Frequency and Consequence Modeling of Rare Events using Accident Databases

Meel A., O'Neill L. M., and Seider* W. D.
Department of Chemical and Biomolecular Engineering
University of Pennsylvania
Philadelphia, PA 19104-6393

Oktem U.
Risk Management and Decision Center, Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6340

Keren N.
Department of Agricultural and Biosystems Engineering
Iowa State University
Ames, IA 50011-3080

Prepared for Presentation at
American Institute of Chemical Engineers
2006 Spring National Meeting
8th Process Plant Safety Symposium
Orlando, Florida
April 24-26, 2006

* Corresponding author: Email: seider@seas.upenn.edu,, Ph: 215-898-7953

Frequency and Consequence Modeling of Rare Events using Accident Databases

Meel A., O'Neill L. M., and Seider* W. D.
Department of Chemical and Biomolecular Engineering
University of Pennsylvania
Philadelphia, PA 19104-6393

Oktem U.
Risk Management and Decision Center, Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6340

Keren N.
Department of Agricultural and Biosystems Engineering
Iowa State University
Ames, IA 50011-3080

Abstract:

Accident databases (NRC, ATSDR's, RMP, and others) contain records of incidents (e.g., releases and spills) that have occurred in United States chemical plants during recent years. For various chemical industries, Kleindorfer and coworkers [1] summarize the accident frequencies and severities in the RMP*Info database. Also, Anand and coworkers [2] use data mining to analyze the NRC database for Harris County, Texas.

Classical statistical approaches are ineffective for low frequency, high consequence events because of their rarity. Given this information limitation, this paper uses Bayesian theory to forecast incident frequencies, their relevant causes, equipment involved, and their consequences, in specific chemical plants. Systematic analyses of the databases also help to avoid future accidents, thereby reducing the risk.

More specifically, this paper presents dynamic analyses of incidents in the NRC database. Probability density distributions are formulated for their causes (e.g., equipment failures, operator errors, etc.), and equipment items are utilized within a particular industry. Bayesian techniques provide posterior estimates of the cause and equipment-failure probabilities. Cross-validation techniques are used for checking the modeling, validation, and prediction accuracies. Differences in the plant- and chemical-specific predictions with the overall predictions are demonstrated. Furthermore, the NRC database is exploited to model the rate of occurrence of incidents in various chemical and petrochemical companies using Bayesian theory. Extreme value theory is used for consequence modeling of rare events by formulating distributions for events over a threshold value. Finally, the fast-Fourier transform is used to estimate the risk within an industry utilizing the frequency and severity distributions.

* Corresponding author: Email: seider@seas.upenn.edu,, Ph: 215-898-7953

1. Introduction

Since the accidents at Flixborough, Seveso, and Bhopal, the reporting of abnormal events in the chemical industries has been encouraged to collect accident precursors. Efforts to increase the reporting of near-misses, with near-miss management audits, have been initiated by the Wharton Risk Management Center [3]. In addition, the AIChE Center for Chemical Process Safety (CCPS) has facilitated the development of a Process Safety Incident Database (PSID) to collect and share incident information, permitting industrial participants access to the database, while sharing their collective experiences [4]. Finally, the Mary Kay Safety Center at Texas A&M University (TAMU) [2, 5] has been gathering incident data in the chemical industries.

An incident/accident database, involving oil, chemical, and biological discharges into the environment in the U.S. and its territories, is maintained by the National Response Center (NRC) [6]. To record accidents, European industries submit their data to a common server [7], while chemical companies in the United States are required to report accident data every five years under the RMP Rule [1, 8]. The latter applies only when one or more of the chemicals on the EPA list is stored above the threshold amount indicated for each chemical.

Several researchers have been analyzing and investigating incident databases to identify common trends and to estimate risks. For example, Chung and Jefferson [9] have developed an approach to integrate accident databases with computer tools used by chemical plant designers, operators, and maintenance engineers, permitting accident reports to be easily accessed and analyzed. In addition, Sonnemans et al. [10] have investigated 17 accidents that have occurred in a petrochemical industry in the Netherlands and have demonstrated qualitatively that had accident precursor information been recorded, with proper measures to control future occurrences, these accidents could have been foreseen and even prevented. Furthermore, Sonnemans and Korvers [11] observe that even after observing accident precursors and disruptions, the operating systems inside companies often fail to prevent accidents. The results of yet another analysis feature the lessons learned from the major accident and near-miss events in Germany from 1993-96 [12, 13]. Finally, Elliott et al. [14] analyze the frequency and severity of accidents in the RMP database with respect to socioeconomic factors and found that larger chemically-intensive companies are located in counties with larger African-American populations and with both higher median incomes and higher levels of income inequality. Note that accident precursors have been studied also in railways, nuclear plants, health science centers, aviation, finance companies, and banking systems.

On the risk estimation frontier, Kirchsteiger [15] discusses the strengths and weaknesses of probabilistic and deterministic methods in risk analysis using illustrations associated with nuclear and chemical plants. It is argued that probabilistic methods are more cost-effective, giving results that are easier to communicate to decision and policy makers. In addition, Goossens and Cooke [16] describe the application of two risk assessment techniques involving: (i) formal expert judgment to establish quantitative subjective assessments of design and model parameters, and (ii) system failure analysis, with

accident precursors, using operational evidence of system failures to derive the failure probability of the system. Furthermore, a HORAAM (human and organizational reliability analysis in accident management) method is introduced to quantify human and organizational factors in accident management using decision trees [17].

In this work, statistical methods are introduced to estimate the operational risk of individual chemical companies using the accident precursors reported by the NRC, with the risk estimated as the product of the frequency and the number of consequences. First, the frequency of abnormal events of the individual companies, on a yearly basis, is formulated using Bayesian theory. Later, a loss-severity distribution of the abnormal events is modeled using extreme value theory (EVT). Subsequently, the operational risk of the individual chemical industries is computed by performing fast-Fourier transforms (FFT) of the frequency and loss-severity distributions to obtain the aggregate loss distribution – although these calculations are being completed. The Bayesian theory upgrades any prior information using data to increase the confidence level in modeling the frequency of abnormal events; decreasing the uncertainty in decision-making with annual information upgrades [18]. Through EVT, both extreme and unusually rare events, which characterize incidents reported in the chemical industries, are modeled effectively. Note that EVT has been applied in structural, aerospace, ocean, and hydraulic engineering [19]. Herein, EVT is introduced to measure the operational risk in the chemical industries.

This approach to measuring risks in specific companies provides a quantitative framework for decision-making at higher levels. Using the platform provided, chemical industries should be encouraged to collect accident precursor data more regularly. Through implementation of this dynamic risk assessment methodology, improved risk management strategies should result. Also, the handling of third party investigations should be simplified after accidents.

1.1 Overview

More specifically, herein, the NRC database is explored for seven companies, including petrochemical and specialty chemical manufacturers, in Harris County, Texas. Figure 1 shows the algorithm to be described for calculating the operational risk of a chemical company. Initially, Bayesian models for the frequency of abnormal events are developed. Note that significant differences in the prediction of abnormal events are observed for the individual companies, as compared with predictions obtained when the incidents from all of the industries are lumped together. In addition, Bayesian models are developed to provide the frequency distribution of the day of the week on which the incidents occur, the equipment types involved, the causes behind the incidents, the chemicals involved, the equipment reliability, and the human reliability. Furthermore, the failure probabilities of the process units, as well as the causes of the incidents, are predicted. To obtain the loss-severity distribution, EVT is applied to the NRC database by formulating a quantitative index for the loss as a weighted sum of the different types of consequences. Then, FFTs are applied to obtain the aggregate loss distribution, also known as the aggregate severity distribution. Relevant information, like the Value at

Risk (VaR), is obtained from the loss distribution. Again, these calculations are being completed.

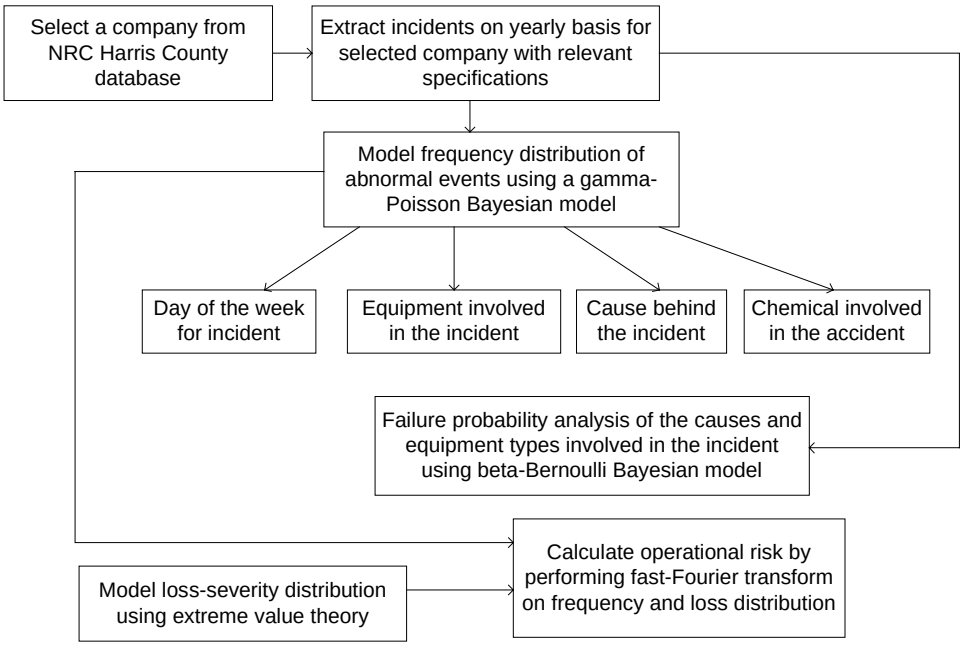


Figure 1. Algorithm to calculate the operational risk of a chemical company

2. Modeling the frequency of abnormal events

This Bayesian model formulates the frequency of occurrence of an abnormal event in a time interval for a company. The possible number of occurrences of an abnormal event in each time interval is a non-negative, integer-valued outcome that can be formulated using the *Poisson* distribution for y:

$$y \sim p(y = y_l) = \left\{ \frac{\lambda^{y_l} e^{-\lambda}}{y_l!} \right\}, \qquad y_l \in \{I^1\}, y_l \geq 0,\ \lambda > 0 \tag{1a}$$

where y_l is the number of abnormal events in time interval l, and λ is the average number of abnormal events in the time intervals, with the expected value, $E(y)$, and variance, $V(y)$, equal to λ. Due to uncertainty, the prior distribution for λ is assumed to follow a *Gamma* distribution, $\lambda \sim Gamma(\alpha, \beta)$:

$$p(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}, \qquad \alpha > 0,\ \beta > 0 \tag{1b}$$

From Baye's theorem, the posterior distribution, $p(\lambda \mid Data)$, is:

$$p(\lambda \mid Data) \propto l(\lambda \mid Data)\, p(\lambda) \propto (\lambda^s e^{-N_t \lambda})(\lambda^{\alpha-1} e^{-\beta\lambda}) \propto \lambda^{(\alpha+s)-1} e^{-(\beta+N_t)\lambda} \tag{1c}$$

where $Data = (y_0, y_1, \ldots, y_{N_t})$, $s = \sum_{l=0}^{N_t} y_l$, N_t is the number of times intervals, and $l(\lambda \mid Data)$ is the *Poisson* likelihood distribution. Note that $p(\lambda \mid Data)$ is also a *Gamma* distribution, $Gamma(\alpha+s,\ \beta+N_t)$, because λ is distributed according to $Gamma(\alpha, \beta)$,

which is a conjugate prior to the *Poisson* distribution. The mean of the posterior distribution is the weighted average of the means of the prior and likelihood distributions:

$$\frac{\alpha + s}{\beta + N_t} = \frac{\beta}{\beta + N_t}\left(\frac{\alpha}{\beta}\right) + \frac{N_t}{\beta + N_t}\frac{s}{N_t} \tag{1d}$$

and the variance of the posterior distribution is $(\alpha + s)/(\beta + N_t)^2$.

The predictive distribution to estimate the number of abnormal events in the next time interval, y_{N_t+1}, conditional on the observed *Data*, is discussed by Meel and Seider [20]. This gives a predictive mean, $(\alpha + s)/(\beta + N_t)$, and predictive variance, $(\alpha + s)/(\beta + N_t)[1+1/(\beta + N_t)]$, and consequently, the posterior and predictive means are the same, while the predictive variance exceeds the posterior variance.

Model-checking using predictive distributions. To check the accuracy of the model, the number of abnormal events in interval l, y_l, is removed, leaving the data, $y_{-l} = (y_0,\ldots, y_{l-1}, y_{l+1},\ldots, y_{N_t})$, over N_t -1 time intervals. Then, the current Bayesian model applied to y_{-l} is used to predict y_l. Finally, y_l and $E[y_l \mid y_{-l}]$ are compared, and predictive z-scores are used to measure their proximity:

$$z_l = \frac{y_l - E[y_l \mid y_{-l}]}{\sqrt{V[y_l \mid y_{-l}]}} \tag{2}$$

For a good model, the mean and standard deviation of $z = (z_0,\ldots, z_{N_t})$ should approach zero and one, respectively.

3. Analysis of NRC database

The NRC database contains reports on all of the oil, chemical, radiological, biological, and etiological discharges into the environment anywhere in the United States and its territories [6]. A typical incident report includes the chemical involved, the cause of the incident, the equipment involved, the date of the incident, the volume of the chemical release, and the extent of the consequences. Herein, the incidents reported for Harris County, Texas, for fixed facilities during 1990-2002 are analyzed to determine their frequencies and consequences (loss or severity). This dataset was obtained from the Mary Kay Safety Center at TAMU, which filtered the NRC database for Harris County, taking care to eliminate duplications of incidents when they occurred. More specifically, the filtered dataset by Anand et al. [2] is used herein for further processing.

The equipment is classified into 13 major categories: electrical equipment (E_1), pumps/compressors (E_2), flare stacks (E_3), heat-transfer equipment (E_4), hoses (flexible pipes) (E_5), process units (E_6), process vessels (E_7), separation equipment (E_8), storage vessels (E_9), pipes and fittings (E_{10}), unclassified equipment (E_{11}), relief equipment (E_{12}), and unknowns (E_{13}). The Harris County database includes several causes of the

incidents, including equipment failures (EF), operator errors (OE), unknown causes (U), dumping, and others, with the EF and OE causes being the most significant.

3.1 Statistical analysis of incidents at chemical companies

Table 1 shows incidents extracted from the NRC database for the seven companies located in Harris County. The total number of incidents, N_{total}, and the number of incidents of equipment failures, N_{EF}, operator errors, N_{OE}, and due to unknown causes, N_U, are listed during the years 1990-2002. In addition, from the 13 equipment categories, the number of process units, N_{PU}, process vessels, N_{PV}, storage vessels, N_{SV}, compressors/pumps, $N_{C/P}$, heat-transfer equipment, N_{HT}, and transfer-line equipment, N_{TL}, are included.

Table 1. Number of incidents for seven companies in the NRC database

Companies	Type	N_{total}	N_{EF}	N_{OE}	N_U	N_{PU}	N_{PV}	N_{SV}	$N_{C/P}$	N_{HT}	N_{TL}
A	Petrochemical	688	443	56	101	59	50	101	86	58	121
B	Petrochemical	568	387	48	88	110	37	69	127	47	56
C	Specialty chemical	401	281	35	46	45	93	61	10	28	77
D	Petrochemical	220	122	24	16	25	47	16	36	27	15
E	Specialty chemical	119	77	21	8	13	12	22	11	12	23
F	Specialty chemical	83	57	14	7	6	6	21	8	10	18
G	Specialty chemical	18	9	2	5	1	1	1	1	3	2

For each of the seven companies, several predictions of abnormal events for future years are carried out utilizing data from previous years, including the prediction of the total number of incidents, N_{total}, incidents associated with each equipment type, and incidents associated with each cause.

Figures 2a and 2b show the predictions of the number of incidents for companies B and F which are chosen arbitrarily to illustrate the variations in the predictive power of the models. In these figures, the number of incidents for the year n are forecasted using the gamma-Poisson Bayesian techniques based on the number of incidents from 1990 to n-1, where n = 1991, 1992, …, 2002. These are compared to the number of incidents that occurred in year n for companies B and F, respectively.

In the absence of information to model the prior distribution for the year 1990, α and β are assumed to be 0.001, providing a relatively flat distribution in the region of interest; that is, a non-informative prior distribution. Note that information upon which to base the prior parameters would enhance the early predictions of the models. This has been illustrated for a beta-Bernoulli Bayesian model, using informative and non-informative prior distributions, showing the sensitivity of the predictions to the prior values [20]. For company B, using non-informative prior distributions, either the numbers of incidents are close to the predicted numbers or higher than those predicted. However, for company F, the numbers of incidents are close to or less than those predicted.

When examining the results for the seven companies, the sizable variations in the number of incidents observed in a particular year are attributed to several factors, including the management and planning efforts to control the incidents, it being assumed that no

significant differences occurred to affect the reporting of the incidents from 1990-2002. Therefore, when the number of incidents is less than those predicted, it seems clear that good incident-control strategies were implemented within the company. Similarly, when the number of incidents is higher than those predicted, the precursor data yields a warning to consider enhancing the measures to reduce the number of incidents in the future.

A good agreement between the numbers of incidents predicted and observed indicates a *stable equilibrium* is achieved with respect to the predictive power of the model. Such a state is achieved when the numbers of incidents and their causes do not change significantly from year-to-year. Note, however, that even as stable equilibrium is approached, efforts to reduce the number of incidents should continue. This is because, even when successful measures are taken year after year (that reduce the number of incidents), the predictive values are usually conservative, lagging behind until the incidence rates converge over a few years.

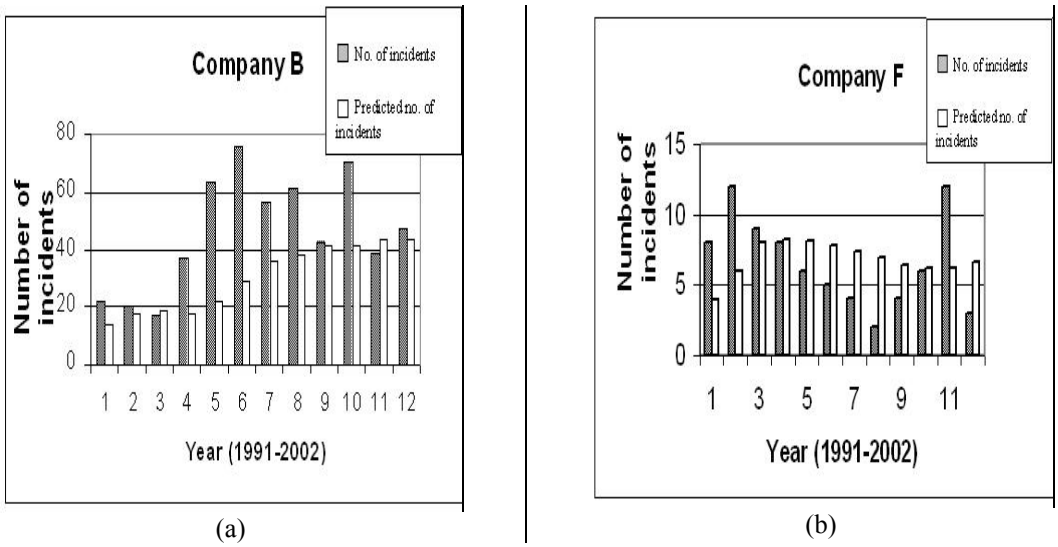
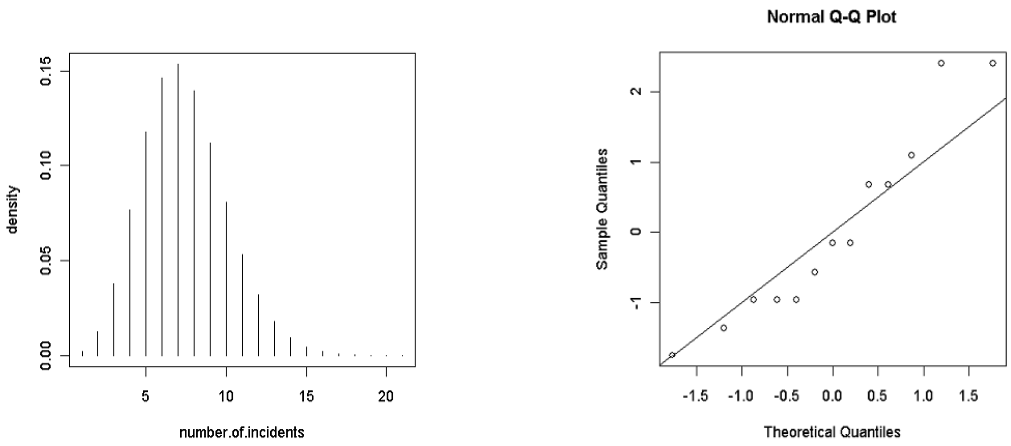


Figure 2. Total number of incidents: (a) Company B, (b) Company F

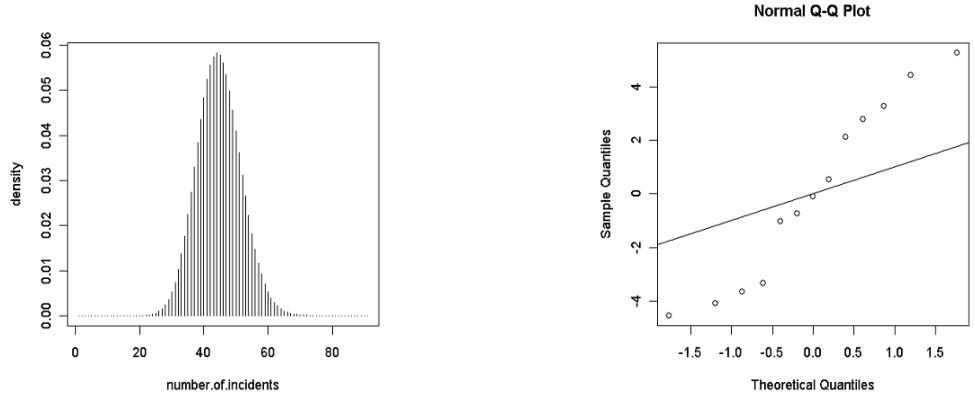
Next, the results of the Bayesian model checking using the *R* software package [21] to compute predictive distributions are presented in Q-Q plots. For company F, Figure 3a shows the density profile of incidents, while Figure 3b shows the normal Q-Q plot, which compares the distribution of z (Eq. (2)) to the normal distribution (represented by the straight line), where the elements of z are represented by circles. The sample quantiles of z (ordered values of z, where the elements, z_i, are called quantiles) are close to the theoretical quantiles (equally-spaced data from a normal distribution), confirming the accuracy of the model predictions. Most of the values are in good agreement, except for two outliers at the theoretical quantiles, 1.0 and 1.5.



(a) (b)

Figure 3. Company F: (a) Density of incidents, (b) Q-Q plot

Figures 4a and 4b show the density profile of incidents and the Q-Q plot for company B. Comparing Figures 4a and 3a, the number of incidents at company B are much higher than at company F. In addition, the variation in the number of incidents in different years is higher at company B (between ~25-65) than at company F (between ~0-15). Note that the circles on the Q-Q plot in Figure 4b depart more significantly from the straight line, possibly due to the larger year-to-year variation in the number of incidents as well as the appropriateness of the of gamma-Poisson distribution. The circles below the straight line correspond to the safe situation where the number of incidents is less than that predicted. However, the circles above the straight line, with the number of incidents higher than those predicted, provide a warning.



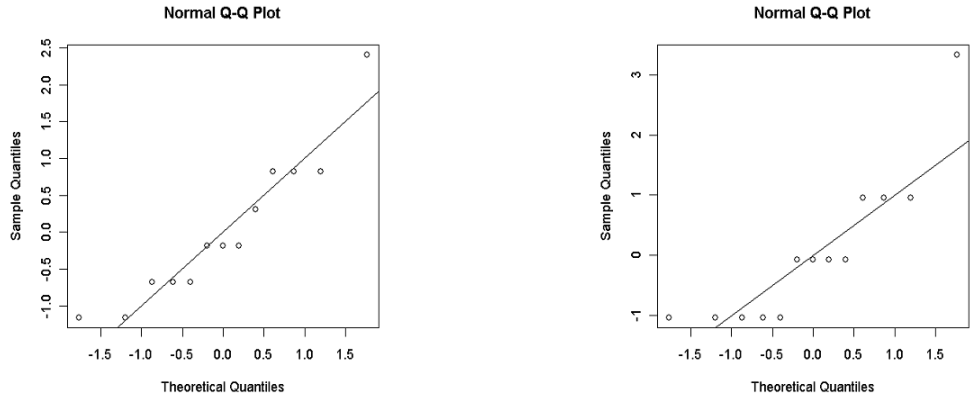
(a) (b)

Figure 4. Company B: (a) Density of incidents, (b) Q-Q plot

To reduce the departures of the circles from the straight line, it is also possible to use a hierarchical Bayesian model. Alternatively, other distributions for the likelihood function, for instance, the negative binomial or the binomial distribution instead of the Poisson distribution can be explored. Note that the Poisson, negative binomial, and binomial distributions belong to the same family, but have a different relationship between the mean and variance of the distributions. The negative binomial distribution is useful when the data mean is less than the variance. The binomial distribution is suitable when the mean exceeds the variance, and the Poisson distribution is preferred when the mean and the variance are in close proximity. Furthermore, the higher number of incidents reported drifts the posterior mean towards the data mean. This is desirable, although poorer predictions may be due to the choice of the distribution. These effects will be examined more closely in future studies.

3.2 Statistical analysis of incident causes and equipment types

In this analysis, for each company, Bayesian models are formulated for each cause and equipment type. Because of the large variations in the number of abnormal events (incidents) observed over the years, the performance of the gamma-Poisson Bayesian models differ significantly. For company F, Figures 5a and 5b show the Q-Q plots for equipment failures and for operator errors, respectively. Figure 5a shows better agreement with the model because the variation in the number of incidents related to equipment failures is small, while the variation in the number of incidents related to operator errors is more significant. This is consistent with the expectation that equipment performance varies less significantly than operator performance over time.

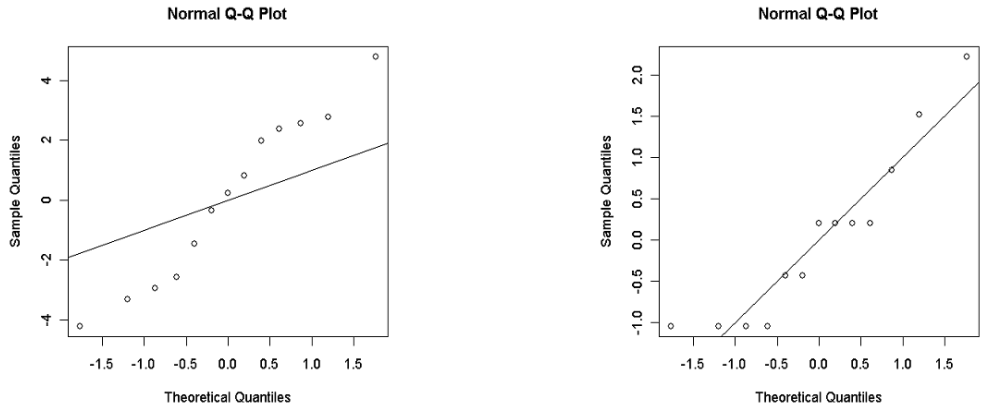


(a) (b)

Figure 5. Company F: (a) Equipment failures, (b) Operator errors

Figures 6a and 6b show the Q-Q plots for equipment failures and for operator errors, respectively, at company B. When comparing Figures 5a and 6a, the predictions of the numbers of equipment failures at company B are poorer than at company F. This is similar to the predictions for the total numbers of incidents at company B, as shown in Figure 4b, compared with those at company F, as shown in Figure 3b. Yet, the

predictions for the operator errors are comparable at companies F and B, and consequently, the larger variation in reporting incidents at company B are attributed to the larger variation in the numbers of equipment failures.



(a) (b)

Figure 6. Company B: (a) Equipment failures, (b) Operator errors

Figures 7a - 7d show the Q-Q plots for incidents associated with the process units, storage vessels, heat-transfer equipment, and compressors/pumps at company A. Figures 7a and 7c are consistent with the small variation in the numbers of incidents per year for process units and heat-transfer equipment, while the variations in the numbers of incidents per year associated with storage vessels and compressors/pumps are more significant in Figures 7b and 7d. Similar analyses show comparable trends for all of the equipment types and causes at each company.

3.3 Statistical analysis of chemicals involved

For each company, an attempt was made to identify trends for each of the top five chemicals associated with the largest number of incidents in the NRC database. However, no specific trends for a particular chemical associated with a higher number of incidents in all of the companies were observed. This could be because the different products are produced in varying amounts by the different companies. It might be preferable to carry out the analysis for a company that manufactures similar chemicals at different locations or for different companies that produce similar products.

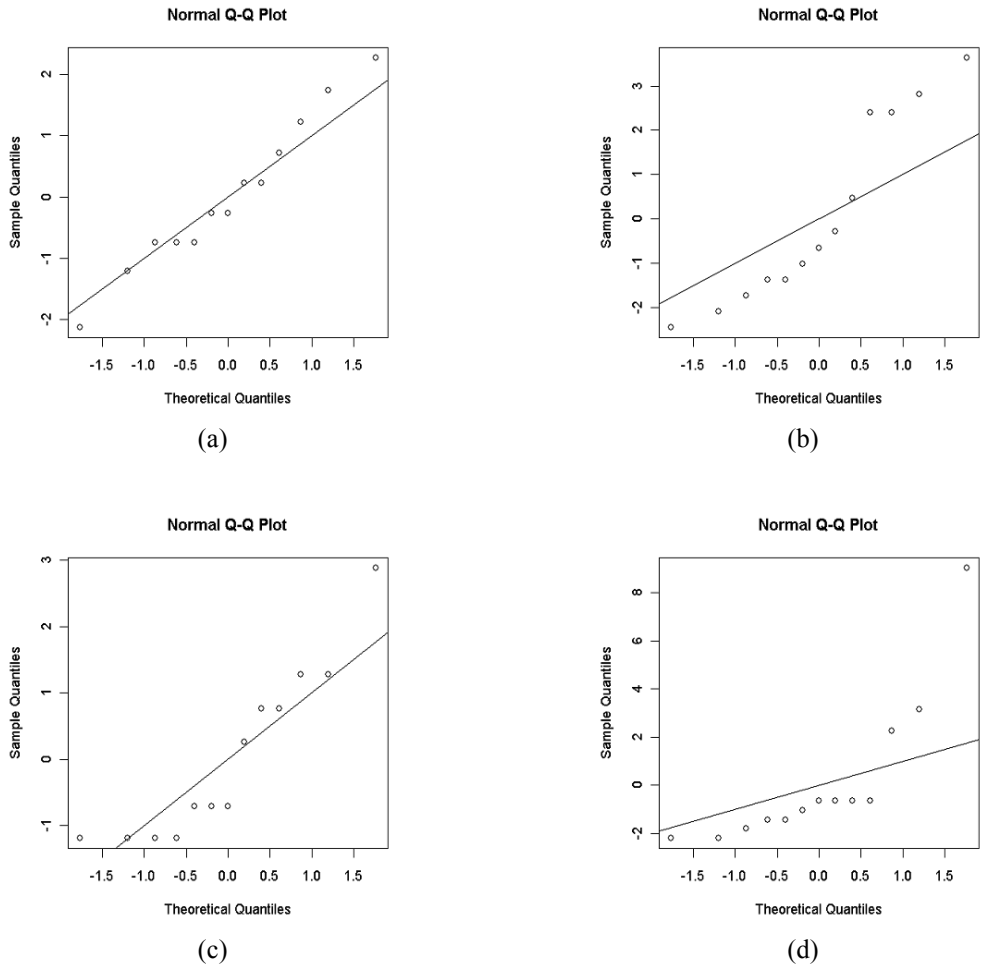


Figure 7. Company A: (a) Process units, (b) Storage vessels, (c) Heat-transfer equipment, and (d) Compressors/pumps

3.4 Statistical analysis of the day of the week

For each of the seven companies, Table 2 summarizes the model checking of the Bayesian predictive distributions, with the mean and variance of z displayed. Again, the predictions improve with the total number of incidents observed for a company. As seen, the mean and variance of z indicate that higher deviations are observed on Wednesdays and Thursdays for almost all of the seven companies. Lower deviations occur at the beginning of the week and over the weekends, which may be attributed to the alertness of the operators at the beginning of the week and fewer operations on the weekends.

To provide a better understanding of this important phenomenon, additional data through operator surveys should be obtained. Furthermore, the higher values of the means and variances for company G on Friday and Saturday suggest that the number of data points is inadequate to generate a reliable Bayesian model. Other analyses that relate the causes

of the incidents to the days of the week can be carried out to identify more specific patterns

Table 2. Q-Q plot properties for day of the week analysis of incidents

	Mon	Tue	Wed	Thru	Fri	Sat	Sun
A	0.027, 1.5	0.015, 1.06	0.032, 1.55	0.046, 1.9	0.023, 1.31	0.022, 1.23	0.055, 1.93
B	0.032, 1.53	0.047, 1.8	0.06, 2.12	0.058, 2.05	0.035, 1.55	0.027, 1.25	0.033, 1.46
C	0.027, 1.28	0.024, 1.21	0.047, 1.67	0.048, 1.62	0.031, 1.33	0.019, 1.002	0.039, 1.48
D	0.15 2.3	0.165, 2.7	0.2, 2.96	0.2, 3.22	0.13, 2.44	0.126, 2.22	0.27, 3.4
E	0.038, 1.06	0.037, 1.19	0.086, 1.66	0.078, 1.64	0.11, 1.89	0.07, 1.46	0.036, 0.96
F	0.034, 1.06	0.06, 1.27	0.04, 1.08	0.87, 0.05	0.035, 0.98	0.043, 1.01	0.07, 1.22
G	0.06, 1.09	0.14, 1.29	0.14, 1.29	0.14, 1.29	7.84, 29.26	15.82, 58.48	0.23, 1.96

3.5 Rates of equipment failures and operator errors

In this section, for an incident, the probabilities of the involvement of each of the 13 equipment types and the probabilities of their causes (e.g., equipment failures, operator errors, and others) are modeled. The tree in Figure 8 shows, for each incident, the possible causes, and for each cause, the possible equipment types. Note that alternatively the tree could show, for each incident, the possible equipment types followed by the possible causes. x_1, x_2, x_3 are the probabilities of causes EF, OE, and O for an incident, and d_1, d_2, d_3 are the cumulative number of incidents at the end of each year. e_1, e_2, e_3, …, e_{13} are the probabilities of the involvement of equipment types, E_1, E_2, …, E_{13}, in an incident through different causes, where $M_1 + N_1 + O_1$, $M_2 + N_2 + O_2$, $M_3 + N_3 + O_3$, …, $M_{13} + N_{13} + O_{13}$ are the cumulative number of incidents associated with each equipment type.

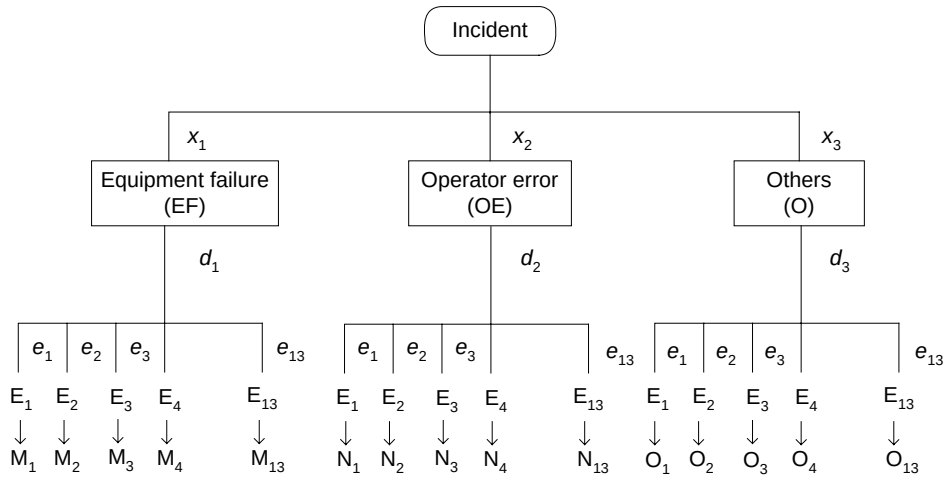


Figure 8. Equipment involved and cause analysis for an incident

The prior distributions of the probability of x_i are modeled using *Beta* distributions with parameters a_i, b_i:

$$f(x_i) \propto (x_i)^{a_i-1}(1-x_i)^{b_i-1}, \quad i=1,\ldots,3 \tag{3}$$

having means = $a_i/(a_i+b_i)$ and variances = $a_ib_i/(a_i+b_i)^2(a_i+b_i+1)$. These conjugate *Beta* prior distributions are updated using *Bernoulli's* likelihood distribution to obtain the posterior distribution of the probability of x_i:

$$f(x_i \mid Data) \propto (x_i)^{a_i-1+d_i}(1-x_i)^{b_i-1+\sum_{k=1,\neq i}^{3} d_k} f(x_i) \tag{4}$$

The posterior distributions, which are also *Beta* distributions having parameters, $a_i + d_i$, and $b_i + \sum_{k=1,\neq i}^{3} d_k$, change at the end of each year as d_i change. a_1 and b_1 are assumed to be 1.0 and 1.0 to give a flat, non-informative, prior distribution; a_2 and b_2 are assumed to be 0.998 and 1.002 to give an non-informative, prior distribution; and a_3 and b_3 are 0.001 and 0.999. Consequently, the mean prior probabilities of EF, OE, and O are 0.5, 0.499, and 0.001, respectively. The posterior means and variances are obtained over the years 1990-2002 for each of the seven companies.

Figures 9a-c show the probabilities of the causes EF, OE, and O for an incident at company F. Using the data at the end of each year, the probabilities increase from 0.5 for equipment failures, decrease from 0.499 for operator errors, and increase from 0.001 for the others, with operator errors approaching slightly higher values than those for the others. Similarly, analyses for equipment types are carried out using *Beta* distributions, $f(e_i)$ and $f(e_i|data)$, with the *data*, $M_1 + N_1 + O_1$, $M_2 + N_2 + O_2$, $M_3 + N_3 + O_3$, …, $M_{13} + N_{13} + O_{13}$. Figure 9d shows, for an incident, that the probability of the involvement of the process vessels (PV) decreases over time. Similarly, the probabilities for the other equipment types approach stable values after a few years with occasional departures from their mean values.

3.5.1 Equipment and human reliabilities

By comparing the causes of incidents between the equipment failures and operator errors, insights regarding equipment and human reliabilities are obtained. In general, for all of the companies, incidents involving equipment failures exceed incidents involving operator errors, even though the OE/EF ratio changes at the end of each year. For petrochemical companies, the ratio is much lower than for specialty chemical companies, as seen in Table 3.

Table 3. OE/EF ratio for the petrochemical (P) and specialty chemical (S) companies

Company	A (P)	B (P)	C (S)	D (P)	E (S)	F (S)	G (S)
OE/EF ratio	0-0.3	0-0.22	0-0.75	0-0.5	0-0.667	0-0.667	0-0.5

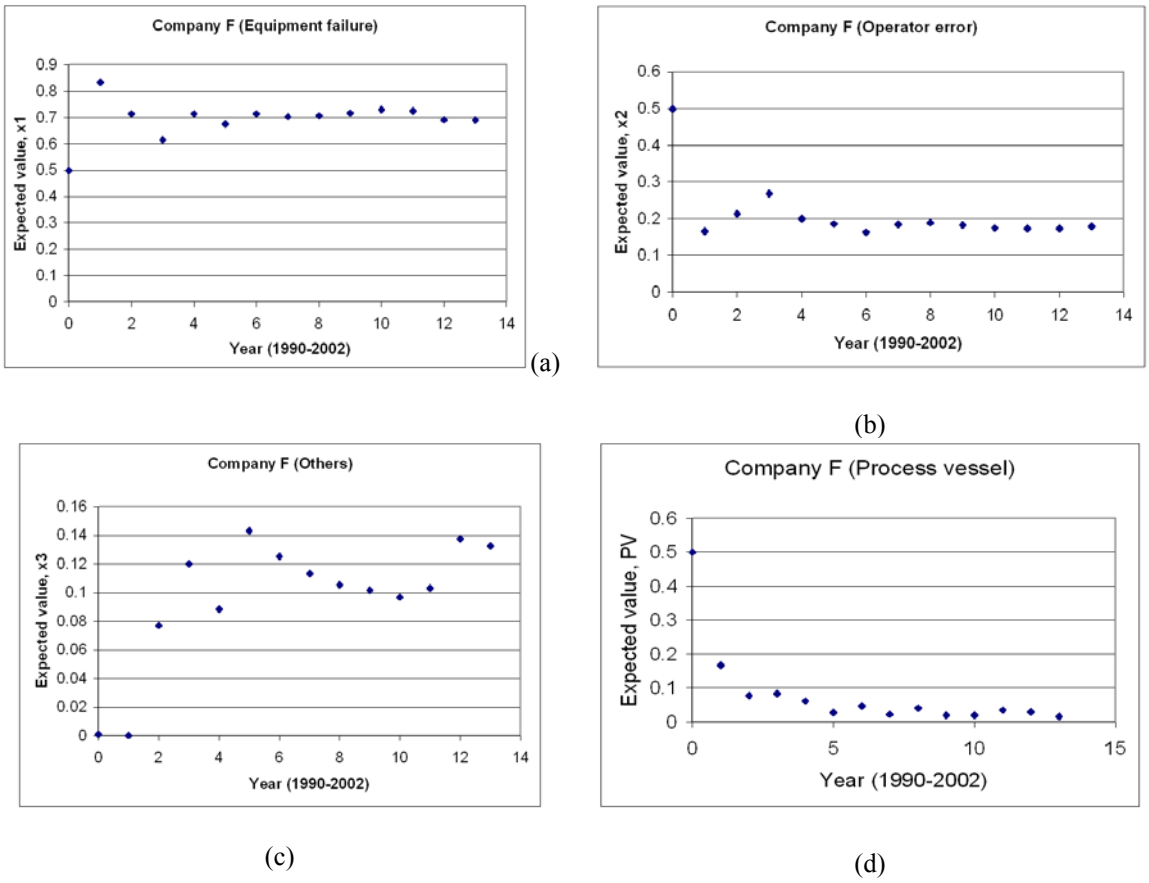


Figure 9. Probabilities for company F: (a) EF, (b) OE, (c) others, (d) PV

3.6 Specialty chemicals and petrochemicals

To identify trends in the manufacture of specialty chemicals and petrochemicals, data for companies C, E, F, and G are combined and compared with the combined data for companies A, B, and D. Note that this is advantageous when the data for a single company are insufficient to identify trends, and when it is assumed that the lumped data for each group of companies are identically and independently distributed. For these reasons, all of the analyses in Sections 3.1 - 3.5 were repeated with the data for specialty chemical and petrochemical manufacturers lumped together. Because the number of data entries in each lumped data set are increased, the circles on the Q-Q plot lie closer to the straight line. However, the cumulative predictions for the specialty chemical and petrochemical manufacturers differ significantly from those for the individual companies. Hence, it is important to carry out company specific analyses. Nevertheless, when insufficient data are available for each company, the cumulative predictions for specialty chemical and petrochemical manufacturers are preferable. Furthermore, when insufficient lumped data are available for the specialty chemicals and petrochemical manufacturers, trends may be identified by combining the data for all of the companies.

3.7 Loss-severity modeling using extreme value theory

For rare events with extreme losses, it is important to identify those that exceed a high threshold. Extreme value theory (EVT) is a powerful and fairly robust framework to study the tail behavior of a distribution. Embrechts et al. [19] provide an overview of extreme value theory as a risk management tool, discussing its potential and limitations. In another study, McNeil [22] examines the estimation of the tails of the loss-severity distributions and the estimation of quantile risk measures for financial time-series using extreme value theory. Herein, EVT is employed to develop a loss-severity distribution for the seven chemical companies.

The distribution of excess values of losses, x, over a high threshold, u, is defined as:

$$F_u(y) = \mathbf{Pr}\{X - u \le y \mid X > u\} = \frac{F(y+u) - F(u)}{1 - F(u)}, \qquad x \in X \tag{5}$$

which represents the probability that the value of x exceeds the threshold, u, by at most an amount, y, given that x exceeds the threshold u, where F is the cumulative probability distribution. For sufficiently high threshold u, the distribution function of the excess may be approximated by the generalized Pareto distribution (GPD), and consequently $F_u(y)$ converges to GPD as the threshold becomes large. The GPD is:

$$G(x) = \begin{cases} 1 - \left(1 + \xi \dfrac{x}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - e^{-x/\beta} & \text{if } \xi = 0 \end{cases} \tag{6}$$

where ξ is the shape parameter and the tail index is $\alpha = \xi^{-1}$. Note that the GPD reduces into different distributions depending on ξ. The distribution of excesses may be approximated by the GPD by choosing ξ and β and setting a high threshold u. The parameters of the GPD can be estimated using various techniques; for example, the maximum likelihood method and the method of probability-weighted moments.

3.7.1 Loss-severity distribution of NRC database

A software package, Extreme Value Analysis in MATLAB (EVIM), is used to obtain the parameters of the GPD for the NRC database [23]. Because few incidents have high severity levels, the incidents analyzed for the seven companies are assumed to be identically and independently distributed (iid). The loss for an incident, L, is calculated as a weighed sum of the numbers of evacuations, injuries, hospitalizations, fatalities, and damages:

$$L = w_{\mathrm{e}} N_{\mathrm{e}} + w_{\mathrm{i}} N_{\mathrm{i}} + w_{\mathrm{h}} N_{\mathrm{h}} + w_{\mathrm{f}} N_{\mathrm{f}} + w_{\mathrm{d}} N_{\mathrm{d}} \tag{7}$$

where w_e = \$100, w_i = \$10,000, w_h = \$50,000, w_f = \$2,000,000, and w_d = 1, with N_d reported in dollars. For the NRC database, the threshold value, u, is chosen to be \$1,000.

As expected, the NRC database has few incidents that have a significant loss. Only 92 incidents among those reported (~3,000) had monetary loss ($L > 0$) and even fewer exceeded the threshold. Note that to obtain a satisfactory prediction of the GPD parameters, usually 100 data points are needed. Therefore, the cumulative distribution of the losses for the NRC database in Figure 10 would be improved with additional data. To achieve this, data from additional companies in Harris County could be included. However, the performance in Figure 10 is considered to be satisfactory. The parameters, $\xi = 0.8688$ and $\beta = 1.7183 \times 10^4$, are computed using the maximum likelihood method.

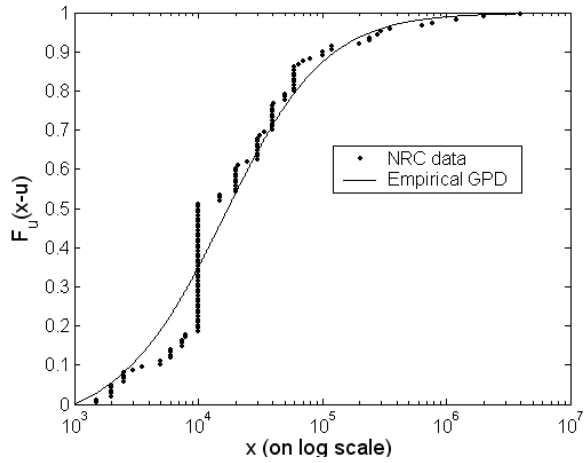


Figure 10. Loss distribution of NRC database

4. Operational risk

Calculations are being completed to estimate the Value-at-Risk (VaR) from the aggregate loss distribution by performing fast-Fourier transforms on the frequency and loss-severity distributions.

5. Conclusions

Statistical models to analyze accident precursors in the NRC database have been developed. They:

1. provide Bayesian models that facilitate improved company-specific estimates, as compared with lumped estimates involving all of the specialty chemical and petrochemical manufacturers.

2. identify Wednesday and Thursday as days of the week in which incidents are more likely to occur.

3. are effective for testing equipment and human reliabilities, indicating that the OE/EF ratio is lower for petrochemical than specialty chemical companies.

4. are expected to be beneficial for obtaining the Value-at-Risk (VaR) from the loss-severity distribution using EVT. Note that the VaR calculations are being completed.

Consistent reporting of incidents is crucial for the reliability of this analysis. In addition, the predictive errors are reduced when: (i) sufficient incidents are available for a specific company to provide reliable means, and (ii) less variation occurs in the number of incidents from year-to-year. Furthermore, to obtain better predictions, it helps to select distributions that better represent the data, properly modeling the functionality between the mean and variance of the data.

References

1. Kleindorfer PR, Belke JC, Elliott MR, Lee K, Lowe RA, Feldman HI. Accident epidemiology and the US chemical industry: Accident history and worst-case data from RMP*Info. Risk Anal. 2003; 23:865-881.
2. Anand S, Keren N, Tretter MJ, Wang Y, O'Connor TM, Mannan MS. *Harnessing data mining to explore incident databases*. *7th Annual Symposium, Mary Kay O'Connor Process Safety Center*. 2004. College Station, TX.
3. Phimister JR, Oktem U, Kleindorfer PR, Kunreuther H. Near-miss incident management in the chemical process industry. Risk Anal. 2003; 23:445-459.
4. CCPS. Process Safety Incident Database (PSID). http://www.aiche.org/CCPS/ActiveProjects/PSID/index.aspx
5. Mannan MS, O'Connor TM, West HH. Accident history database: An opportunity. Environ. Prog. 1999; 18:1-6.
6. NRC. National Response Center http://www.nrc.uscg.mil/nrchp.html.
7. Rasmussen K. The experience with Major Accident Reporting System from 1984 to 1993. European Commission, Joint Research Center, EUR 16341 EN. 1996;
8. RMP. 40 CFR Chapter IV, Accidental Release Prevention Requirements; Risk Management Programs Under the Clean Air Act Section 112(r)(7); Distribution of Off-Site Consequence Analysis Information. Final Rule, 65 FR 48108 2000;
9. Chung PWH, Jefferson M. The integration of accident databases with computer tools in the chemical industry. Comp. Chem. Eng. 1998; 22:S729-S732.
10. Sonnemans PJM, Korvers PMW, Brombacher AC, van Beek PC, Reinders JEA. Accidents, often the result of an 'uncontrolled business process' - a study in the (Dutch) chemical industry. Qual. Reliab. Eng. Intern'l. 2003; 19:183-196.
11. Sonnemans PJM, Korvers PMW. Accidents in the chemical industry: Are they foreseeable? J. Loss Preven. Proc. Ind. 2006; 19:1-12.
12. Uth HJ. Trends in major industrial accidents in Germany. J. Loss Preven. Proc. Ind. 1999; 12:69-73.
13. Uth HJ, Wiese N. Central collecting and evaluating of major accidents and near-miss-events in the Federal Republic of Germany - results, experiences, perspectives. J. Hazard. Mat. 2004; 111:139-145.

14. Elliott MR, Wang Y, Lowe RA, Kleindorfer PR. Environmental justice: frequency and severity of US chemical industry accidents and the socioeconomic status of surrounding communities. J. Epidem. Commun. Health 2004; 58:24-30.
15. Kirchsteiger C. Impact of accident precursors on risk estimates from accident databases. J. Loss Preven. Proc. Ind. 1997; 10:159-167.
16. Goossens LHJ, Cooke RM. Applications of some risk assessment techniques: Formal expert judgement and accident sequence precursors. Safety Sci. 1997; 26:35-47.
17. Baumont G, Menage F, Schneiter JR, Spurgin A, Vogel A. Quantifying human and organizational factors in accident management using decision trees: The HORAAM method. Reliab. Eng. Sys. Safety 2000; 70:113-124.
18. Robert CP. The Bayesian Choice. Springer-VerlagNew York: 2001.
19. Embrechts P., Kluppelberg C., Mikosch T. Modelling Extremal Events. SpringerBerlin: 1997.
20. Meel A, Seider WD. Plant-specific dynamic failure assessment using Bayesian theory. Submitted to Chem. Eng. Sci. 2005;
21. Gentleman R, Ihaka R, Bates D, Chambers J, Dalgaard J, Hornik K. The R project for Statistical Computing. http://www.r-project.org/ 2005;
22. McNeil AJ. Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bulletin 1997; 27:117-137.
23. Gencay R, Selcuk F, Ulugulyagci A. EVIM: A software package for extreme value analysis in MATLAB. Stud. Nonlin. Dynam. Economet. 2001; 5:213-239.