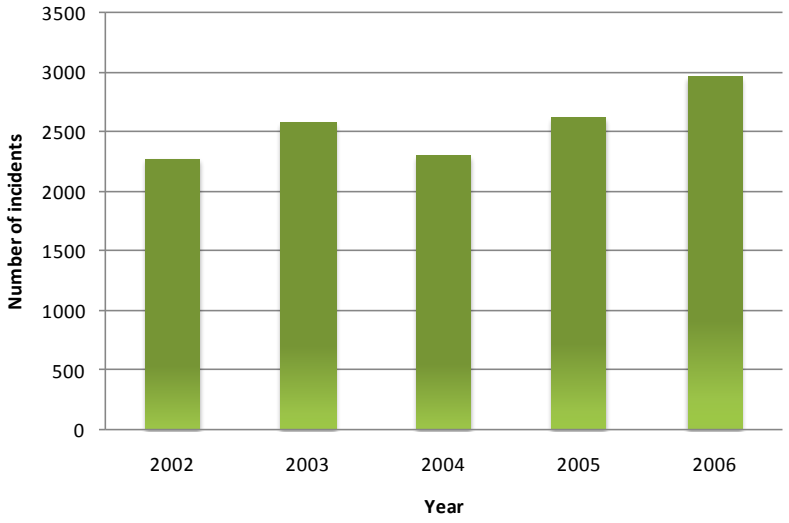
4.1 Statistical Analysis

The statistical analysis provided fundamental information of the HSEES data, focused the research to statistically significant variables, and helped identify patterns in the data. The data went through a cleaning process, where incidents related to emission releases, duplicate entries, and incident cases with high numbers of missing attributes were eliminated. The cleaning process set boundaries so the analysis only focused on incidents related to loss of containment that occurred in the industrial area.

4.1.1 Trend Analysis

Figure 8 shows the number of incidents that occurred from 2002 to 2006. The average number of chemical process related incidents that occurred from 2002 to 2006 was approximately 2,500 incidents per year (only include data from 15 states, only incidents occurring within the chemical facilities boundaries are included, petroleum-only incidents are excluded). Throughout this period there was an average of 15 states participating in HSEES incident collection system. In order to observe the tendency of an incident's occurrence, these figures need to be normalized by the number of facilities participating each year. However this type of information was not provided by the database. Assuming that the number of facilities reporting to the HSEES each year was

constant, it can be observed that the number of chemical process related incidents rose slightly.



Year	2002	2003	2004	2005	2006
Participating states	15	15	15	15	14

Figure 8: Chemical process related incidents and number of participating states

Figure 9 shows the distribution of major and minor injuries from 2002 to 2006. The HSEES severity classification was used to define major and minor injuries (HSEES collection form). Major injuries are defined as any incident consequence where the victims were transported to, admitted to and treated at a hospital, or transported to, not admitted to, and treated at the hospital. Minor injuries are defined as incident consequences, in which the victims experienced injuries within 24 hours, sought private physician service within 24 hours, were treated on the scene or received first-aid help.

The number of injuries, both major and minor, reported to HSEES increased from 98 in 2002 to 466 in 2006. The number of injuries from 2002 to 2005 was relatively steady,

and there was a sharp increase in reported injuries in 2006, up 150% compared to the previous years. These increments could be either a sign of an actual rise in the number of injuries, an increase in the number of facility participating in the HSEES reporting system or an indicator of growing safety awareness in the industries which made the participating facilities report more incidents to the HSEES system. In order to understand this observation, information on the number of facilities reporting and the amount of regulation changes made in this period needs to be available and accounted for.

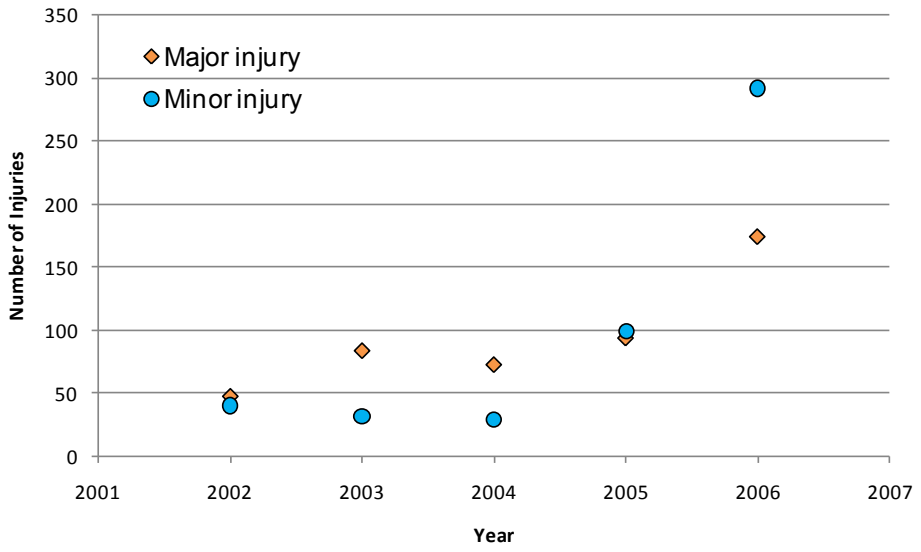


Figure 9: Number of injuries reported to HSEES in 2002 – 2006

As shown in Figure 10, the number of fatalities reported to HSEES fluctuated from 5 in 2002 to 3 in 2006. There was no obvious trend observed. The analysis can be enhanced using the descriptions of the incidents with fatalities. The text description can be analyzed using text mining to study the common factors which indicate fatalities (Bunn et al., 2008).

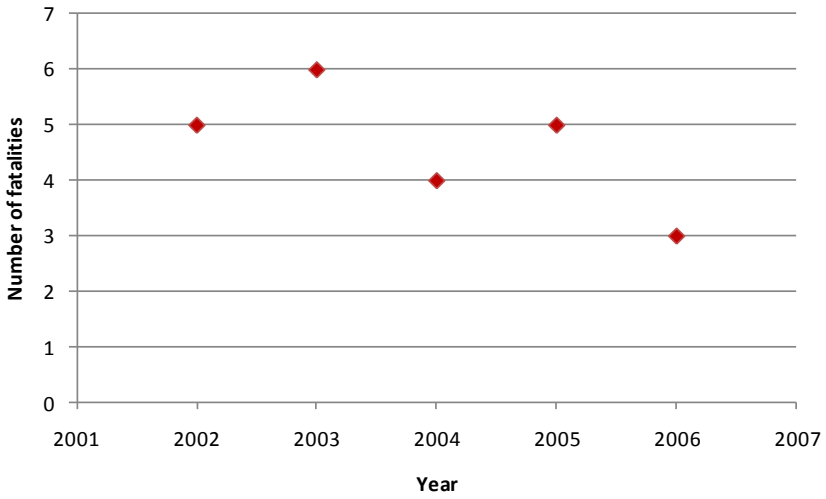


Figure 10: Number of fatalities reported to HSEES in 2002-2006

Figure 11 shows the types of incidents reported to HSEES. The types of HSEES incidents were categorized based on the physical state of the chemical released such as vapor releases (which include vapor, gas and aerosol) and spill releases (which include liquid and solid). The events following the chemical release such as fire or explosion were also classified in the same variables as vapor and spill releases. This means that HSEES treats fires and explosions as causal factors for the incidents instead as consequences. The major type of incidents involved vapor releases, which consist of 79% of the total incidents, and spill releases which consist of 19% of the total incidents. Fire and explosion releases each have slightly less than 1% of the total number of incidents.

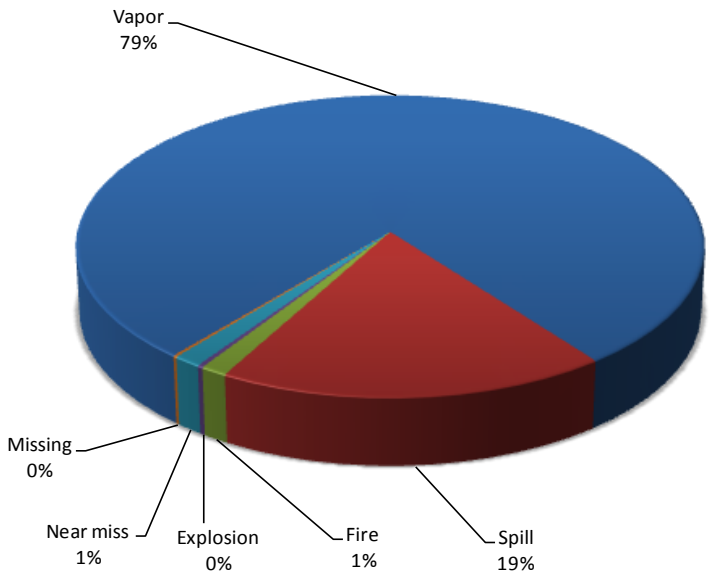


Figure 11: Types of incidents reported to HSEES

Figure 12 shows the distribution of the contributing causes of the incidents reported to HSEES. Equipment failure was the mode of immediate contributing causes, comprising 63% of the total number of incidents, followed by human error at 15%, deliberate damage at 14%, natural disaster at 4% and unknown and others at less than 5%. The current analysis can be further enhanced by using the text comment variable, which provides a succinct description of the incidents. The utilization of text comments can produce analysis of the types of releases, equipment where the release came from and the process involved in the incidents.

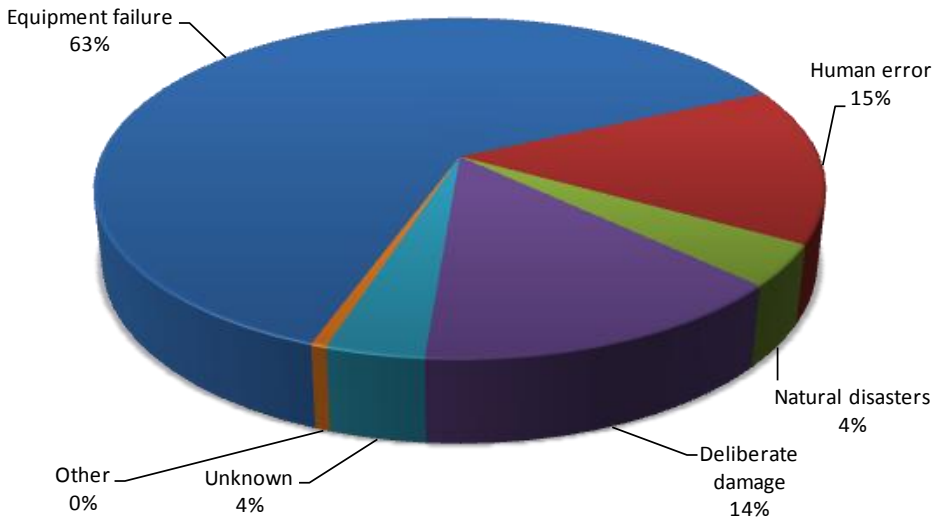


Figure 12: Contributing causes reported to HSEES

Figure 13 shows the Pareto chart of the ten chemicals frequently released. The release of these chemicals constituted about 35% of the total incidents used in this research. The majority of the incidents involved Ammonia release, approximately 1,536 incidents or 12% of the total number of incidents. Incidents involving Benzene, Chlorine, Freon and Vinyl Chloride followed at 516, 479, 378 and 347 incidents, or 4%, 3.8%, 35 and 2.7% of the total number of incidents, respectively. The remaining chemicals were released in a relatively low numbers of incidents and could not all be mentioned due to the hundreds of variety of chemicals involved.

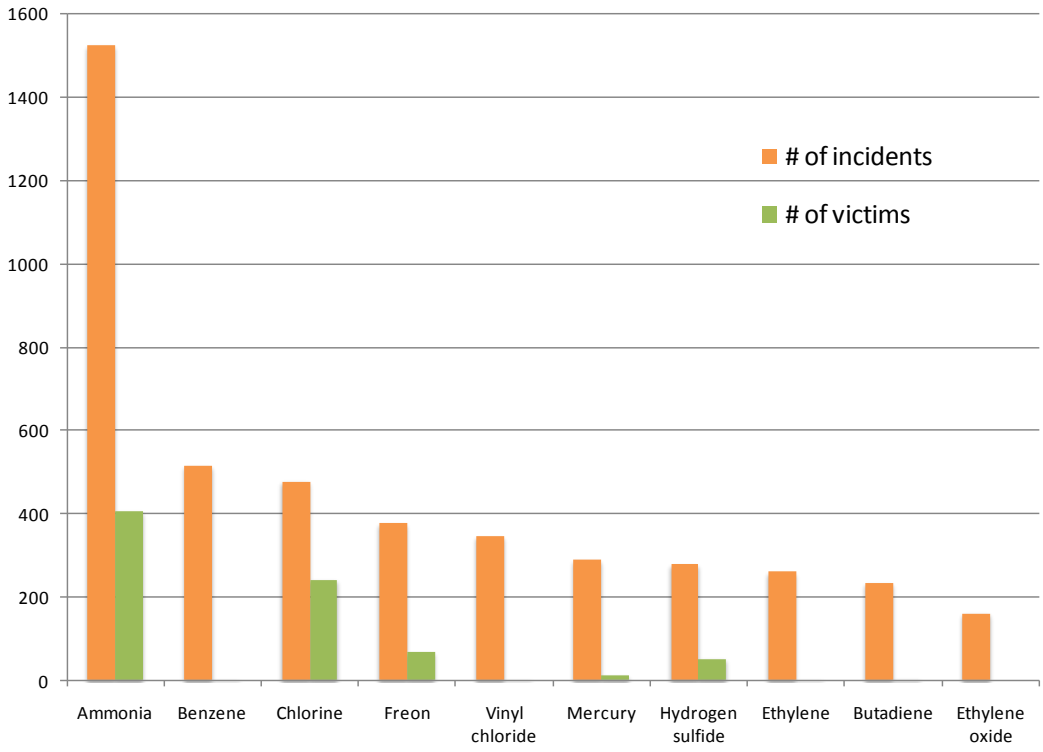


Figure 13: Chemicals frequently reported to HSEES in 2002-2006

Figure 14 shows the ratios between the number of victims and the number of incidents derived from Figure 13. Chlorine had the highest ratio of number of victims to number of incidents at 0.5, which translated to 1 victim for every 2 Chlorine-related incidents. Ammonia had a ratio of 0.26 which translated to 1 victim for every 4 Ammonia-related incidents. Similar ratios for Hydrogen sulfide, Freon and Mercury are 0.19, 0.18, and 0.05 respectively.

This ratio however cannot be used directly as a measure of injury rate for each chemical because there are several factors that influence the consequence of incidents, such as the chemical dose exposed to the victims, including both concentration and time of exposure, the toxicity of the chemicals, personal protection equipment worn by the victims, etc.

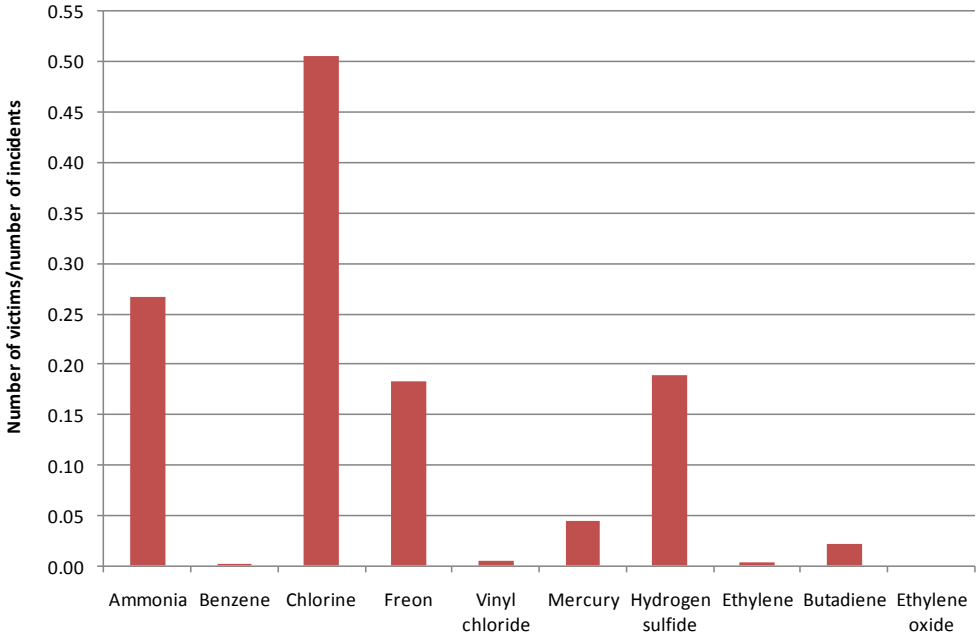


Figure 14: The ratio of the numbers of victims and the number of incidents

The previous analysis can be enhanced using information about the source of the release, either the type of process or the type of equipment. However, HSEES did not collect this type of information. This can be added to points of consideration for developing recommendations on ways to improve chemical incident databases.

The variable 'quantity of chemicals released' was distributed through a large range of quantities due to the large variety in the quantities reported to HSEES and the uncertainty of this variable. Figure 15 shows the distribution of the release quantity and the number of victims for each quantity category. The most common release quantity was in the range of 100 to 999 lbs, where 3,591 incidents, approximately 28% of the total number of incidents, were reported in this category. The release quantity category of 10 to 99 lbs and 1,000 to 9,999 lbs, approximately 21% and 18% of the total number of incidents reported, were also reported among the frequently released quantities.

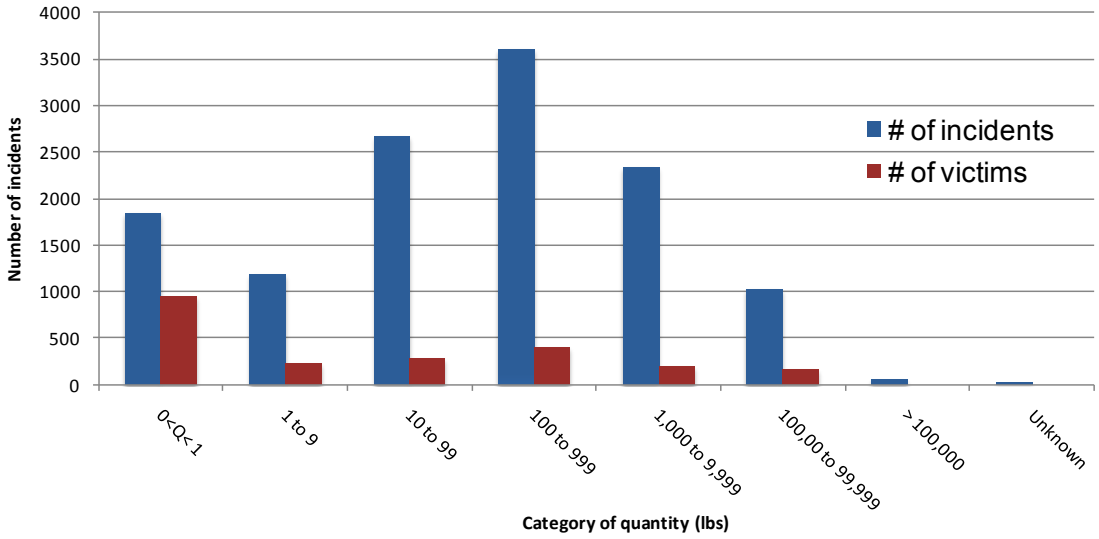


Figure 15: Distribution of the release quantity in 2002-2006

Figure 16 shows the ratios between the numbers of victims and the number of incidents for each release quantity category derived from Figure 15. The incidents where a chemical was released under a quantity of less than 1 lb had the highest number of victims over number of incidents ratio. It is interesting to consider the fact that HSEES did not require chemical released less than 10 lbs or 1 gallon to be reported; however the analysis showed a relatively high number of incidents and number of victims reported in this category. This may indicate that the industry and the public in general were becoming more conscious about the importance of reporting incidents regardless of the incident reporting criteria.

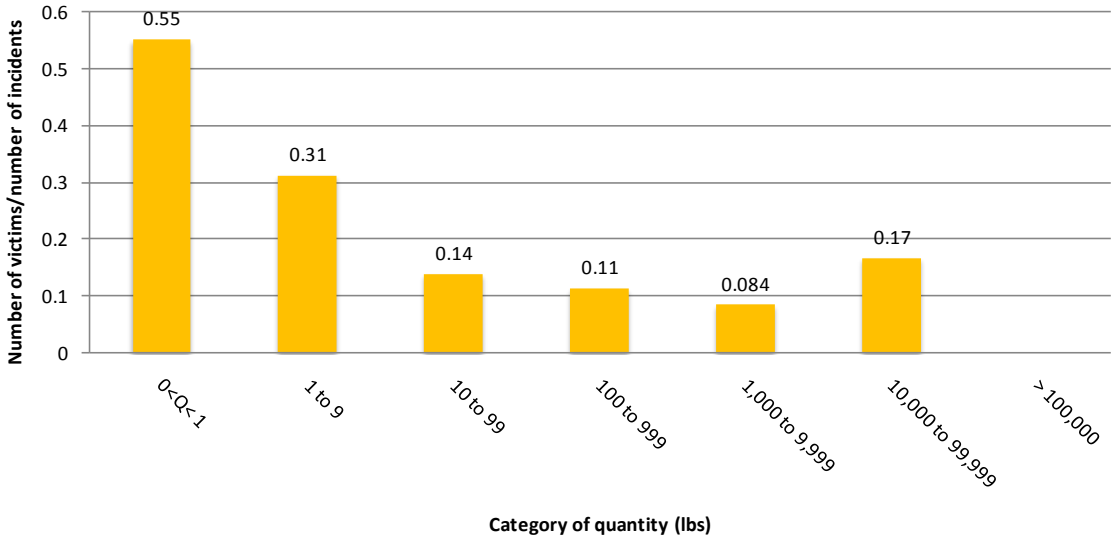


Figure 16: Ratio of number of victims and number of incidents in HSEES incidents

The next highest ratios of number of victims over number of incidents were incidents with chemical releases in quantities of 1 to 9 lbs, 10,000 to 99,999 lbs, and other categories. In order to evaluate how the quantities of the chemicals released influences the outcome of the incidents, severity factors need to be included in the analysis. Such analysis will be discussed in the next section.

4.1.2 Relationship between Release Quantity and Severity of Incidents

The research aims at studying the relationship between the quantities of the chemical released and the severity of injury of incidents. Provided that other variables are comparable, it is assumed that the quantity of chemical released is proportional to the incident consequences in terms of the adverse effects on human health, involving the number of injuries or fatalities. Thus, the HSEES severity data was used to justify this assumption. The types of severity of health effects of the incidents reported to HSEES were given in Table 4.

Table 4: Severity of health effects of the incidents reported to HSEES

Severity	Translation	Severity Index
Treated on the scene or received first-aid	Minor injury	0.1
Observation at the hospital, no treatment	Minor injury	1
Injuries within 24 hours, reported by officials	Minor injury	1
Seen by private physician within 24 hours	Minor injury	1
Unknown	Minor injury	1
Treated at the hospital, not admitted	Major injury	10
Treated at the hospital, admitted	Major injury	50
Death on scene or after arrival at the hospital	Fatality	100

The severities of the incidents were evaluated based on a scoring system, which gave quantification to the consequences of the incidents (MKOPSC, 2006). The scoring system does not reflect the value of life or injury but merely helps differentiate the significance of each type of incident. Figure 17 shows the relationship between the quantity of the release and the severity of the consequence of the incidents, which was normalized by the number of respective incidents.

With the assumption that all incidents have comparable causal factors, chemicals and conditions, it can be presumed that an incident with a higher release quantity would have a higher severity score. However, the incidents where chemicals were released under the category of < 1 lb have the highest normalized severity score, followed by the incidents releasing chemicals under the category 1 to 9 lbs. This indicates that causal factors such as the dose exposed to the victims, the distance of the victims from the source of release, the type of personal protection equipment (PPE) worn by the victims and other factors affect the outcome of the incidents. This finding can be used to justify the importance of in-depth incident investigations into incidents with high severity scores. A further study of the relationship between the release quantity and the severity of the incident's consequence is presented in section 4.2.3.

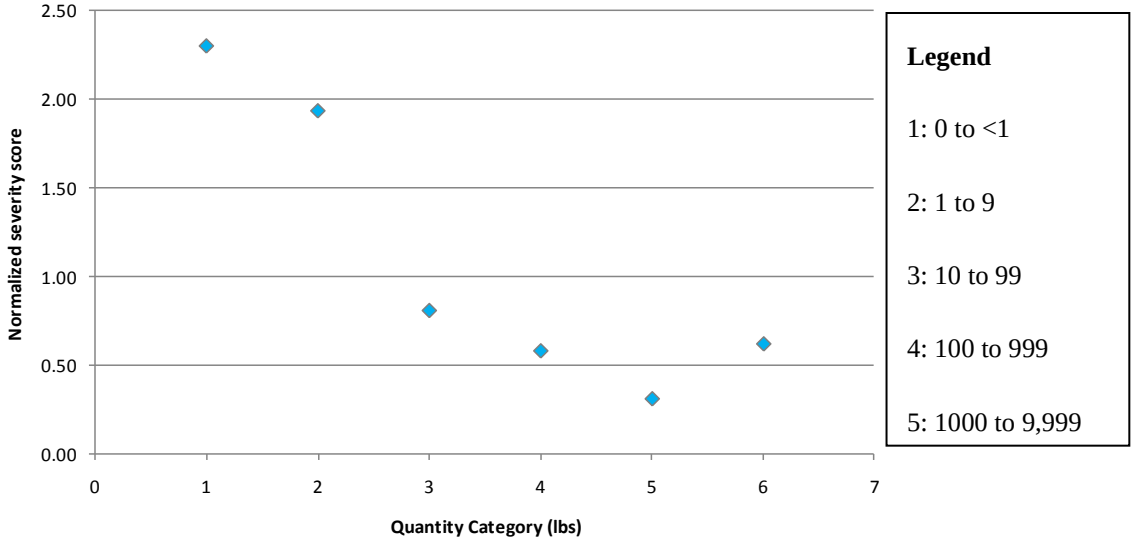


Figure 17: Severity score of the incidents based on the release quantity

4.2 Data Mining Results

Data mining is a method used to analyze large data, using multivariate analysis to discover patterns, associations or relationships between variables in the data. There are several data mining techniques used in this research such as cluster analysis and the association rule, each of which will be further explained in the following sections. All the data mining analysis was performed with the aid of STATISTICA™ data mining software.

4.2.1 Cluster Analysis Results

The cluster analysis was performed using several combinations of variables to produce meaningful clusters. The clusters shown in the following figures were generated using four variables that describe: 1) chemicals released, 2) physical state of the chemicals released, 3) equipment or area involved in the incidents and 4) industry, where the

incident occurred. Through iteration, the cluster analysis produced 2 clusters with attributes shown in Figures 18 through 20. The generated clusters were not completely discernable from one another, because all attributes can be found in each cluster. However, the clusters can still be differentiated from each other by evaluating and comparing their major attributes.

Figure 18 shows the incident data segmented based on the equipment involved in the respective incident. Incidents included in cluster 1 (shown in blue) can be characterized as incidents where the chemical was released from piping. Incidents involving piping consists of 52% of the total incidents in cluster 1. Incidents included in cluster 2 (shown in red) primarily reported ancillary process equipment and process vessels as the source of the chemical release. Both of these pieces of equipment consist of 65.7% and 28.6% of the total incidents in cluster 2.

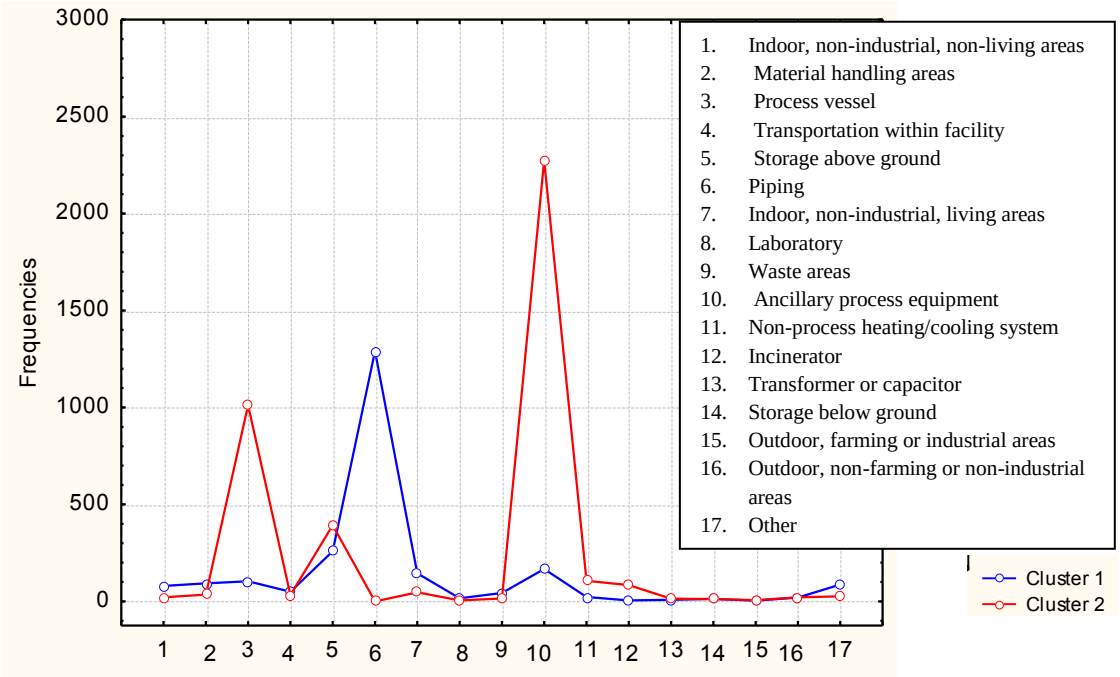


Figure 18: Clusters viewed from the perspective of equipment involved

Figure 19 shows the segmented incidents viewed from the perspective of the physical state of the chemicals released. Incidents grouped as cluster 1 (shown in blue) primarily released chemicals in their liquid phase, whereas spill releases consist of 50% of the total incidents in this cluster. Incidents in cluster 1 also have a number of vapor releases which consist of 40% of the total incidents included in this cluster. In comparison, 90% of incidents grouped in cluster 2 involved chemicals released in a vapor state. Based on the clustering, it can be concluded that incidents in cluster 1 released chemicals in their liquid state while incidents in cluster 2 released chemicals in their vapor state.

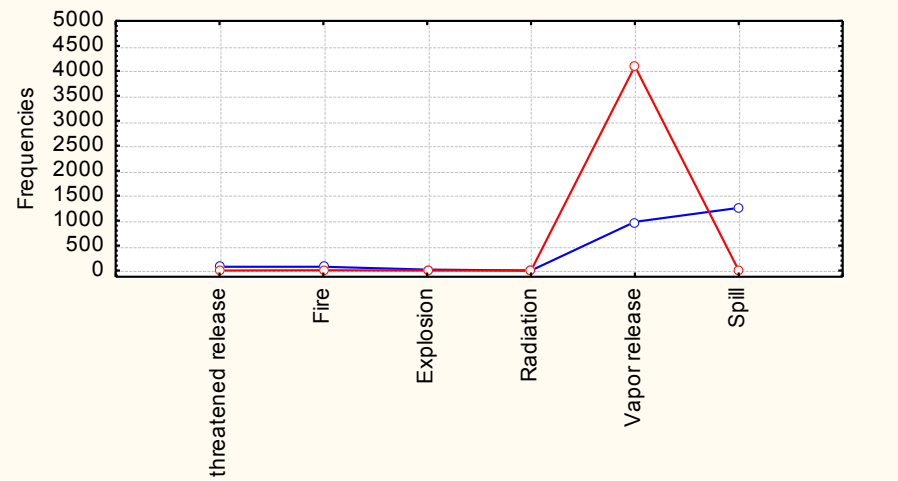


Figure 19: Clusters viewed from the physical state of the chemicals released

Figure 20 shows the segmented incidents viewed from the industry where it occurred. Segmenting the incidents using this variable did not give discernable clusters because comparable composition of the attributes was observed for both clusters.

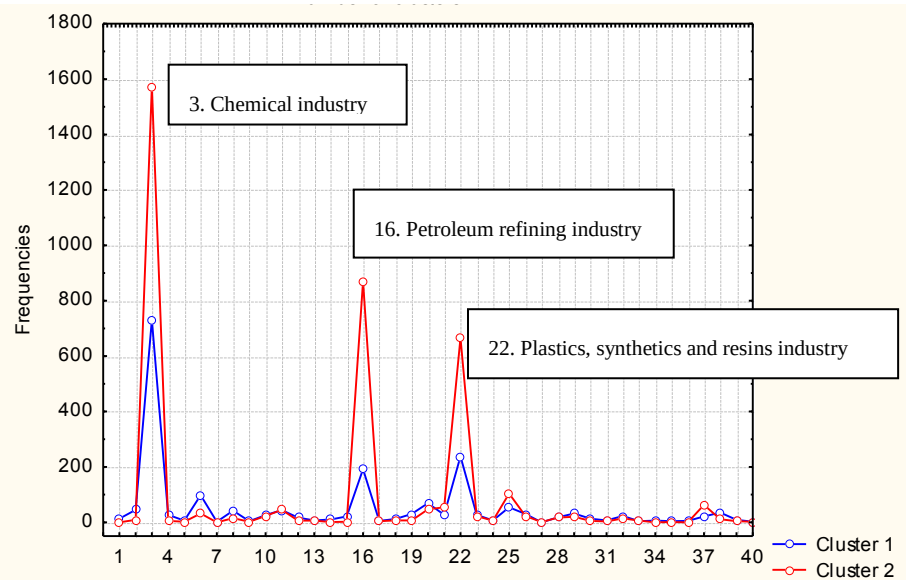


Figure 20: Clusters based on industry type

The incidents were also segmented using a variable that describes the chemical released. The frequency graph cannot be generated because there were too many chemicals covered in the analysis.

The significance of clustering is that one can expect a certain scenario by having information on one or more variables of an incident. In this case, the cluster analysis pointed out that the data can be segmented into two clusters with the following characteristics: incidents which occurred in the piping system which usually released as liquid (spill) or incidents which occurred in the process vessel or ancillary equipment and released as vapor. Furthermore, the cluster members can be used in further analysis to produce more accurate results because of their homogeneous characteristics.

4.2.2 Association Rule Results

The association rule method was performed using four variables that describe the chemicals released, equipment or area in the facility from where the chemical was released, the physical state of the chemical released and the type of industry where the incident occurred. The computation of the association rule was performed using data mining software, STATISTICA version 8.0 and which required the user to input predefined support and confidence values in order to generate the pattern.

Table 5 shows the rules or patterns identified using the association rule method viewed from the perspective of the industry type where the incidents occurred. As can be observed in pattern numbers 4 through 6, the probability of incidents occurring in the chemical industry and also involving ancillary process equipment was calculated as 0.072. The confidence of this pattern, ratio between the probability of incident and this particular set of attributes occurring to the probability of all incidents occurring in the chemical industry, was observed as 0.39.

The other observed patterns include the probability of incidents occurring in the chemical industry which also involved process vessels and piping. The confidence values for both patterns were observed in the same value of 0.23. This means that the likelihood of chemical industry incidents involving process vessels is comparable to that of piping.

Table 5: Incident pattern in regard to the industry where the incidents occurred

No.	A	B	P (A∩B)	P (B | A)
1.	Chemical industry	Vapor release	0.15	0.85
2.	Petroleum refining	Vapor release	0.078	0.94
3.	Plastics, synthetics, and resins	Vapor release	0.063	0.89
4.	Chemical industry	Ancillary process equipment	0.072	0.39
5.	Chemical industry	Process vessel	0.041	0.23
6.	Chemical industry	Piping	0.042	0.23
7.	Petroleum refining	Ancillary process equipment	0.049	0.59
8.	Plastics, synthetics, and resins	Process vessel	0.021	0.29
9.	Plastics, synthetics, and resins	Ancillary process equipment	0.031	0.44

Table 6 shows the incident pattern using variables that described the equipment involved and the type of release. The patterns are presented in the dependent probability format. The probability that ammonia could be released from piping was calculated at 0.16 and was the highest compared to similar piping incidents involving benzene and chlorine where the probability values were calculated at 0.052 and 0.057 respectively.

Table 6: Probability of incident B after incident involving equipment A occurred

No.	A	B	P (A∩B)	P (B | A)	P (A)
1.	Piping	Chlorine	0.011	0.057	0.19
2.	Piping	Benzene	0.01	0.052	0.19
3.	Piping	Ammonia	0.031	0.16	0.19
4.	Ancillary equipment	Benzene	0.013	0.041	0.32
5.	Ancillary equipment	Ammonia	0.027	0.087	0.31
6.	Ancillary equipment	Hydrogen sulfide	0.01	0.033	0.30
7.	Process vessel	Vinyl chloride	0.012	0.068	0.18
8.	Storage above ground	Ammonia	0.017	0.202	0.08

Table 7: Probability of incident B after release of A

No.	A	B	P (A∩B)	P (B | A)	P(A)
1.	Chlorine	Piping	0.011	0.29	0.04
2.	Ammonia	Piping	0.031	0.26	0.12
3.	Benzene	Piping	0.01	0.25	0.04
4.	Benzene	Ancillary process equipment	0.012	0.31	0.04
5.	Ammonia	Ancillary process equipment	0.027	0.22	0.12
6.	Hydrogen sulfide	Ancillary process equipment	0.01	0.46	0.02
7.	Vinyl chloride	Process vessel	0.012	0.45	0.03

These probability values serve as the base condition to estimate the frequency of incident occurrences. In order to specify the probability values, a service factor that takes into account the frequency of use, bulk of containment, and other specific condition of the equipment can be applied. By observing the occurrences of similar attributes from the opposite perspective, as shown in Table 7, the lift values of piping incidents for different chemicals can be calculated.

The lift values for piping and ancillary equipment are shown in Table 8. The lift value quantifies the probability of incident involving chemical Y to occur, given incident involving equipment X has occurred, in relative to the probability of incidents involving equipment X. Using the example from Table 8, it can be observed that the probability of an incident involving ammonia to occur, given a piping incident has occurred is 1.4 times more higher in comparison to the probability of piping incidents.

Table 8: Lift value for piping and ancillary equipment

Chemicals	Lift value	
	Piping	Ancillary equipment
Ammonia	1.4	0.7
Benzene	1.3	1.0
Chlorine	1.5	n/a
Hydrogen sulfide	n/a	1.4

The lift values in Table 8 were compared with the lift values obtained from previous research (Anand, 2005). Both of the current and previous studies used piping as the equipment variable, and the lift values found for three chemicals was compared. As shown in Table 9, there are slight discrepancies between the values obtained from the current research and the previous one. The discrepancies may come from the number of data that was taken into account in the analysis as well as the source of the data used. The previous research used the NRC data which was limited to Harris County data, while the current research was using the HSEES data from 14 states. This suggests that using lift values is limited to facilities or processes in that respective database.

Table 9: Comparison of lift value for piping

Chemicals	Lift value for Piping	
	Present research	(Anand, 2005)
Chlorine	1.5	2
Ammonia	1.4	1.5
Benzene	1.3	1

Dependent probabilities and lift values shown in Tables 5 through 8 can be used as probability inputs in preliminary risk assessments or in general decision making during industrial operations. These values are derived from historical records and provide the likelihood of incidents occurring. Although these values may not be directly applicable for all processes available in the process industry, they can be used to an extent by facilities that participated in the HSEES system. The dependent probabilities can also be used for updating the prior probability of events using observations recorded in the database and the Bayes Theorem. This application however is limited to particular events that the prior probability and the posterior probability were derived from.

4.2.3 Classification and Regression Tree Results

This section is a continuation of section 4.1.2, where the relationship between the predictor variables, such as release quantity and distance of victims in respect to the release source and the dependent variable of incident severity is further studied. In determining the relationship, the first step is creating scatter plots using the variables of interest. The scatter plots exhibit the characteristics of the data and indicate which model is suitable for data fitting. As shown in Figure 21 and Figure 22, the scatter plots indicated that there was no obvious relationship between the variables. The repetitive values of the predictors and severity score exist because they were derived from categorical responses, where the middle point of the categorical response was used to represent the value of each respected variable.

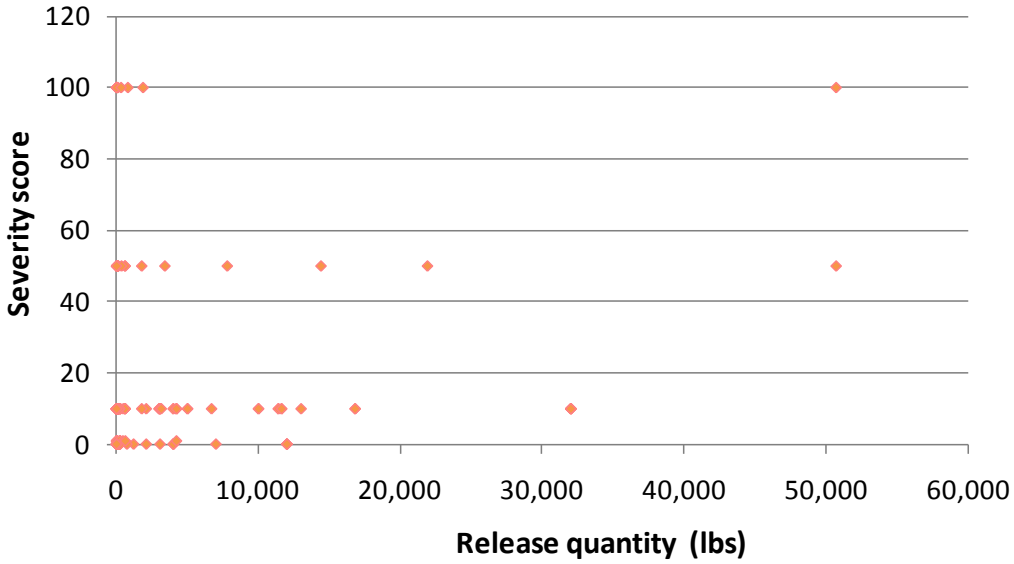


Figure 21: Scatter plot of quantity of release with severity score

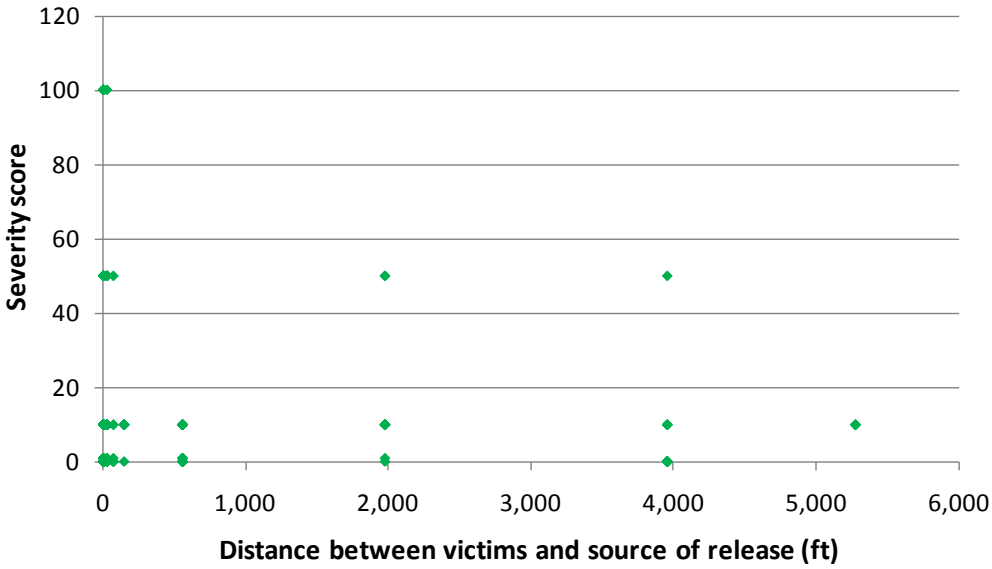


Figure 22: Scatter plot of distance of victims in respect to source of release

Based on the scatter plots, the classification and regression tree (CRT) was selected to fit the incident data due to its ability to perform piecewise regressions. The regression function of the CRT is used to estimate the severity of the incident using predictor variables of release quantity and distance between victims and the source of release. Figure 23 shows the CRT which is used to describe a total of 1,152 process related incidents. The release quantity of 42,000 lbs was the critical release quantity, a point at which the tree started splitting into branches. Incidents having releases beyond this amount were estimated to have an average severity score of 57 (node 20), and in this case the distance between the victims and the source of release did not affect the average severity score.

On the other hand, the severity of incidents with releases less than 42,000 lbs was affected by the distance of the victims to the source of release. If the distance of the victims to the source of release was less than 113 ft, then the average severity score was estimated at 14, otherwise it was estimated at 7. Following the split arrows, if the release quantity is less than 1.5 lbs then go left to node 1, which represents the incidents with estimated average severity score of 7. For incidents with a release quantity of more than 1.5 lbs, go right to node where the average severity score was estimated at 15, and so on. Table 10 shows all of the nodes generated by the CRT and the values of the predictors and dependent variable.

From 13 of the 19 terminal nodes, it can be observed that there is a proportional relationship between the release quantity and the incidents' average severity score, and there is an inverse proportional relationship between the distances of the victims from the source of the release and the incidents' average severity score. These relationships are reasonable due to the fact that a higher quantity of chemical releases in the same space means a higher concentration and larger distance between victims and the source of the release and higher dilution effects, thus lower concentrations.

Nodes 1-2 and 14-16 show that the increase in release quantity, given constant distance between the victims and the source of release, resulted in a higher average severity score. From nodes 3-4 and 10-11, it can be observed that the larger the distance between victims and the source of release, given the same release quantity, resulted in lower average severity score. Nodes 6-7 and nodes 8-9 are incidents with increasing release quantity and a constant distance between victims and the source of release. For each pairs of nodes 6-7 and 8-9, there is a proportional relationship between the release quantity and incidents' average severity score. However, if we look at nodes 6-9 as a continuous scheme, the proportional relationship does not apply anymore. This indicates that there may be other predictor variables which affect the incidents' severity, particularly for the incidents belonging to these nodes. Therefore, identifying and integrating the other significant predictors can improve the CRT results.

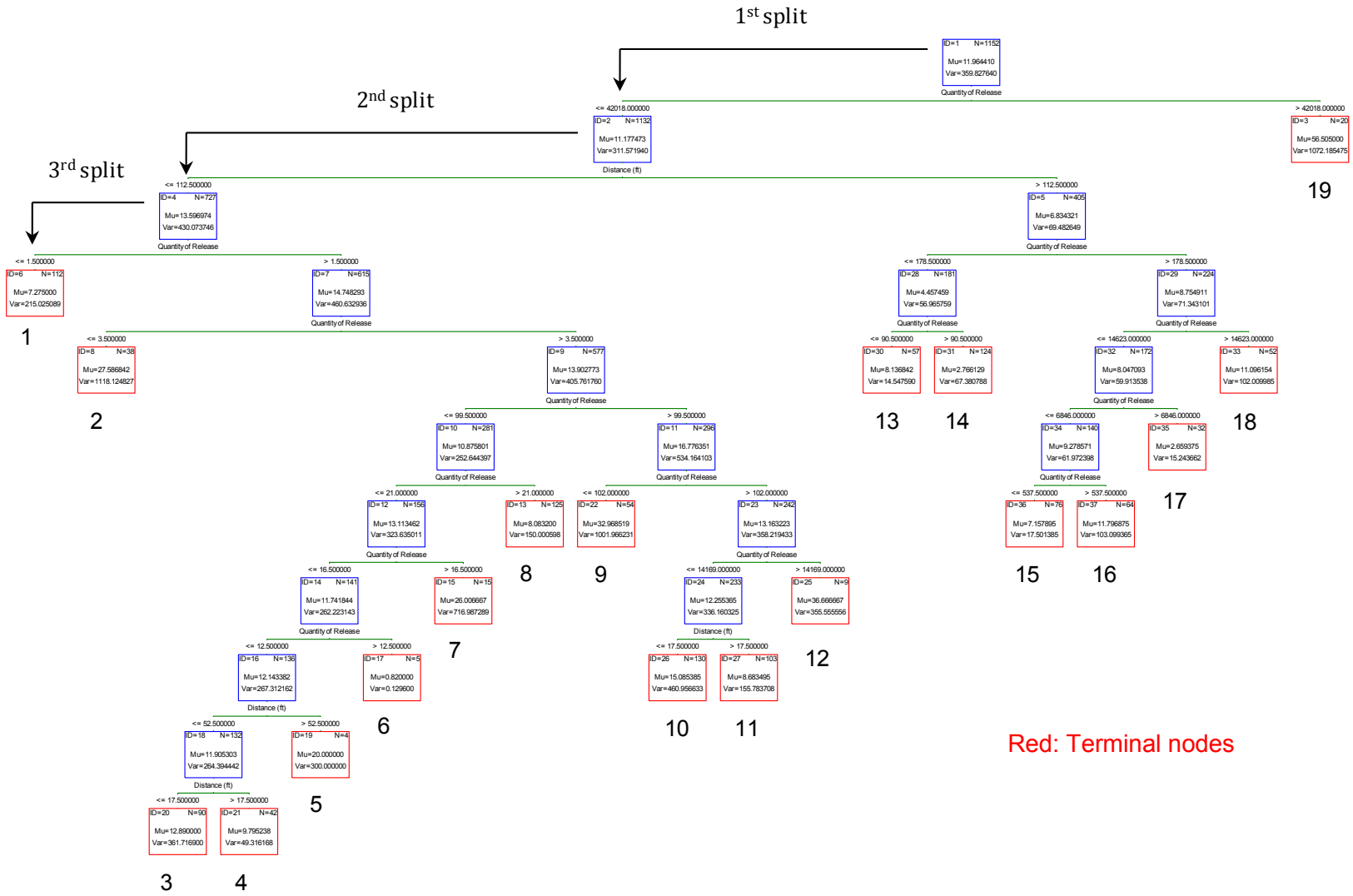


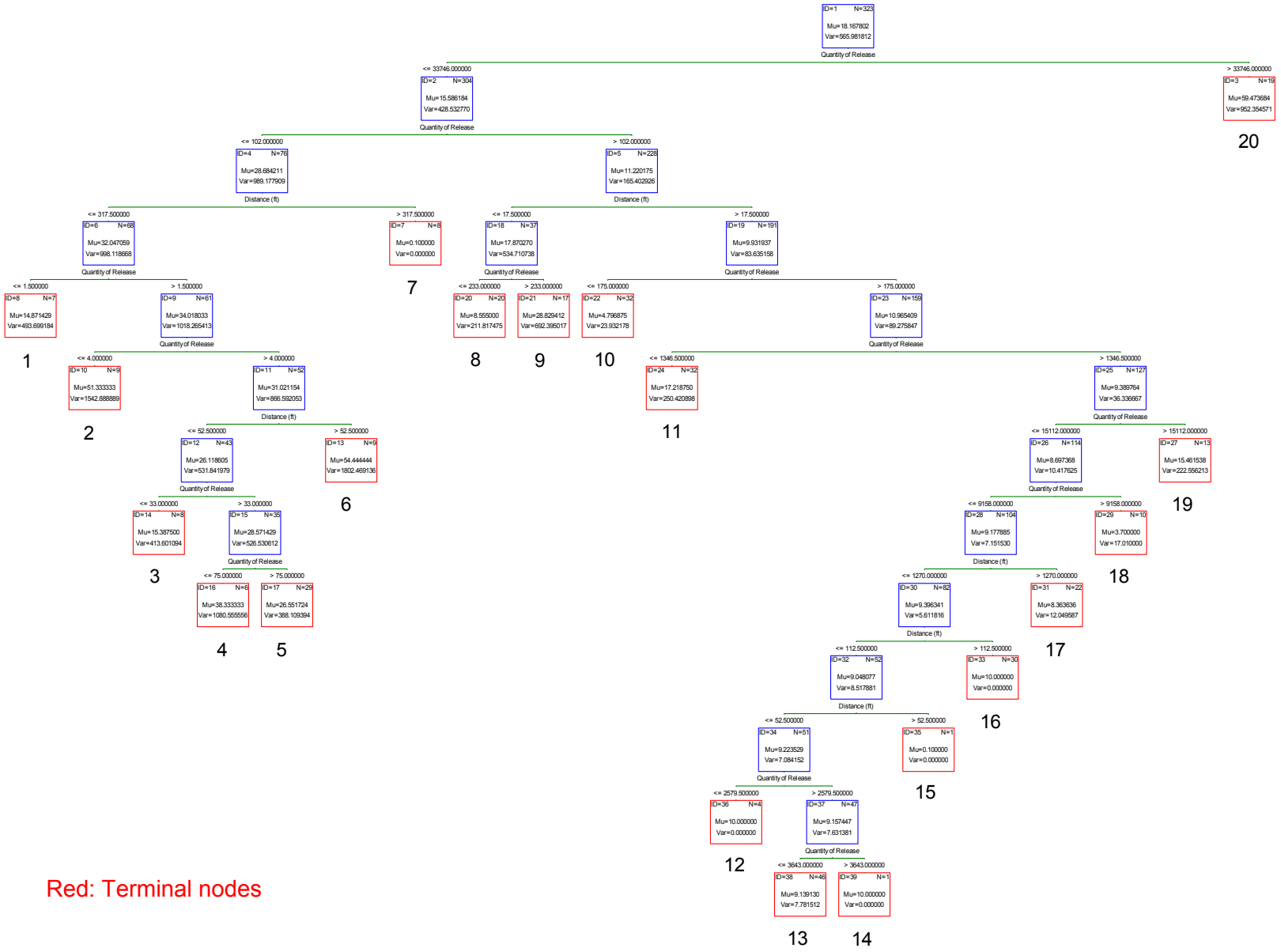
Figure 23: CRT for chemical process related incidents

Table 10: CRT summary for chemical process related incidents

First Splits	Node	Number of incidents (N)	Release quantity (lbs)	Distance (ft)	Average severity score (Mu)
Left	1	112	Q ≤ 1.5	D ≤ 112.5	7
	2	38	Q ≤ 3.5	D ≤ 112.5	28
	3	90	3.5 < Q ≤ 12.5	D ≤ 17.5	13
	4	42	3.5 < Q ≤ 12.5	17.5 < D ≤ 52.5	10
	5	4	3.5 < Q ≤ 12.5	52.5 < D ≤ 112.5	20
	6	5	12.5 < Q ≤ 16.5	D ≤ 112.5	1
	7	15	16.5 < Q ≤ 21	D ≤ 112.5	26
	8	125	21 < Q ≤ 99.5	D ≤ 112.5	8
	9	54	99.5 < Q ≤ 102	D ≤ 112.5	33
	10	130	102 < Q ≤ 14,169	D ≤ 17.5	15
	11	103	102 < Q ≤ 14,169	17.5 < D ≤ 112.5	10
	12	9	14,169 < Q ≤ 42,018	D ≤ 112.5	37
	13	57	Q ≤ 90.5	D > 112.5	8
	14	124	90.5 < Q ≤ 178.5	D > 112.5	3
	15	76	178.5 < Q ≤ 537.5	D > 112.5	7
	16	64	537.5 < Q ≤ 6,846	D > 112.5	12
	17	32	6,846 < Q ≤ 14,623	D > 112.5	3
	18	52	14,623 < Q ≤ 42,018	D > 112.5	11
Right	19	20	Q ≥ 42,018	-	57

Then, the CRT was performed to more specific data that included 323 incidents where fires and explosions had occurred. Figure 24 shows the critical predictor for this tree was the release quantity of 33,746 lbs. For a release quantity above this amount, the average severity score of the incidents was estimated at 60, otherwise it was estimated at 16. Following the same principle as the previous CRT, this tree produced 20 terminal nodes which are summarized in Table 11.

Nodes 1-2, 3-4, 8-9, 10-11 and 18-19 demonstrate the proportional relationship between the release quantity and the average severity score, given a constant distance between victims and the source of release. Nodes 12 and 13 have similar predictor variable values and therefore, similar values for the average severity score. Nodes 16-17 show that within the same release quantity range, the severity score decreased as the distance between victims and the source of the release increased.



Red: Terminal nodes

Figure 24: CRT for fire and explosion incidents

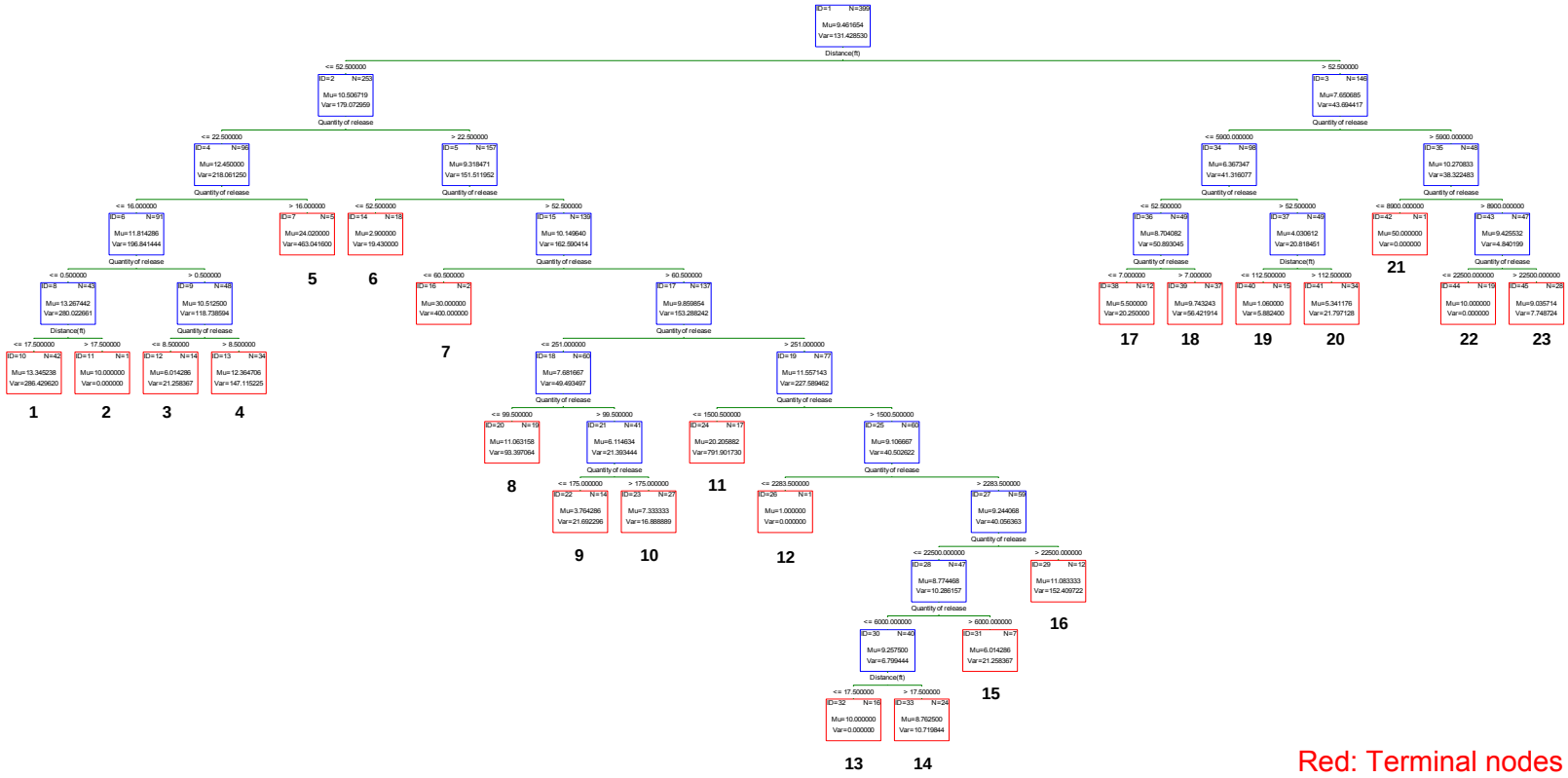
Table 11: CRT summary for fire and explosion incidents

First Splits	Node	Number of incidents (N)	Release quantity (lbs)	Distance (ft)	Average severity score (Mu)
Left	1	7	Q ≤ 1.5	D ≤ 317.5	15
	2	9	1.5 < Q ≤ 4	D ≤ 317.5	51
	3	8	4 < Q ≤ 33	D ≤ 52.5	15
	4	6	33 < Q ≤ 75	D ≤ 52.5	38
	5	29	75 < Q ≤ 102	D ≤ 52.5	27
	6	9	4 < Q ≤ 102	52.5 < D ≤ 317.5	54
	7	8	-	D ≥ 317.5	0.1
	8	20	102 < Q ≤ 233	D ≤ 17.5	9
	9	17	233 < Q ≤ 33,746	D ≤ 17.5	29
	10	32	Q ≤ 175	D > 17.5	5
	11	32	175 < Q ≤ 1,346.5	D > 17.5	17
	12	4	1, 346.5 < Q ≤ 2,579.5	D ≤ 52.5	10
	13	46	1,346.5 < Q ≤ 3,643	17.5 < D ≤ 52.5	9
	14	1	3,643 < Q ≤ 9,158	17.5 < D ≤ 52.5	10
	15	1	1,346.5 < Q ≤ 9,158	52.5 < D ≤ 112.5	0.1
	16	30	1,346.5 < Q ≤ 9,158	112.5 < D ≤ 1,270	10
	17	22	1,346.5 < Q ≤ 9,158	D > 1,270	8
	18	10	9,158 < Q ≤ 15,112	D > 17.5	4
	19	13	15,112 < Q ≤ 33,746	D > 17.5	15
Right	20	19	Q ≥ 33,746	-	60

Finally, the CRT was performed using 339 cases of ammonia incidents. Figure 25 shows the CRT for the ammonia data, where the first partitioning of the data occurred at a distance of 53 ft. For incidents where the distance between the victims and the source of release was less than 53 ft, go left, otherwise go right. This CRT produced 23 terminal nodes which represent the incident outcomes as shown in Table 12. Each node has incidents with frequency (N), and the analysis considered only nodes with N ≥ 5. In general, it can be observed that the release quantity proportionally increases with the increase of severity score. Node 3-5, 6 and 8, 9-11, 15-16, and 17-18 show that the increase in release quantity resulted in the increase in the average of the severity score, given the distance of victims to the release was constant. The previous result justifies the positive relationship between release quantity and severity of incidents, which was assumed at the beginning.

Node 13 and 14 show that within the same range of release quantity of $2{,}283.5 < Q \leq 6{,}000$ lbs, the average severity score of the incidents was similar in the value of 10 and 9, respectively. The incident severity of node 14 was lower compared to that of node 13, because the distance between victims and source of release in node 14, $17.5 < D \leq 52.5$ ft, was larger compared to the distance of node 13, $D \leq 17.5$ ft. This justifies the inverse proportional relationship between the distance between victims and the source of release and the incidents' severity that was presumed. Other nodes did not show obvious relationships.

The CRT was successfully used to describe the relationship between the predictor variables, release quantity and distance of victims from source of release, as well as the dependent variable, incident severity, in a semi-quantitative manner. There are cases, where the release quantity and distance between the victims and source of release alone were not enough to estimate the severity of the incidents. This indicates that there are other factors that need to be taken into account for as predictor variables in the analysis in order to estimate the severity of the incidents accurately.



Red: Terminal nodes

Figure 25: CRT for ammonia incidents

Table 12: CRT summary for ammonia incidents

First Splits	Node	Number of incidents (N)	Release quantity (lbs)	Distance (ft)	Average severity score (Mu)
Left hand	1	42	Q ≤ 0.5	D ≤ 17.5	13
	2	1	Q ≤ 0.5	17.5 < D ≤ 52.5	10
	3	14	0.5 ≤ Q ≤ 8.5	D ≤ 52.5	6
	4	34	8.5 < Q ≤ 16	D ≤ 52.5	12
	5	5	16 < Q ≤ 22.5	D ≤ 52.5	24
	6	18	22.5 < Q ≤ 52.5	D ≤ 52.5	3
	7	2	52.5 < Q ≤ 60.5	D ≤ 52.5	30
	8	19	60.5 < Q ≤ 99.5	D ≤ 52.5	11
	9	14	99.5 < Q ≤ 175	D ≤ 52.5	4
	10	27	175 < Q ≤ 251	D ≤ 52.5	7
	11	17	251 < Q ≤ 1,500	D ≤ 52.5	20
	12	1	1,500 < Q ≤ 2,283.5	D ≤ 52.5	1
	13	16	2,283.5 < Q ≤ 6,000	D ≤ 17.5	10
	14	24	2,283.5 < Q ≤ 6,000	17.5 < D ≤ 52.5	9
	15	7	6,000 < Q ≤ 22,500	D ≤ 52.5	6
	16	12	Q ≥ 22,500	D ≤ 52.5	11
Right hand	17	49	Q ≤ 7	D ≥ 52.5	6
	18	37	7 < Q ≤ 52.5	D ≥ 52.5	10
	19	15	52.5 < Q ≤ 5,900	52.5 < D ≤ 112.5	1
	20	34	52.5 < Q ≤ 5,900	D ≥ 112.5	5
	21	1	5,900 < Q ≤ 8,900	D ≥ 52.5	50
	22	19	8,900 < Q ≤ 22,500	D ≥ 52.5	10
	23	28	Q ≥ 22,500	D ≥ 52.5	9

4.3 Text Mining Results

Text mining was performed to analyze the HSEES comments variable. The data mining tools used to analyze the text mining results were feature selection and cluster analysis.

4.3.1 Feature Selection Results: Predictor Variables

The number of text documents used in the text mining process was 3,316 and the number of words selected was 181. The SVD process was performed using the inverse document frequency to reduce the initial 181 x 3,316 word-document matrix. The same

process resulted in 27 components of singular value. The scree plot visualizes the results of the SVD, as shown in Figure 26. The elbow of the plot is located after component 4. Thus, components 1 through 4 represent 20 % of the variance among the data.

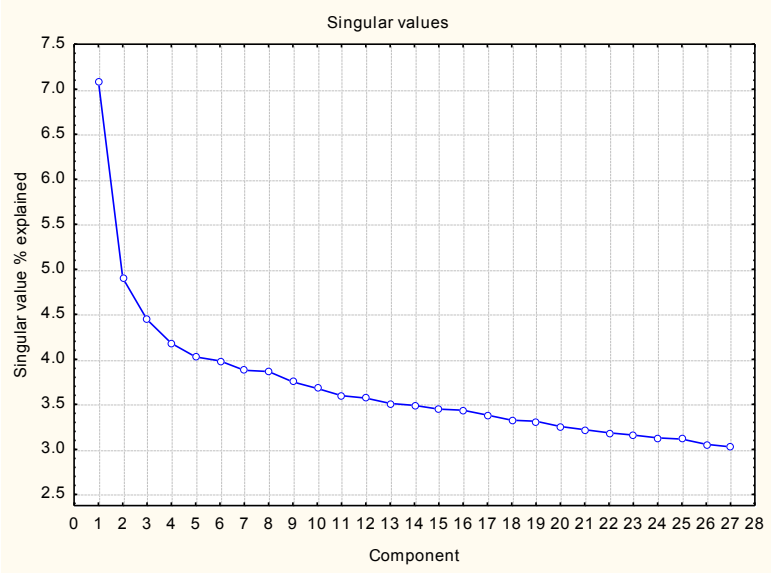


Figure 26: SVD scree plot

Then the words were mapped in the reduced dimension as shown in Figures 27 through 29. The groups of words that are distinguished from the large majority of words group are the words of importance and the proximity of the words on the plot represents their close relationships (Statsoft, 2008).

From Figure 27, it can be observed that ammonia, valve, tank, leak, flare and line were the important words. Through the closeness of the words shown in this figure, it can be observed that whenever the word ammonia is mentioned, the report would also include the words leak, valve and tank. This implied that ammonia incidents reported to HSEES generally can be linked to a leaking valve or tank.

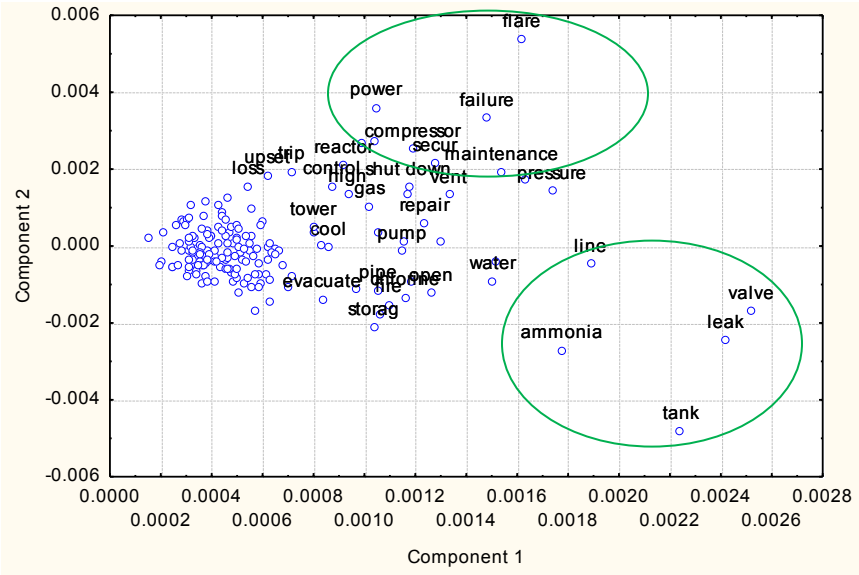


Figure 27: Scatter plot of component 1 and 2

Flare, failure, power, compressor, reactor, maintenance and pressure were the other important words that are also imminent to one another. The grouping of these words is reasonable due to the fact that flaring is usually conducted to handle overpressure or equipment (compressor, reactor) failure.

Figure 28 illustrates the scatter plot between components 3 and 4, where it shows that valve, pressure, leak, open, relief valve, cooling, heat exchanger, tube, leak, water and tower as the important words. The proximity of these words implies that there were significant amounts of incident reports of heat exchangers leaking in the tube side where water had leaked. The plot also revealed that there had been many reports of incidents where overpressure had occurred that led to the opening of a relief valve.

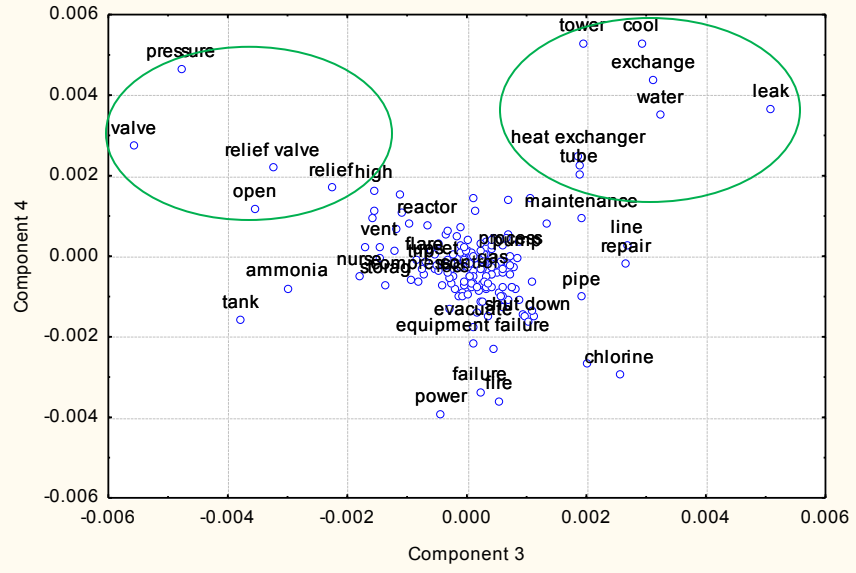


Figure 28: Scatter plot component 3 and 4

The numeric indices obtained from the extracted word frequencies and SVD process now can be further analyzed using various data mining tools such as the feature selection tool. The feature selection tool is used to identify the best predictors for the words of interest (dependent variable) in predictive modeling.

Figures 29 through 38 show the 10 best predictors for selected dependent variables. The words trip, restart, shutdown, flare, refrigerant, control, and failure were the best predictors for the variable ‘compressor’ as shown in Figure 29. Referring back to the incident description, it was evident that most compressor related incidents were caused by a trip or was shut down due to equipment failure or power outage. Then, the compressor line was isolated and sent to flare system. In many of the cases, the compressor was part of the refrigeration system, and this explained the presence of the word refrigerant.

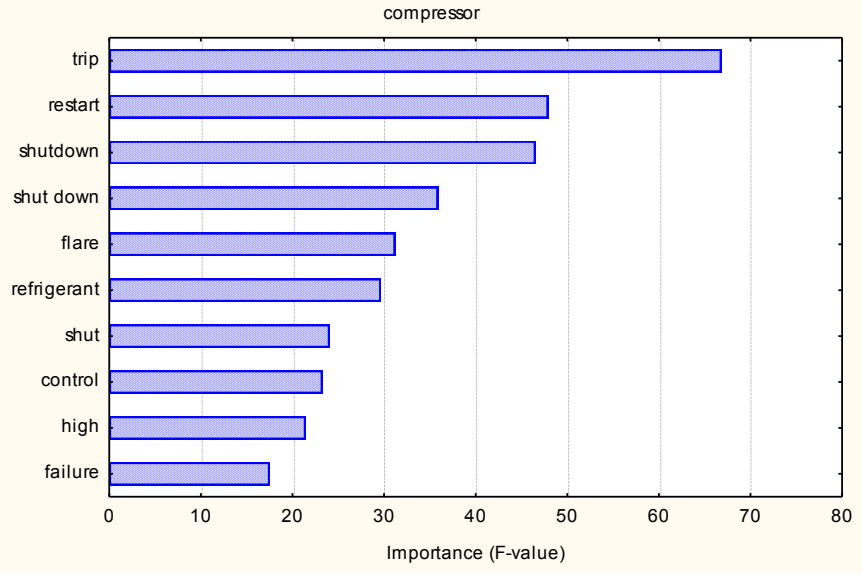


Figure 29: Importance plot using compressor as dependent variable

Figure 30 shows that seal, failure, isolate, discharge, maintenance, replace, stop, block, and water are the best descriptors for pump. In many of the reported incidents, seal failure was the prominent cause of pump failures. The word discharge referred to the failure (leaking or blocked) of a discharge line or chemical discharge. From the incident descriptions, it can be observed that these incidents often occurred during maintenance work. This can be used as an indication that the work procedure during maintenance needs to be evaluated.

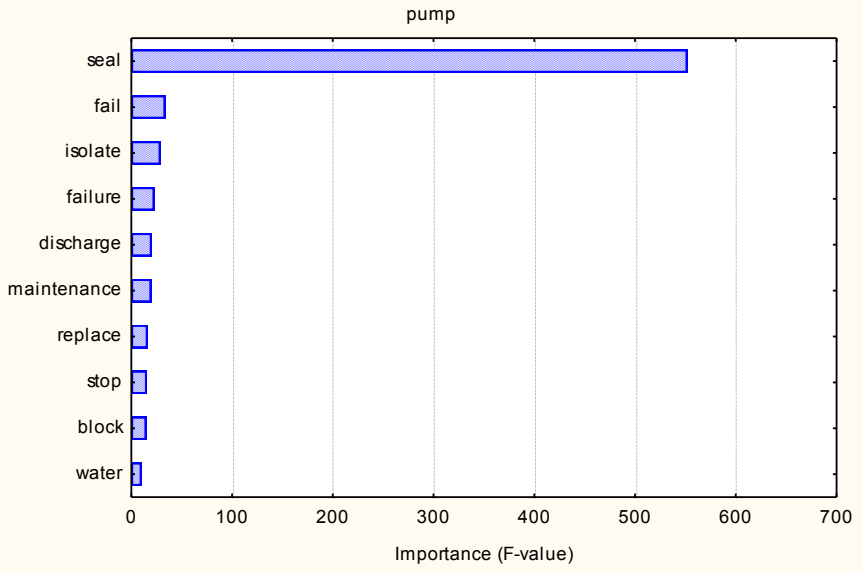


Figure 30: Importance plot using pump as dependent variable

As shown in Figure 31, the variable reactor was strongly related to the words temperature, high, shutdown, flare, vent, trip, active, upset, flow, and tank. Based on the incident descriptions, the reactor-related incidents are often due to high temperatures in the reactor that leads to shutting down the system, venting the chemical, or isolating the reactor system. It can also be observed from the plot that there were several instances where the word ‘reactor’ was present, and words such as trip, process upset, overflow or tank were also present.

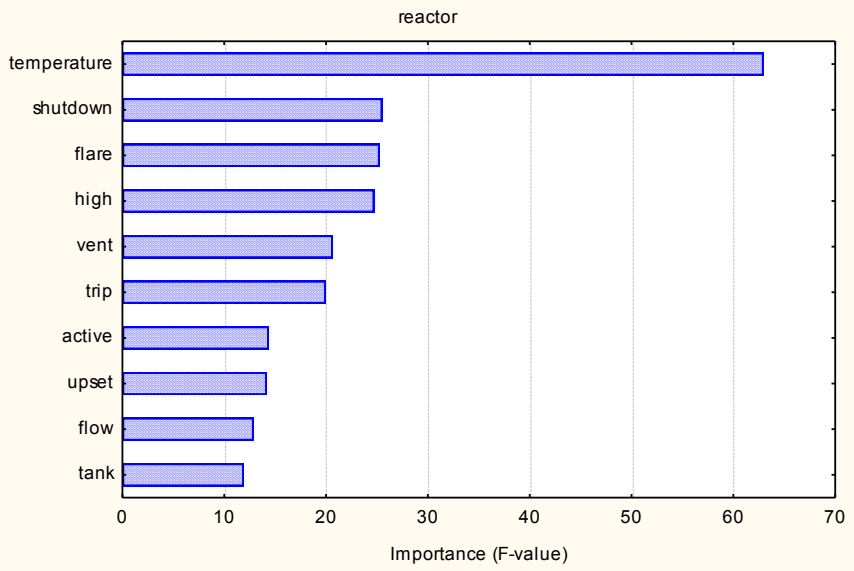


Figure 31: Importance plot using reactor as dependent variable

Figure 32 shows the best predictors for the word ‘boiler’. The details of HSEES reports confirmed that the boiler-related incidents were often initiated by trips which then caused a steam pressure swing and fuel gas to be vented. The faulty switches on boiler levels were reported several times as the cause of the trip.

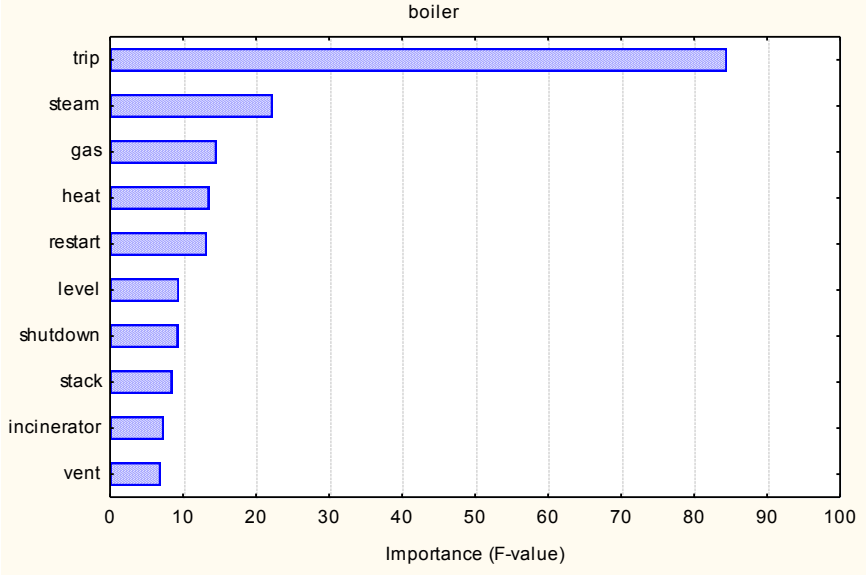


Figure 32: Importance plot using boiler as dependent variable

Figure 33 shows that the words trip, vent, stack, loss, feed, malfunction, hydrogen, leak, active and gas were the best descriptors when incinerator is the dependent variable. The reported mode of incinerator failures was trip, which resulted in the incidental release of stack gas to the atmosphere. Incinerator malfunctions were described as another mode of incinerator failure.

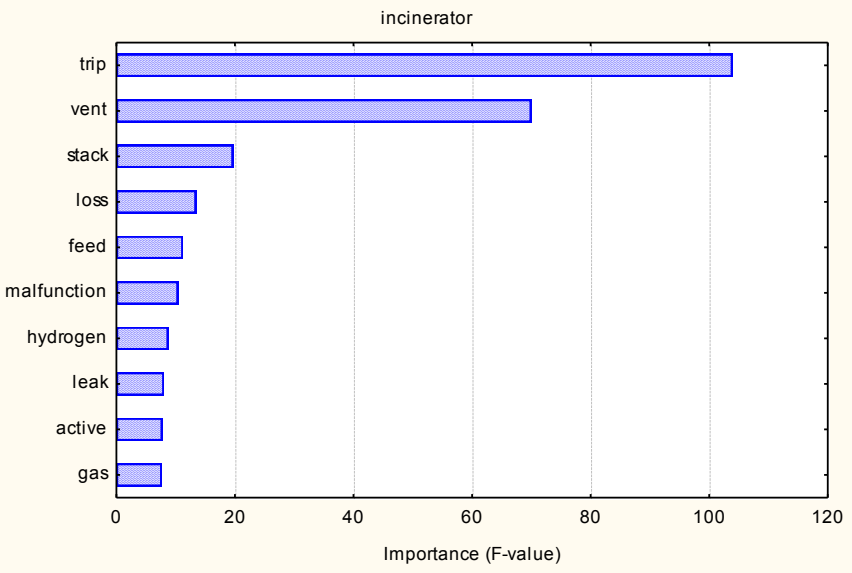


Figure 33: Importance plot using incinerator as dependent variable

As shown in Figure 34, the word pipe had descriptor words such as hole, leak, broke, crack, fit, evacuation, victim, soil, chlorine, and flange. The incident descriptions revealed that the pipe-related incidents were usually due to a leak through a hole, crack or break in the pipeline. Some of these pipe failure modes resulted from corrosion problems and occurred during maintenance activities. Several incidents also reported the release of chlorinated water from the pipe, evacuation as a consequence of the chemical release, and maintenance as the company response to the leak.

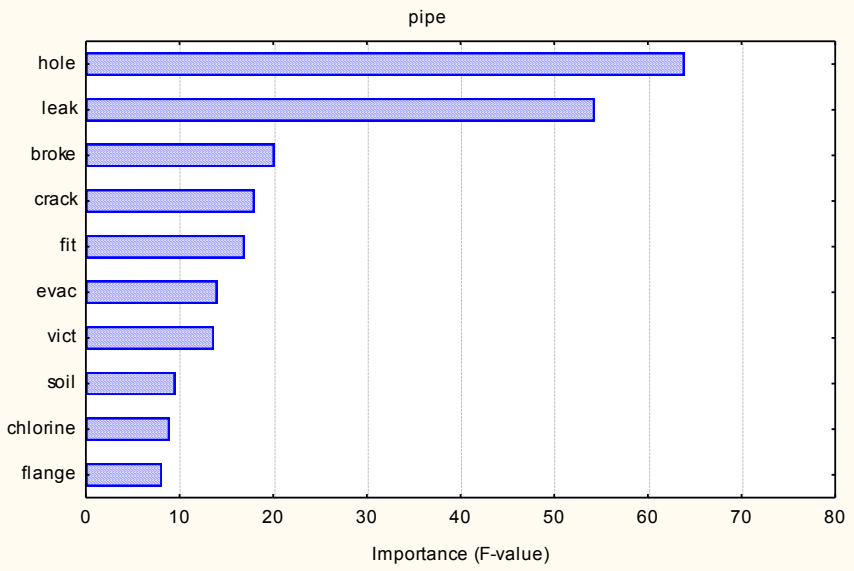


Figure 34: Importance plot using pipe as dependent variable

Figure 35 shows that the word gasket can be predicted using words such as flange, blew, fail, rupture, transfer, vapor, line, reactor, and detect. The incident descriptions showed that the incidents mentioning gasket were highly related to flanges on pipelines and many times the gasket on the flanges failed (ruptured) and blew away. Other significant gasket incident reports involved failure of gaskets in the manways of the reactors.

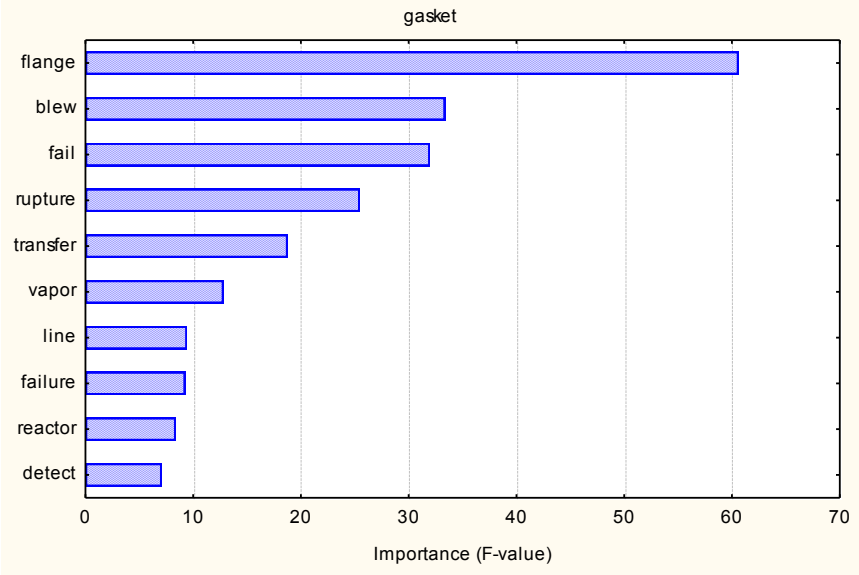


Figure 35: Importance plot using gasket as dependent variable

The word fit pointed out in Figure 36 refers to the fitting of pipes. As can be observed, the words crack, repair, pipe, tube, leak, ammonia, hose, block, broke, pump were among the best predictors for incidents related to fitting. The common failure modes for these incidents were crack and loose fittings which led to chemical leak. Ammonia in particular was frequently reported in incidents related to fittings leak.

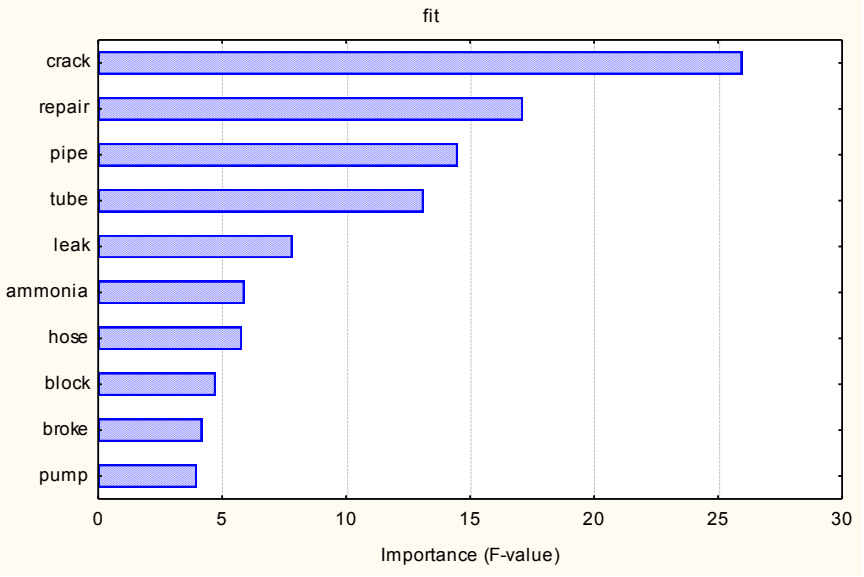


Figure 36: Importance plot using fit or fitting as dependent variable

As shown in Figure 37, the word hose was strongly related to the word nurse, which referred to nurse tanks. The nurse tanks were often reported to store ammonia. The word transfer was also common due to the transfer of fluids using a hose, *e.g.*, from a tank to a truck.

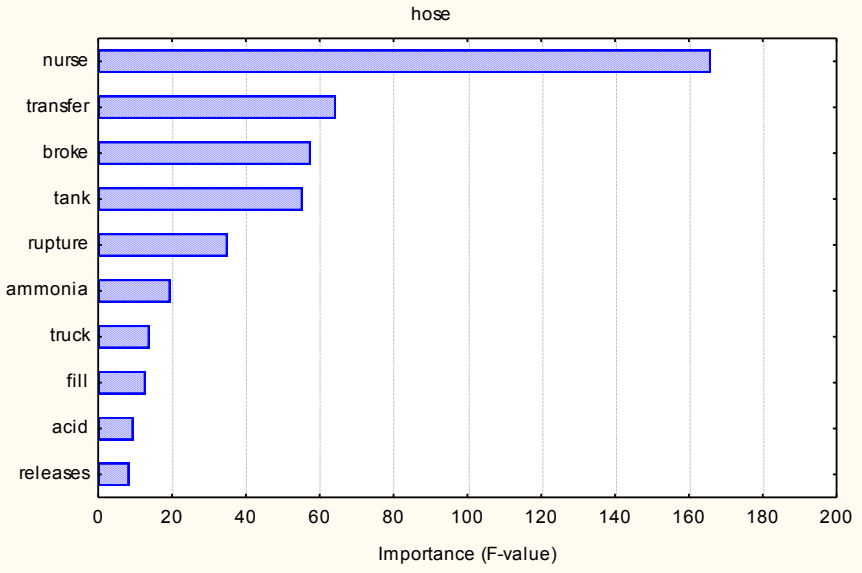


Figure 37: Importance plot using hose as dependent variable

Figure 38 shows the importance plot using ammonia as the dependent variable. Whenever ammonia was reported, the incident description would also include the word smell, referring to the pungent odor of ammonia. The ammonia-related incidents reported to HSEES commonly related to refrigeration systems, which explained the high F-values of the word ‘refrigerant’. From the plot, it can also be assumed that a significant number of ammonia incidents resulted in evacuation of the nearby population.

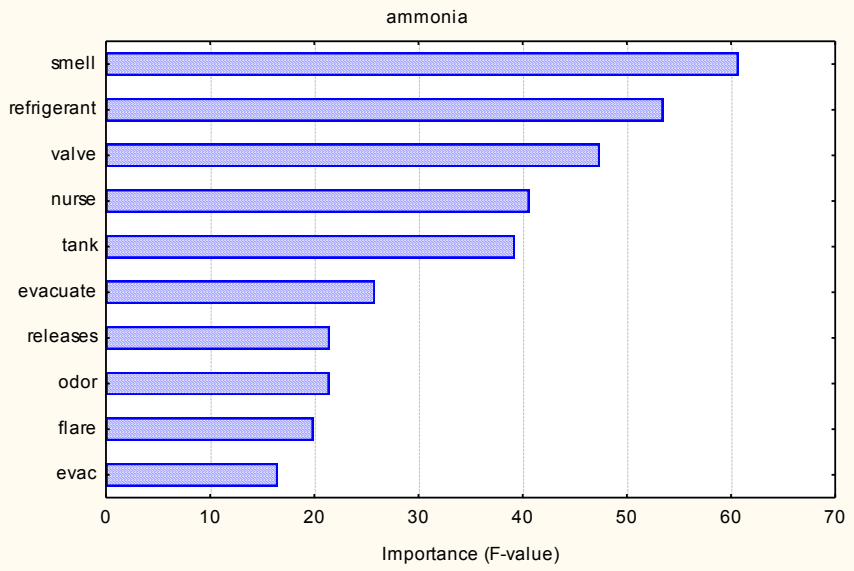


Figure 38: Importance plot using ammonia as dependent variable

The previous figures describe the usefulness of importance plots where one can get the significant predictor variables of a particular word or variable of interest. Furthermore, these predictors can be used in predictive modeling for the dependent variable. The predictor words were identified based on their relative frequencies and importance in the incident documents. Using the predictor words, scenarios of incidents with different types of failure modes can be developed. Incident scenarios can be used for many purposes, such as inputs for process hazard analysis (PHA), quantitative risk assessments (QRA), etc. In order to have a comprehensive view of the incident scenarios, they have to be developed and structured to emulate the actual events. The current variables in HSEES may not provide all the necessary elements to build detailed incident scenarios.

Several recommendations can be proposed to improve the quality of the data and information contained in the database. This will be further discussed in section 4.4. Overall, this section demonstrated the application of text mining to the HSEES text variable. The text mining results captured the information contained in the incident database. The findings from text mining analysis not only were aligned with the findings from the statistical analysis but also provided more details, which added more value to the analysis.

4.3.2 Cluster Analysis with Text Inputs Results

Cluster analysis was performed using words and document components obtained from the SVD process. The numbers of clusters were assigned so that the members of each cluster provided a meaningful group of words that can be used as elements to describe the incidents. The text data was segmented into 3 clusters and the profiles of the 3 clusters were evaluated for meaningful groupings. If these clusters did not provide meanings, the process was iterated until meaningful groups were found. The final attempt was grouping the data into 8 clusters. Finally, after observing and evaluating the profiles of the clustering, the incident data with 6 clusters provided the most

meaningful clusters. The characterizing words of each cluster are shown in Table 13 and the succinct descriptions of the clusters are shown in Table 14.

Table 13: Cluster analysis using text inputs

Cluster	%	Characterizing words
1	4.4	Cool, heat exchanger, isolate, leak, tower, tube, water
2	2.8	Ammonia, open, pressure, tank, valve
3	3.9	Chlorine, evacuate, facility, fire, pipe, spill, storage
4	12.1	Air, compressor, control, equipment failure, flare, gas, high, line, maintenance, occur, power, process, pump, reactor, repair, secure, shutdown, vent
5	20.4	Anhydrous, block, change, close, column, condense, drum, excess, flow, high, indicate, isolate, level, left, liquid, malfunction, monitor, nurse, overpressure, product, refrigerant, releases, relief valve, relieve, rupture, safety valve, steam, stop, temperature, trip, upset, vapor, vessel
6	56.4	Acid, active, bag, blew, boiler, broke, broken, burner, chemical, clean, condition, contractor, crack, cylinder, detect, discharge, drain, electric, emergency, employees, enter, error, evacuation, faulty, feed, fill, fire department, fit, flange, freon, hole, hose, hydrogen, incinerator, injury, inside, intent, investigate, laboratory, load, located, loss, material, mercury, odor, old, outage, outside, oxide, pipeline, place, plan, plug, possible, problem, receive, reduction, remove, replace, resident, restart, roof, room, scene, scrubber, seal, service, severe, sewer, site, smell, soil, stack, start up, storm, taken, thermal, transfer, treatment, truck, victim, waste

Table 14: Clusters descriptions

No.	Cluster Description
1	Heat exchanger related incidents: leak as the common mode of failure, particularly in the tube side; water as prominent leaking agent; isolation of the equipment was the facility response.
2	Ammonia incidents included scenarios where pressure built up due to improper filling of the tanks, full or partial opening of the valve caused chemical release during operation
3	Chlorine incidents where it spilled from pipes or was released from storage tank causing facility-wide evacuations.
4	Incidents due to equipment failure: compressor incidents due to malfunctioning control valve or power trip, high pressure or temperature of the reactor resulted in releasing gas (chemical) through flare, reactor upset led to shutdown and the chemical released through flare, pump failure, maintenance as the facility response, etc
5	Condenser incidents where a leak was detected as cause of loss of containment, incidents related to condensation: cold temperature condensed moisture in airline, excessively high temperatures, leaks which released refrigerant, high level in the compressor drum tripping the equipment caused by malfunctioning instrumentation. Incident involving blocked flow, high level that led to overpressure, activation of relief valve. Incident occurred while performing change in the equipment: unexpected residual released to the environment. Nurse tank incidents: a temperature change influenced pressurization of the tank causing valve to open. Tank or line rupture due to overpressure, puncture or thermal shock or safety disk rupture due to overpressure. Steam was used to dilute emission at the flare, air monitoring due to chemical release, anhydrous ammonia leaked due to valve failure.
6	Incidents related to thermal oxidizer, boiler, incinerator, scrubber, chlorine reduction burner, plug, pipeline, drain, transfer hose or line and flange failure. The equipment blew, broke, cracked, or leaked through holes and resulted in loss of containment. The failure resulted from electrical problems such as power outage or damaged electrical box. Chemical discharged in these types of incidents were hydrogen, freon, mercury, oxide, and acids. Most of the equipment involved in the incidents was restarted to continue operation. Most of the releases occurred during planned activities such as equipment start up, cleaning and maintenance. The incidents in this cluster also cover releases that went to the storm sewer system.

Cluster 1 contains group of words which implies incidents involving heat exchangers, as shown in Table 14. Referring to the incident descriptions, generally the heat exchangers experienced failure due to a leak particularly in the tube side of the equipment. Water was reported as the prominent leaking agent in this incident group and facilities performed isolation of the equipment as a response to the incidents.

Cluster 2 contains group of words that describe incidents related to ammonia releases. The possible scenarios that can be derived from this importance plot are ammonia releases from tanks due to overpressure or ammonia release from an accidentally opened valve. More detailed scenarios of ammonia incidents included pressure built up due to improper filling of the tanks and full or partial opening of the valve causing release during operation.

Cluster 3 is characterized mostly by chlorine incidents where it spilled from storage tanks or leaked from pipelines. Due to its toxicity, many of the incidents resulted in a facility wide evacuation. The word fire in cluster 3 was considered peculiar because chlorine does not pose a fire hazard. Referring back to the incident descriptions, the word fire actually referred to phrases such as fire department or fire fighters. This indicates that the synonyms and phrases list needs to be edited to capture this phrase.

Cluster 4 covers incidents which were primarily related to compressor, reactor, and pump failures. Control valve malfunctions and power trips were reported as common precursors in compressors incidents. Instrumentation malfunctions resulted in high levels of liquid in the compressor drum that subsequently tripped the compressor. The reactor incidents usually resulted from high pressures or temperatures, where the gases (chemicals) were released through flares. The pump failure usually linked to seal failure, which allowed the chemical to leak to the environment. Maintenance implied maintenance work activities as the facilities' response to the incidents.

Cluster 5 consists of incidents related to condensers and vessels containing liquid in particular nurse tanks. For all this equipment, leaks were detected as the major cause of loss of containment events. Reported condenser incidents included loss of cooling water in the condenser overheads, causing excessive temperature and pressure

increases, and leaks in the flanges or tube sides, releasing refrigerants to the environment.

Nurse tank incident descriptions stated that drastic changes in temperature influenced the pressurization of the tank and triggered the activation of the relief valve. Other types of vessels and lines experienced ruptures due to overpressure, puncture, or thermal shock. Overpressure was a consequence of either blocked flow or high liquid levels in the vessels. This cluster also includes incidents where cold temperature was reported as the one of the contributing factors to failure due to line freezing and the condensation of air moisture in the airline.

Cluster 5 also covers incidents that occurred while performing changes to the equipment. Releases occurred when residuals of the chemicals were not completely discharged or purged from the system. In order to reduce the concentration of the chemical, steam was used as a diluting agent at the flare and the air was monitored.

Cluster 6 contains various incidents that do not belong in other clusters; hence it is comprised of incidents with different characteristics. Included in this cluster are incidents involving process units and parts, such as thermal oxidizers, boilers, incinerators, scrubbers, chlorine reduction burners, plugs, pipelines, drains, transfer hoses or lines and flanges. The equipment blew, broke, cracked, or leaked causing a loss of containment. Electrical problems such as power outages and electrical breaker failures were reported as precursors to the equipment failures.

Chemicals that were discharged in significant numbers were hydrogen, freon, mercury, oxide, and acids. Most of the chemicals discharge occurred during a planned activity such as equipment start up, cleaning and maintenance. The incidents in this cluster also cover chemical discharges that ran off into the storm sewer system.

In comparison with the cluster analysis that used structured data from section 4.2.1, the cluster analysis that used text inputs gave far more meaningful results because it contained words that can be structured to form incident elements and descriptions. The cluster that used text inputs showed a number of typical scenarios which also validates the cluster results. The present analysis coincides with the typical scenarios occurring in

the industry, and it can also be developed to identify rare incident scenarios where information is still limited.

4.4 Recommendation for Chemical Incident Databases

The HSEES chemical incident database is an example of the benefits and versatility of a chemical incident database. The previous HSEES database contained numerous variables that were applicable to process safety analysis. However, some analysis was limited by the information available from HSEES. For example, the equipment, chemical and industry type patterns generated using the association rules are still relatively general. There are two improvements that can be made, 1) to include more informative variables such as variables that describe the process units pertinent to the incidents and 2) to organize the variables and their attributes into taxonomy suitable to process safety.

The process unit information should characterize the type of process held in the vicinity where the loss of containment had occurred. For example, an incident occurred in a fertilizer plant where process vessels and ammonia were reported as equipment involved in the chemical release. Instead of reporting the previous variables, which were the existing variables used by HSEES, it would be more informative to report that the incident had occurred in an ammonia refrigeration unit of the fertilizer plant.

Chemical incident database users, especially those with process safety backgrounds, can create relevant associations with the process conditions, standard supporting equipment used in the processes, the relative locations of the process units in the facilities (*e.g.*, downstream or upstream of a reactor), utilities complementing the process units. This type of information will create more possibilities for utilization of the incident reports. The variables should also be organized and structured using a taxonomy tailored to process safety so that the users have a common understanding of the incidents. The Mary Kay O'Connor Process Safety Center is currently working on developing a process safety taxonomy which can be applied to incident database systems.

The prominent incident structure in the industry is the Barrier model. This model structures the incident by initiating events, protection barriers and subsequent events or consequence. The possible initiating events turn into process deviation if the control systems of the prevention barrier fail. Subsequently, if process deviation continues, loss of control leads to incidents (Markowski, 2006). The application of incident taxonomy and structured incident scenarios can improve the database significantly and diversify the potential use of chemical incident databases.

Gathering information on the time duration of the incidents which then can be translated into the length of exposure to the victims would be very useful. As mentioned, the adverse effects of chemical exposure to human health should be analyzed based on the combination effect of concentration and length of exposure. The text entries in HSEES can be designed so that the reporters can provide brief descriptions of the incidents in a structured manner. This can be accomplished by creating list of questions for the HSEES data collection form that ensures desired information about the incidents are covered in the text descriptions.

Furthermore, the analysis of data can be only as good as the quality and the quantity of data itself and because of that, the collection of incident data should be performed so that all required fields are filled properly. The current HSEES data has relatively adequate variables for analysis however there are many missing or blank fields. Improvement on the follow-up actions on the incident reporting is essential in order to complete the data and to generate comprehensive analysis. The incident collection system should invest time, effort and money to ensure that proper and required information are provided for each reported incident.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The analysis of the HSEES chemical incident database has been performed using statistical means, data, and text mining. The results shown in this research prove that data and text mining methodologies are powerful ways to process large data and to obtain valuable information from it. Data mining exhibits potential for analyzing data describing the severity of injuries or fatalities that were experienced by victims. Text mining shows great potential for analyzing incident scenarios and possible severity information, given the required data is provided. Several conclusions can be drawn from this research:

- Trend analysis provided an overall picture of the HSEES chemical incident data from 2002-2006 and served as a reference to guide the data and text mining analysis. The trend analysis results indicate that the number of incidents increased, as well as the number of injuries caused by those incidents. The most prominent type of release and contributing cause were vapor release and equipment failure, respectively. Among frequently released chemicals are ammonia and chlorine. The chemicals were released primarily in the quantity category of $100 > Q \geq 1{,}000$ lbs.
- Cluster analysis using structured data did not produce discernable clusters. Clustering using text inputs produced meaningful groups of incidents.
- The relationship between the release quantity, distance between the victims and the location of the source of release, and the severity of the incident can be evaluated using classification and regression trees. The analysis shows that the relationship between release quantity and distance between the victims and the source of release were proportional and inversely proportional to the severity of the incidents, respectively. However, in several cases, the relationships did not apply because there were more factors that affected the severity of the incidents.

- The association rule produced probability values of incidents involving particular equipment and chemicals; piping and ancillary equipment, and ammonia, chlorine, benzene and hydrogen sulfide, respectively. The two lift values produced were lower than the lift values produced from previous research.
- Text mining was performed to identify typical scenarios that were reported to the HSEES chemical incident database. The scenarios were built using feature selection tools and cluster analysis.

Data and text mining shows that there is more that can be done with chemical incident data than trend analysis. The quality and extent of the analysis strongly depends on the depth and the accuracy of the data. Therefore, there is much to be improved on the existing chemical incident data collection system in order to obtain desired information needed by the process industry. This research has also pointed out several areas for improving the chemical incident database, which focused on gathering data that has in-depth information.

5.2 Recommendations for Future Work

There are many areas of this research that can be improved further. The cluster analysis using response variables can be performed using more variables in order to include more characteristics of the incidents. The association rules can be performed further by segmenting the incident data based on the type of industry as well as the process involved in the incident, so the resulting probability and lift values become more specific. The classification and regression tree analysis can be enhanced by taking into account the type of equipment that was involved in the incidents. This information can be retrieved from the text variable. The data can be also segmented based on type of the release and the type of chemicals released in order to get a more detailed description of the incident pathways.

The start words list of the text mining can be edited to contain the words describing the cause of incidents (*e.g.*, gasket, tank, overpressure, corrosion), safety barriers which failed during the incidents (*e.g.*, safety interlock failure, alarm malfunctions), and the

consequence of incidents (*e.g.*, fire, explosion, relief, flare). The text variables can also be joined with the categorical variables in order to capture more information from the database.

REFERENCES

Al-Qurashi, F. (2000). Development of a relational chemical process safety database and applications to safety improvements. M.S. Thesis, Texas A&M University.

Anand, S. (2005). Novel applications of data mining methodologies to incident databases. M.S. Thesis, Texas A&M University.

ATSDR. (2004). *Hazardous substances emergency events surveillance protocol*. Atlanta, GA: Division of Health Studies, Agency Toxic Substances Disease Registry, U.S. Department of Health and Human Services.

ATSDR. (2006). *HSEES Annual Report 2006*. Atlanta, GA: Division of Health Studies, Agency for Toxic Substances and Disease Registry, US Department of Health and Human Service.

Berry, MW & Browne, M. (1999). *Understanding search engines: mathematical modeling and text retrieval*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Bunn, TL., Slavova, S & Hall, L. (2008). Narrative text analysis of Kentucky tractor fatality reports. *Accident Analysis & Prevention*, *40* (2): 419-425.

Cerrito, P. (2006). *Introduction to data mining using SAS Enterprise Miner™*. Cary, NC: SAS Institute Inc.

Edelstein, H. (1999). *Introduction to data mining and knowledge discovery*. Potomac, MD: Two Crows Corporation.

Han, J & Kamber, M. (2006). *Data mining: concepts and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Hand, DJ., Mannila, H & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: MIT Press.

Jones, S., Kirschsteiger, C & Bjerke, W. (1999). The importance of near miss reporting to further improve safety performance. *Journal of Loss Prevention in the Process Industries*, *12*: 59-67.

Khan, SS. (2010). Active and knowledge-based process safety incident retrieval system. M.S. Thesis, Texas A&M University.

Mannan, MS., O'Connor, TM & West, HH. (1999). Accident history database: an opportunity. *Environmental Progress*, *18* (1): 1-6.

Markowski, AS. (2006). *Layer of protection analysis for the process industries*. Lódz, Poland: Polska Akademia Nauk.

Matignon, R. (2007). *Data mining using SAS Enterprise Miner™*. Hoboken, NJ: John Wiley and Sons, Inc.

MIT. (2002). *Singular value decomposition tutorial*. Retrieved January, 2011, from http://web.mit.edu/be.400/www/SVD.

MKOPSC. (2006). *Propane incident data collection project phase III & IV: full report* College Station: Mary Kay O'Connor Process Safety Center, Texas A&M University.

MKOPSC. (2009). *Developing a roadmap for the future of national hazardous substances incidents surveillance* College Station: Mary Kay O'Connor Process Safety Center, Texas A&M University.

Nisbet, R., Elder, JF & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Oxford, UK: Academic Press/Elsevier.

Obidullah, A. (2006). Use of incident databases for cause and consequence analysis and national estimates. M.S. Thesis, Texas A&M University.

Prem, KP., Ng, D & Mannan, MS. (2010). Harnessing database resources for understanding the profile of chemical process industry incidents. *Journal of Loss Prevention in the Process Industries*, *23* (4): 549-560.

Raja, U & Tretter, M. (2010). Classification of software patches: a text mining approach. *Journal of Software Maintenance and Evolution: Research and Practice*, *23* (2): 69-87.

Statsoft. (2008). STATISTICA (data analysis software system), version 8.0. Tulsa, Oklahoma.

Uth, HJ. (1999). Trends in major industrial accidents in Germany. *Journal of Loss Prevention in the Process Industries*, *12* (1): 69-73.

Veltman, L. (2008). Incident data analysis using data mining techniques. M.S. Thesis, Texas A&M University.

Wakakura, M & Iiduka, Y. (1999). Trends in chemical hazards in Japan. *Journal of Loss Prevention in the Process Industries*, *12* (1): 79-84.

Welles, WL., Wilburn, RE., Ehrlich, JK & Kamara, JM. (2009). New York Hazardous Substances Emergency Events Surveillance (HSEES) data support emergency response, promote safety and protect public health. *Journal of Loss Prevention in the Process Industries*, *22* (6): 728-734.

APPENDIX

Stop words list

event
notification
release
report
small
unit
workers
Q20
Q60
Q65
unknown
call
caller
area
atmosphere
build
one
found
time
respond
went
time
cause
nrc
released
per
due
company
system
result
plant
response
phone
caused
had
material
lbs
employee
personnel
people
home
school
incident
oper
equip
causes
prp
came
allow
hospital
locate
minute
contain
do
provide
quantity
inform
back
made
put
go
will
work

Phrases and synonyms

equipment failure
shut down
start up
relief valve
safety valve
control valve
float valve
loose fit
upset
process vessel
heat exchanger
operating condition
back pressure
pressure wave
water curtain
natural gas
work permit
boiling liquid expanding vapor explosion
permit to work
personal protective equipment
personal protective gear
distributed control system
rail tank
fire fighter
fire water
fire department
runaway reaction
flammable vapor
flexible hose
gas cloud
ignition source
transfer tank
waste water
vapor cloud

Extracted words from text mining process

Word-*i*	Word	Count
1	Leak	712
2	Valve	560
3	Tank	466
4	Line	392
5	Ammonia	343
6	Failure	338
7	Flare	329
8	Process	285
9	Air	278
10	Pressure	278
11	Maintenance	276
12	Repair	232
13	Causing	229
14	Spill	227
15	Fail	223
16	Fire	222
17	Water	217
18	Vent	210
19	Compressor	206
20	Chlorine	194
21	Shutdown	191
22	Pump	181
23	Pipe	179
24	Facility	179
25	Secure	175
26	Open	172
27	Power	166
28	Occur	165
29	Gas	158
30	Equipment failure	156
31	Shut down	149
32	Reactor	137
33	Storage	128
34	Shut	127
35	Evacuate	126
36	Relief valve	120
37	Control	120
38	Malfunction	107
39	Trip	105
40	High	104
41	Chemical	104
42	Close	100
43	Refrigerant	100
44	Upset	99
45	Product	98
46	Cool	98

47	Clean	97
48	Start	96
49	Discovered	95
50	Releases	94
51	Broke	92
52	Oxide	91
53	Vapor	86
54	Seal	85
55	Mercury	83
56	Tower	82
57	Rupture	80
58	Tube	79
59	Plan	78
60	Site	78
61	Remove	78
62	Problem	77
63	Isolate	75
64	Use	75
65	Receive	75
66	Stop	73
67	Two	73
68	Ground	72
69	Exchange	71
70	Level	70
71	Left	68
72	Cylinder	68
73	Stack	68
74	Employees	656
75	Startup	65
76	Electric	65
77	Error	65
78	Drum	65
79	Fd	64
80	Hose	64
81	Pound	64
82	Nurse	63
83	Acid	62
84	Loss	61
85	Estimate	61
86	Boiler	61
87	Steam	60
88	Developed	60
89	Bag	60
90	Hour	59
91	Gasket	59
92	State	58
93	Inside	58
94	Thermal	58
95	Liquid	57
96	Amount	57

97	Monitor	57
98	Block	56
99	Heat	55
100	Overpressure	55
101	Relief	55
102	Room	54
103	Lift	54
104	Hazmat	54
105	Roof	53
106	Feed	53
107	Onto	53
108	Scrubber	53
109	Emission	53
110	Load	53
111	Transfer	53
112	Crack	52
113	Temperature	52
114	Sewer	51
115	Freon	51
116	Flow	50
117	Safety	50
118	Column	50
119	Evac	49
120	Incinerator	49
121	Condense	49
122	Vessel	49
123	Materi	49
124	Hole	48
125	Fit	48
126	Day	48
127	Treatment	48
128	Waste	47
129	Enter	47
130	Lab	47
131	Discharge	46
132	Anhydrous	45
133	Outside	45
134	Blew	45
135	Truck	44
136	Plug	44
137	Service	44
138	Taken	43
139	Change	43
140	Pipeline	43
141	Drain	43
142	Injury	43
143	Contractor	43
144	Burner	43
145	Flange	42
146	Emerge	42

147	Investigate	42
148	Severe	42
149	Intent	42
150	Fill	41
151	Indicate	41
152	Outage	41
153	Safety valve	41
154	Vict	41
155	Resident	40
156	Lost	40
157	Storm	39
158	Replace	39
159	Condition	39
160	Odor	39
161	Active	39
162	Located	39
163	Place	39
164	Faulty	39
165	Scene	38
166	Fire department	38
167	Excess	38
168	Relieve	38
169	Hydrogen	37
170	Notice	37
171	Possible	37
172	Heat Exchanger	37
173	Restart	37
174	Old	36
175	Start up	35
176	Reduction	35
177	Broken	35
178	Smell	34
179	Human	34
180	Detect	34
181	…	…

VITA

Mahdiyati received her Bachelor of Science degree in chemical engineering from Institut Teknologi Bandung, Indonesia in 2006. She started the Master of Science in safety engineering program at Texas A&M University in September 2008 and at the same moment participated in the Mary Kay O'Connor Process Safety Center under the guidance of Dr. Sam Mannan. Her research interests include chemical incident data analysis, consequence modeling, quantitative risk assessment and green engineering. She received her M.S. in safety engineering in May 2011.

Ms. Mahdiyati may be reached at Jack E. Brown Engineering Building, 3122 TAMU Room 200, College Station, Texas, 77843 or through email: *mahdiyatis@gmail.com*.