March 2007

Operational risk assessment of chemical industries by exploiting accident databases

A. Meel
University of Pennsylvania

L. M. O'Neill
University of Pennsylvania

J. H. Levin
University of Pennsylvania

Warren D. Seider
University of Pennsylvania, seider@seas.upenn.edu

U. Oktem
University of Pennsylvania

See next page for additional authors

Postprint version. Published in *Journal of Loss Prevention in the Process Industries*, Volume 20, Issue 2, March 2007, pages 113-127. Publisher URL: http://dx.doi.org/10.1016/j.jlp.2006.10.003

Operational risk assessment of chemical industries by exploiting accident databases

Abstract

Accident databases (NRC, RMP, and others) contain records of incidents (e.g., releases and spills) that have occurred in the USA chemical plants during recent years. For various chemical industries, [Kleindorfer, P. R., Belke, J. C., Elliott, M. R., Lee, K., Lowe, R. A., & Feldman, H. I. (2003). Accident epidemiology and the US chemical industry: Accident history and worst-case data from RMP**Info. Risk Analysis*, 23(5), 865–881.] summarize the accident frequencies and severities in the RMP*Info database. Also, [Anand, S., Keren, N., Tretter, M. J., Wang, Y., O'Connor, T. M., & Mannan, M. S. (2006). Harnessing data mining to explore incident databases, the *Journal of Hazardous Material*, 130, 33–41.] 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. The NRC database is exploited to model the rate of occurrence of incidents in various chemical and petrochemical companies using Bayesian theory. Probability density distributions are formulated for their causes (e.g., equipment failures, operator errors, etc.), and associated equipment items 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, 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 capital at risk within an industry utilizing the *frequency* and *loss-severity* distributions.

Keywords

risk, frequency modeling, consequence modeling, abnormal events, chemical plants

Comments

Postprint version. Published in *Journal of Loss Prevention in the Process Industries*, Volume 20, Issue 2, March 2007, pages 113-127. Publisher URL: http://dx.doi.org/10.1016/j.jlp.2006.10.003

Author(s)

A. Meel, L. M. O'Neill, J. H. Levin, Warren D. Seider, U. Oktem, and N. Karen

8.07f/W(Jul162004) +model JLPP : 1830 Prod.Type:FTP pp.1–15(col.fig.:NIL) ED:ShruthiHJ PAGN:vs SCAN:

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Journal of Loss Prevention in the Process Industries ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Journal of Loss Prevention in the process industries

www.elsevier.com/locate/jlp

Operational risk assessment of chemical industries by exploiting accident databases

A. Meel[a], L.M. O'Neill[a], J.H. Levin[a], W.D. Seider[a,*], U. Oktem[b], N. Keren[c]

[a]Department of Chemical and Biomolecular Engineering, University of Pennsylvania, Philadelphia, PA 19104-6393, USA
[b]Risk Management and Decision Center, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, USA
[c]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA 50011-3080, USA

Received 11 June 2006; received in revised form 17 October 2006; accepted 18 October 2006

Abstract

Accident databases (NRC, RMP, and others) contain records of incidents (e.g., releases and spills) that have occurred in the USA chemical plants during recent years. For various chemical industries, [Kleindorfer, P. R., Belke, J. C., Elliott, M. R., Lee, K., Lowe, R. A., & Feldman, H. I. (2003). Accident epidemiology and the US chemical industry: Accident history and worst-case data from RMP*Info. *Risk Analysis*, *23*(5), 865–881.] summarize the accident frequencies and severities in the RMP*Info database. Also, [Anand, S., Keren, N., Tretter, M. J., Wang, Y., O'Connor, T. M., & Mannan, M. S. (2006). Harnessing data mining to explore incident databases. *Journal of Hazardous Material*, *130*, 33–41.] 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. The NRC database is exploited to model the rate of occurrence of incidents in various chemical and petrochemical companies using Bayesian theory. Probability density distributions are formulated for their causes (e.g., equipment failures, operator errors, etc.), and associated equipment items 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, 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 capital at risk within an industry utilizing the *frequency* and *loss-severity* distributions.

Keywords: Risk; Frequency modeling; Consequence modeling; Abnormal events; Chemical plants

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 (Phimister, Oktem, Kleindorfer, & Kunreuther, 2003). 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 indus-

Abbreviations: Companies A, B, C, D, E, F, G, A, B, C, D, E, F, G; Basic indicator approach, BIA; Capital at risk, CaR; Center for chemical process safety (AIChE), CCPS; Equipment failure, EF; Environmental protection agency, EPA; Extreme value theory, EVT; Fast-Fourier transform, FFT; Heat transfer units, HT; Inverse fast-Fourier transform, IFFT; Internal measurement approach, IMA; Loss distribution approach, LDA; Markov-chain Monte Carlo, MCMC; Major accident reporting system, MARS; National response center, NRC; Others, O; Operator error, OE; Occupational safety and health administration, OSHA; Process safety incident database, PSID; Process safety management, PSM; Process units, PU; Process vessels, PV; Quantile-quantile, Q-Q; Risk management plan, RMP; Standardized approach, SA; Storage vessel, SV; Transfer line, TL

*Corresponding author. Tel.: +1 215 898 7953.
E-mail address: seider@seas.upenn.edu (W.D. Seider).

0950-4230/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.jlp.2006.10.003

Nomenclature

a,b	parameters of *Beta* prior probability distribution
a_i, b_i	parameters of prior probability distribution of cause i for an incident
d_1, d_2, d_3	cumulative number of incidents of causes EF, OE, and O at the end of each year
e_i	probability of involvement of equipment type i
$E(\mu\|Data)$	expected posterior mean of μ
$E(q\|Data)$	expected posterior mean of q
$E(y)$	expected value of number of incidents in a year
$E[Y_i\|Y_{-i}]$	expected value of prediction of incident in year i based on incidents in Y_{-i}
$f(e_i)$	prior probability distribution of involvement of equipment i for an incident
$f(x_i\|Data)$	posterior probability distribution of involvement of equipment i conditional upon *Data*
$f(x_i)$	prior probability distribution of cause i for an incident
$f(x_i\|Data)$	posterior probability distribution of cause i conditional upon *Data*
f_l	discrete *loss-severity* distribution function
$f_z(Z)$	discrete probability distribution function of total loss
$F_u(y)$	cumulative probability distribution for distribution of losses, l, over threshold u
$G(l)$	*Generalized Pareto* distribution of losses
l	loss associated with an incident
$M_i+N_i+O_i$	cumulative number of incidents associated with equipment i at the end of each year
n_p	number of points desired in *total loss* distribution
$N_{C/P}$	number of incidents associated with compressors and pumps
N_d	amount of damage, \$
N_e	number of evacuations
N_{EF}	number of incidents associated with equipment failures
N_f	number of fatalities
N_h	number of hospitalizations
N_{HT}	number of incidents associated with heat-transfer equipment items
N_i	number of injuries
N_{OE}	number of incidents associated with operator errors
N_{PU}	number of incidents associated with process units
N_{SV}	number of incidents associated with storage vessels
N_t	number of years
N_{TL}	number of incidents associated with transfer-line equipment
N_{total}	total number of incidents
N_U	number of incidents associated with unknown causes
$p(\lambda)$	prior distribution of λ
$p(\lambda\|Data)$	posterior distribution of λ given *Data*
$p(q\|Data)$	marginal posterior distribution of q given *Data*
$p(\mu\|Data)$	marginal posterior distribution of μ given *Data*
P_N	probability generating function of the frequency of events, N
p_i, q_i	parameters of prior probability distribution of involvement of equipment i in an incident
q	parameter of the *Negative Binomial* distribution
s	total number of incidents in N_t years
u	threshold value of l for *loss-severity* distribution
$V(y)$	variance of number of incidents per year
w_d	dimensionless damage measure
w_e	dollar amount per evacuation, \$
w_f	dollar amount per fatality, \$
w_h	dollar amount per hospitalization, \$
w_i	dollar amount per injury, \$
x_1, x_2, x_3	probabilities of causes EF, OE, and O for an incident
y_i	number of incidents in year i
z_i	predictive score for incidents in year i
Z	total annual loss for a company

Greek

α, β	parameters for *Gamma* density distribution function
$\beta(a, b)$	*Beta* density distribution with parameters a and b
ϕ_l	characteristic function of the *loss-severity* distribution
ϕ_Z	characteristic function of *total loss* distribution
λ	average annual number of incidents
λ_B	average annual number of incidents for company B with losses greater than u
λ_F	average annual number of incidents for company F with losses greater than u
μ	parameter of the *Negative Binomial* distribution
ξ, β	parameters of the *generalized Pareto* distribution
$\gamma(\alpha, \beta)$	*Gamma* distribution with parameters α and β

Subscript

i	year counter
n	year vector

trial participants access to the database, while sharing their collective experiences (CCPS, 1995). Finally, the Mary Kay Safety Center at Texas A&M University (TAMU) has been gathering incident data in the chemical industries (Anand et al., 2006; Mannan, O'Connor, & West, 1999).

An incident database, involving oil, chemical, and biological discharges into the environment in the USA and its territories, is maintained by the national response center (NRC) (NRC, 1990). While companies participate voluntarily, raising reliability concerns, the NRC database for Harris County, Texas, is acknowledged to be reliable thanks to the conscientious efforts of many chemical companies in reporting incidents. Moreover, the Mary Kay Safety Center has concentrated time and resources toward refining the Harris County database to increase its reliability and consistency.

To record accidents, European industries submit their data to the major accident-reporting system (MARS) (Rasmussen, 1996), while a database for chemical companies in the USA is created from risk management plans (RMP) submitted by facilities subject to Environmental protection agency's (EPA) chemical accidental release prevention and response regulations (Kleindorfer et al., 2003; RMP, 2000).

Several researchers have been analyzing and investigating incident databases to identify common trends and to estimate risks. For example, Chung and Jefferson (1998) 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, Korvers, Brombacher, van Beek, and Reinders (2003) have investigated 17 accidents that have occurred in the Netherlands petrochemical industries 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 possibly prevented. Furthermore, Sonnemans and Korvers (2006) observed that even after recognizing 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 to 1996 (Uth, 1999; Uth & Wiese, 2004). Finally, Elliott, Wang, Lowe, and Kleindorfer (2004) analyzed 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 (1997) discussed 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 (1997) described 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 human and organizational reliability analysis in accident management (HORAAM) method was introduced to quantify human and organizational factors in accident management using decision trees (Baumont, Menage, Schneiter, Spurgin, & Vogel, 2000).

In this work, statistical methods are introduced to estimate the operational risk for seven companies, including petrochemical and specialty chemical manufacturers, using the NRC database for Harris County, with the risk estimated as the product of the frequency and consequences of the incidents. Fig. 1 shows the algorithm for calculating the operational risk of a chemical company. For a company in the database, the incidents are extracted on a yearly basis. Then, the frequency distribution of the incidents is estimated using a γ-*Poisson* Bayesian model. Note that significant differences in the prediction of incidents are observed for the individual companies, as compared with predictions obtained when the incidents from all of the companies are lumped together. The Bayesian theory upgrades prior information available, if any, using data to increase the confidence level in modeling the frequency of incidents, decreasing the uncertainty in decision-making with annual information upgrades (Robert, 2001).

Additional γ-*Poisson* 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, and the chemicals involved. In parallel, the failure probabilities of the process units, as well as the causes of the incidents, are predicted using a β-*Bernoulli* Bayesian model.

Later, a *loss-severity* distribution of the incidents is modeled using extreme value theory (EVT) by formulating a quantitative index for the loss as a weighted sum of the different types of consequences. 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 (Embrechts, Kluppelberg, & Mikosch, 1997). Herein, EVT is introduced to measure the operational risk in the chemical industries.

Finally, the operational risk of the individual chemical industries is computed by performing fast-Fourier transforms (FFT) of the product of the *frequency* and *loss-severity* distributions to obtain the *total loss* distribution and the capital at risk (CaR). This approach to measuring risks in specific companies provides a quantitative frame-

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlpp.2006.10.003

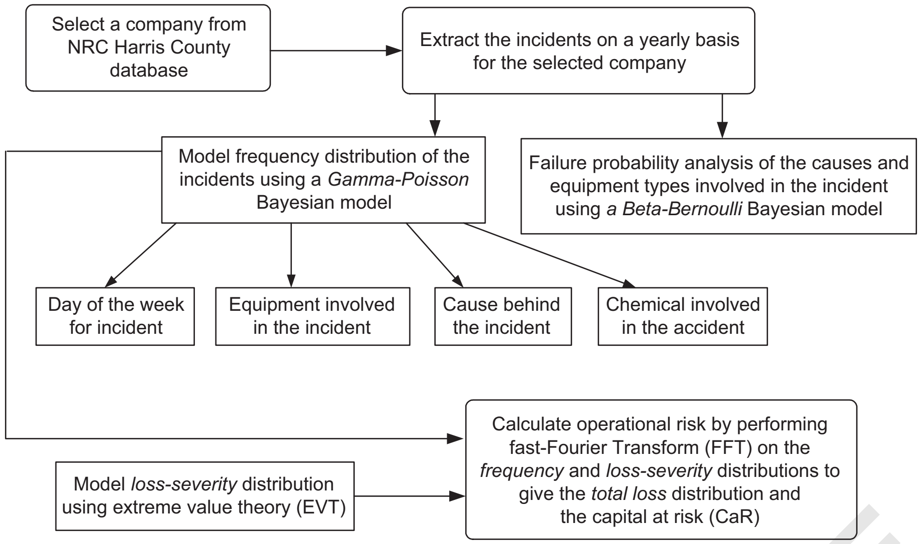

Fig. 1. Algorithm to calculate the operational risk of a chemical company.

work for decision-making at higher levels. Using the platform provided, the 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.

To begin the detailed presentation of this algorithm, Section 2 describes the concepts of Bayesian theory for prediction of the numbers of incidents annually. Then, the NRC database, the Bayesian predictive models, and the *loss-severity* distribution using EVT, are described in Section 3. The CaR calculations using FFTs are discussed in Section 4. Finally, conclusions are presented in Section 5.

2. Modeling the frequency of incidents

Bayesian theory is helpful in formulating the annual frequency of occurrence of incidents for a company. The relationship between the mean and the variance of the annual incidents, over many years, determines the best choice of distribution. For example, the *Poisson* distribution is suitable when the mean and the variance of the data are in close proximity. When the predictions of the *Poisson* distribution are poor, other distributions are used; for instance, the *Negative Binomial* distribution, when the variance exceeds the mean (Bradlow, Hardie, & Fader, 2002).

2.1. Poisson distribution

The annual number of occurrences of an incident is a non-negative, integer-valued outcome that can be estimated using the *Poisson* distribution for y:

$$y \sim p(y = y_i) = \left\{ \frac{\lambda^{y_i} \mathrm{e}^{-\lambda}}{y_i!} \right\}, \quad y_i \in \{I^1\}, y_i \geqslant 0, \quad \lambda > 0, \tag{1a}$$

where y_i is the number of incidents in year i, and λ is the annual average number of incidents, 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 γ-distribution, $\lambda \sim \gamma(\alpha, \beta)$:

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

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

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

where $Data = (y_0, y_1, \ldots, y_{N_t})$, $s = \sum_{i=0}^{N_t} y_i$, N_t is the number of years, and $l(Data|\lambda)$ is the *Poisson* likelihood distribution. Note that $p(\lambda|Data)$ 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 incidents in the next year, y_{N_t+1}, conditional on the observed *Data*, is discussed by Meel and Seider (2006). 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.

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlp.2006.10.003

2.2. Negative binomial distribution

The annual number of occurrences of an incident is a non-negative, integer-valued outcome that can be estimated using the *Negative Binomial* distribution for y:

$$y \sim (q)^{\mu}(1-q)^{y_i} \quad y_i \in \{I^1\}, y_i \geqslant 0, \quad \mu > 0, \quad q \geqslant 0, \tag{1e}$$

where y_i is the number of incidents in year i, and $\mu(1-q)/q$ is the expected annual (mean) number of incidents, $E(y)$, and $\mu(1-q)/q^2$ is the expected variance, $V(y)$. Due to uncertainty, the prior distribution for μ is assumed to follow a *Gamma* distribution, $\mu \sim \gamma(\alpha, \beta)$:

$$p(\mu) \propto \mu^{\alpha-1} e^{-\beta\mu}, \quad \alpha > 0, \quad \beta > 0, \tag{1f}$$

and that for q is assumed to follow a *Beta* distribution, $q \sim \beta(a, b)$:

$$p(q) \propto q^{a-1}(1-q)^{b-1}, \quad a > 0, \quad b > 0. \tag{1g}$$

From Baye's theorem, the posterior distribution, $p(\mu,q|Data)$, is

$$\begin{aligned} p(\mu, q|Data) &\propto l(Data|\mu, q)p(\mu)p(q) \\ &\propto q^{n\mu}(1-q)^{s}(\mu^{\alpha-1}e^{-\beta\mu})q^{a-1}(1-q)^{b-1} \\ &\propto q^{n\mu+a-1}(1-q)^{s+b-1}(\mu^{\alpha-1}e^{-\beta\mu}), \end{aligned} \tag{1h}$$

where $Data = (y_0, y_1, \ldots, y_{N_t})$, $s = \sum_{i=0}^{N_t} y_i$, N_t is the number of years, and $l(Data|\mu,q)$ is the *Negative Binomial* likelihood distribution. The marginal posterior distributions, $p(\mu|Data)$ and $p(q|Data)$, and the posterior means $E(\mu|Data)$ and $E(q|Data)$ are obtained using the Markov Chain Monte-Carlo (MCMC) method in the WINBUGS software (Spiegelhalter et al., 2003). These added calculations are not needed for the *Poisson* distribution, in which the expected value, $E(\lambda|Data)$, is computed easily using Eq. (1d).

2.3. Model-checking

To check the accuracy of the model, the number of incidents in year i, y_i, is removed, leaving the data, $y_{-i} = (y_0, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{N_t})$, over N_t-1 years. Then, a Bayesian model applied to y_{-i} is used to predict y_i. Finally, y_i and $E[y_i|y_{-i}]$ are compared, and predictive z-scores are used to measure their proximity:

$$z_i = \frac{y_i - E[y_i|y_{-i}]}{\sqrt{V[y_i|y_{-i}]}}. \tag{2}$$

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 the NRC database

The NRC database contains reports on the oil, chemical, radiological, biological, and etiological discharges into the environment in the USA and its territories (NRC, 1990). A typical incident report includes the date of the incident, the chemical involved, the cause of the incident, the equipment involved, the volume of the chemical release, and the extent of the consequences. Herein, the incidents reported for Harris County, Texas, for seven specific facilities during the years 1990–2002, are analyzed to determine their frequencies and consequences (loss-severities). 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. (2006), comprised of 7265 records, is used 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 (PV) (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 (intentional and illegal deposition of material on the ground), and others, with the EF and OE causes being the most significant. Herein, the unknown causes (U), dumping, and others are combined and referred to as others (O).

3.1. Prediction of incidents at chemical companies

Table 1 shows the number of incidents extracted from the NRC database for the seven companies. The total number of incidents, N_{total}, and the number of incidents of EF, N_{EF}, OE, 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 incidents of process units, N_{PU}, storage vessels, N_{SV}, compressors/pumps, $N_{\text{C/P}}$, heat-transfer equipment, N_{HT}, and transfer-line equipment, N_{TL}, are included. Note that the large excess of EF compared with the numbers of OE was unanticipated. Perhaps this is due to cost-saving measures that have reduced maintenance budgets, with major repairs postponed until they are deemed to be urgent. Also, because automated equipment often experiences fewer failures than those related to the inconsistencies of the operators, it is likely that many reported EF are indirectly a result of OE.

For each of the seven companies, the numbers of incidents were predicted for future years utilizing data from previous years. Included are the total number of incidents, N_{total}, the number of incidents associated with each equipment type, and the number of incidents associated with each cause. In the remainder of this section, selected results are presented and discussed.

Figs. 2(a) and (b) show the predictions of the number of incidents for companies B and F using *Poisson* distributions 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

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

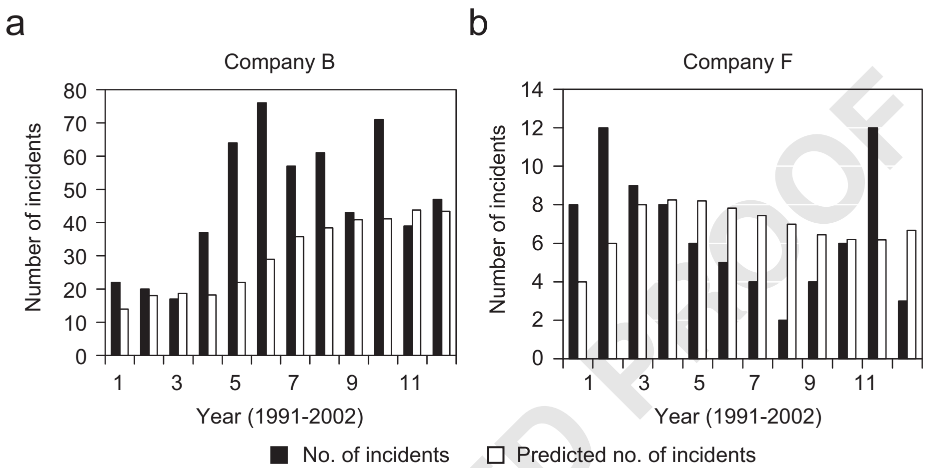

Fig. 2. Total number of incidents: (a) company B, (b) company F.

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 (Meel & Seider, 2006). 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 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 to 2002—although OSHA's PSM standard and EPA's RMP rule were introduced in 1992 and 1996, respectively. 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 that 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.

Next, the results of the Bayesian model checking using the *R* software package (Gentleman et al., 2005) to compute predictive distributions are presented in quantile–quantile (Q–Q) plots. For company F, Fig. 3(a) shows the density profile of incidents, while Fig. 3(b) shows the normal *Q–Q* plot, which compares the distribution of *z*

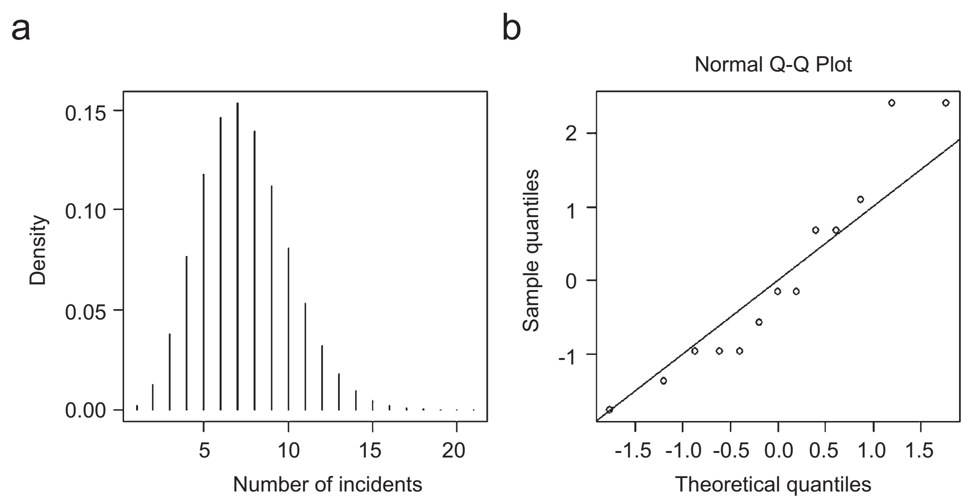

Fig. 3. Company F: (a) density of incidents, (b) Q–Q plot.

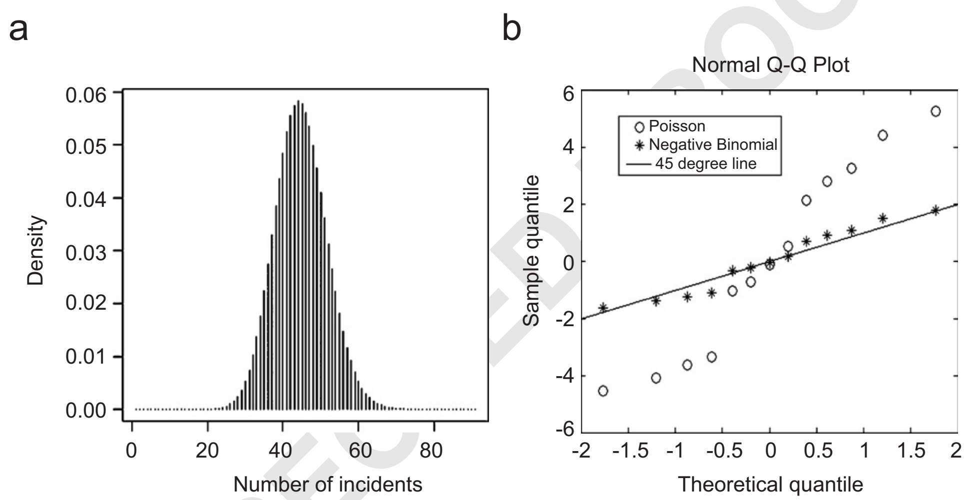

Fig. 4. Company B: (a) density of incidents, (b) Q–Q plot.

(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.

Figs. 4(a) and (b) show the density profile of incidents and the Q–Q plot for company B. Comparing Figs. 4(a) and 3(a), the number of incidents at company B is 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 and 65) than at company F (between ~0 and 15). Note that the circles on the Q–Q plot in Fig. 4(b) 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 higher than predicted, provide a warning.

The predictions in Fig. 4(b) are improved by using a *Negative Binomial* likelihood distribution with *Gamma* and *Beta* prior distributions. The prior distribution for 1990 is obtained using $\alpha = \beta = 0.001$, and $a = b = 1.0$, providing a relatively flat distribution in the region of interest; that is, a non-informative prior distribution. The *Negative Binomial* distribution provides better agreement for company B, while the *Poisson* distribution is preferred for company F.

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 incidents observed

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlp.2006.10.003

over the years, the performance of the *Gamma-Poisson* Bayesian models differ significantly. For company F, Figs. 5(a) and (b) show the *Q–Q* plots for EF and for OE, respectively. Fig. 5(a) shows better agreement with the model because the variation in the number of incidents related to EF is small, while the variation in the number of incidents related to OE is more significant. This is consistent with the expectation that equipment performance varies less significantly than operator performance over time.

Figs. 6(a) and (b) show the *Q–Q* plots for EF and for OE, respectively, at company B. When comparing Figs. 5(a) and 6(a), the predictions of the numbers of EF at company B are poorer than at company F using the *Poisson* distribution, but are improved using the *Negative Binomial* distribution. This is similar to the predictions for the total numbers of incidents at company B, as shown in Fig. 4(b), compared with those at company F, as shown in Fig. 3(b). Yet, the predictions for the OE 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 EF.

Figs. 7(a–d) show the *Q–Q* plots for incidents associated with the process units, storage vessels, heat-transfer equipment, and compressors/pumps at company B using *Poisson* and *Negative Binomial* distributions. The *Negative Binomial* distribution is better for incidents associated with the process units, compressors/pumps, and heat-transfer equipment, while the *Poisson* distribution is preferred for storage vessels.

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 Harris County obtained from the NRC database. However, no specific trends for a particular chemical associated with a higher number of

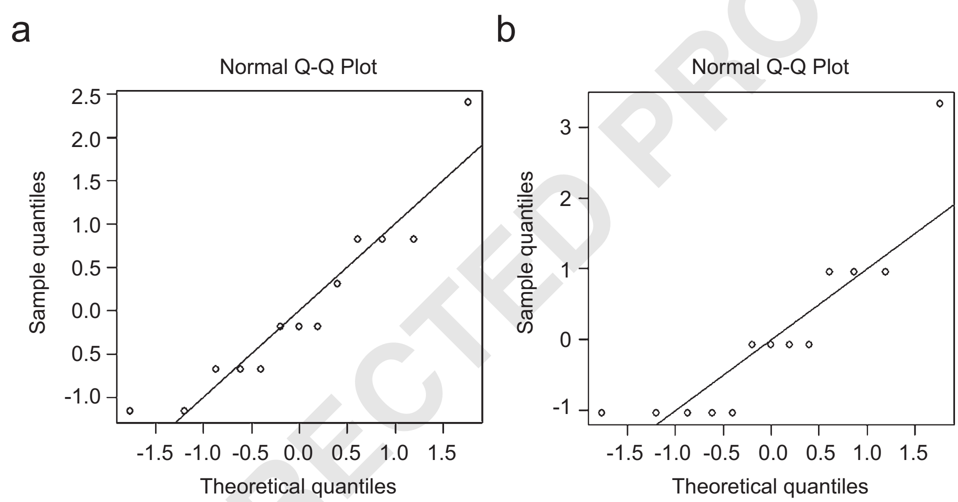

Fig. 5. Company F: (a) equipment failures, (b) operator errors.

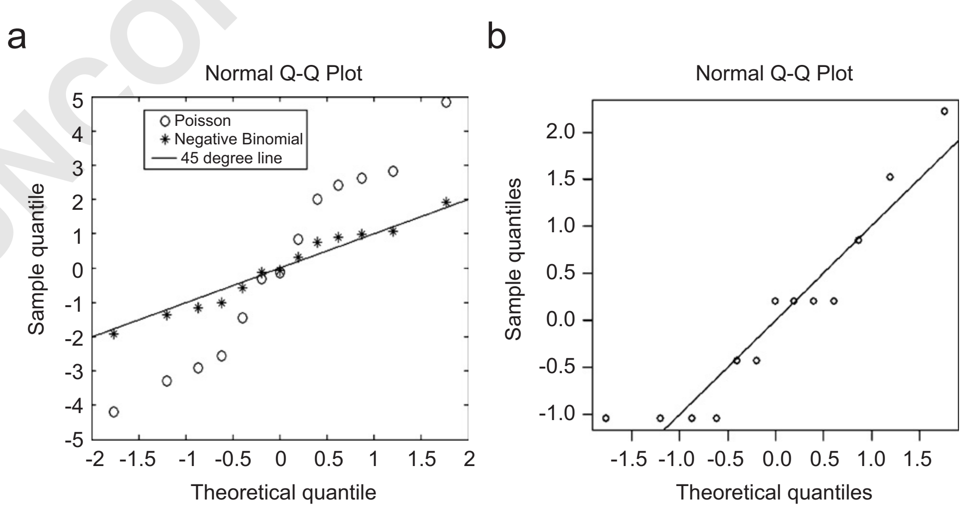

Fig. 6. Company B: (a) equipment failures, (b) operator errors.

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlp.2006.10.003

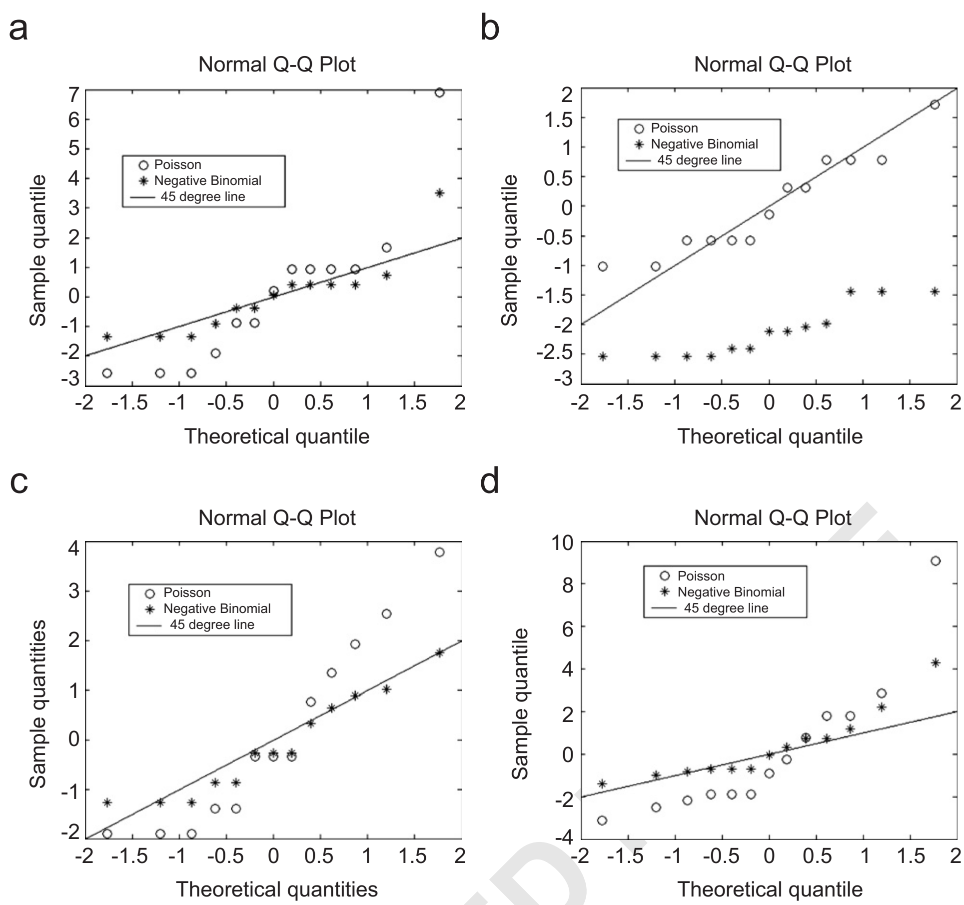

Fig. 7. Company B: (a) process units, (b) storage vessels, (c) Heat-transfer equipment, and (d) compressors/pumps.

incidents in all of the companies were observed. This could be because different products are produced in varying amounts by 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.

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 of the days of the week, with the mean, E, and variance, V, of z tabulated. 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 all of the companies, except company G. Lower deviations occur at the beginning of the week and over the weekends. To understand this observation, more information appears to be necessary; for example, (1) defining the operator shift and maintenance schedules, (2) carrying out operator surveys, (3) determining operator work loads, and (4) relating the data on the causes of the incidents to the days of the week, identifying more specific patterns. Furthermore, the higher means and variances for company G on Friday and Saturday suggest that additional data are needed to generate a reliable Bayesian model.

3.5. Rates of EF and OE

In this section, for an incident, the probabilities of the involvement of each of the 13 equipment types and the probabilities of their causes (EF, OE and O) are modeled. The tree in Fig. 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 numbers 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.

The prior distributions of the probability of x_i are modeled using *Beta* distributions with parameters a_i and 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

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

Entry in each cell–$E(z)$, $V(z)$

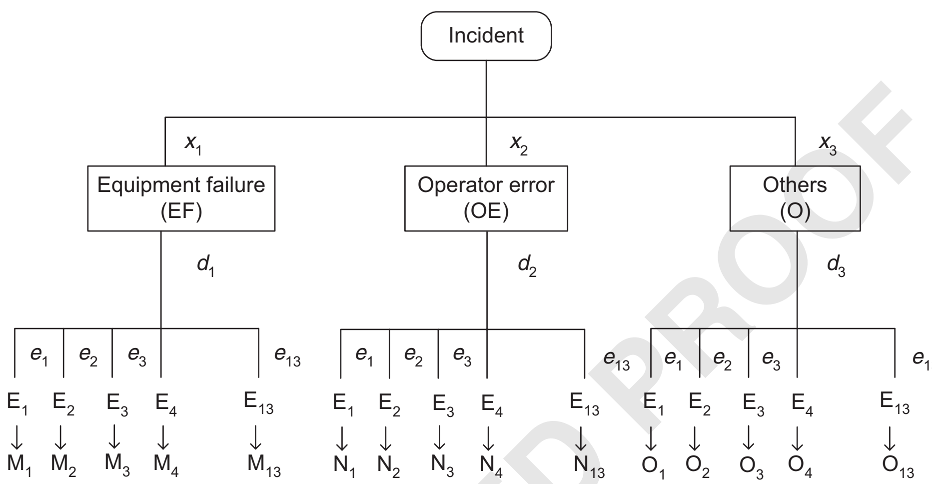

Fig. 8. Tree of causes and equipment types involved in an incident.

are updated using *Bernoulli's* likelihood distribution to obtain the posterior distribution of the probability of x_i:

$$f(x_i|Data) \propto (x_i)^{a_i-1+\mathrm{d}_i}(1-x_i)^{b_i-1+\sum_{k=1,\neq i}^{3}\mathrm{d}_k} f(x_i). \tag{4}$$

The posterior distributions, which are also *Beta* distributions having parameters, $a_i+\mathrm{d}_i$, and $b_i+\sum_{k=1,\neq i}^{3}\mathrm{d}_k$, change at the end of each year as d_i change. a_1 and b_1 are assumed to be 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 a nearly flat, 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, as shown in Fig. 9(a).

The posterior means and variances are obtained over the years 1990–2002 for each of the seven companies. Fig. 9(a) shows the probabilities, x_1, x_2, and x_3, 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 the EF, decrease from 0.499 for the OE, and increase from 0.001 for the others, with the OE approaching slightly higher values than those for the others.

Similarly, analyses for the probabilities of the equipment types, e_1, e_2, …, e_{13}, are carried out using *Beta* distributions, $f(e_i)$ and $f(e_i|Data)$, with the *Data*, $M_1+N_1+O_1$, $M_2+N_2+O_2$, $M_3+N_3+O_3$, …, $M_{13}+N_{13}+O_{13}$. The prior distributions of the probabilities of e_i are modeled using *Beta* distributions with parameters p_i and q_i:

$$f(e_i) \propto (e_i)^{p_i-1}(1-e_i)^{q_i-1}, \quad i=1,\ldots,13, \tag{5}$$

having means $= p_i/(p_i+q_i)$ and variances $= p_iq_i/(p_i+q_i)^2(p_i+q_i+1)$. These conjugate *Beta* prior distributions are updated using *Bernoulli's* likelihood distribution to obtain the posterior distributions of the probabilities of e_i:

$$f(e_i|Data) \propto (e_i)^{p_i-1+M_i+N_i+O_i} \times (1-e_i)^{q_i-1+\sum_{k=1,\neq i}^{13} M_k+N_k+O_k} f(e_i). \tag{6}$$

The posterior distributions, which are also *Beta* distributions having parameters, $p_i+M_i+N_i+O_i$, and $q_i+\sum_{k=1,\neq i}^{3}M_k+N_k+O_k$, change at the end of each year as $M_i+N_i+O_i$ change. The parameters, p_i and q_i, are chosen to give flat, non-informative, prior distributions.

The posterior means and variances are obtained over the years 1990–2002 for each of the thirteen equipment types at each of the seven companies. Fig. 9(b) shows, for an incident, that the probability of the involvement of the PV decreases over time. Similarly, the probabilities for the

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlp.2006.10.003

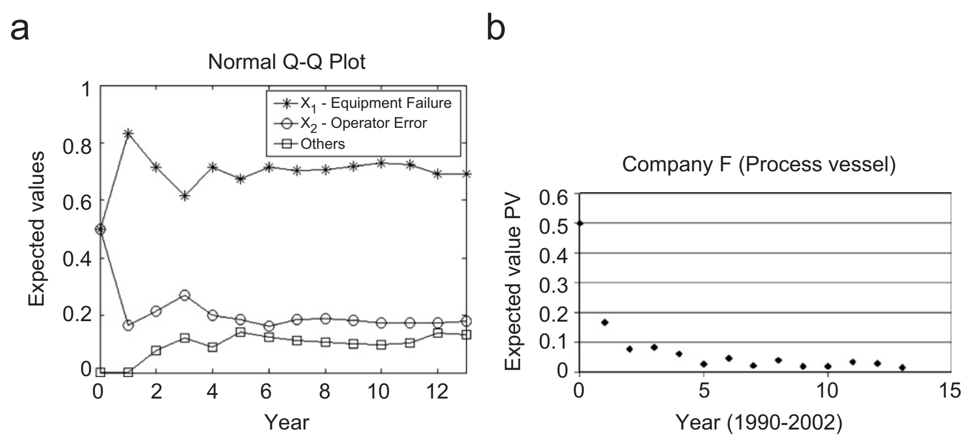

Fig. 9. Probabilities of x_i for company F: (a) EF, OE, and others, (b) PV.

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

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 EF and OE, insights regarding equipment and human reliabilities are obtained. In Table 3, where the range of the annual OE/EF ratio for all of the companies is shown, incidents involving EF exceed incidents involving OE. As mentioned in Section 3.1, the low OE/EF ratios are probably due to the operator bias when reporting incidents. Nevertheless, for petrochemical companies, the ratio is much lower than for specialty chemical companies. This is anticipated because the manufacture of specialty chemicals involves more batch operations, increasing the likelihood of OE.

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 (i.i.d.). 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 datum entries in each lumped data set is 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. Modeling the loss-severity distribution using EVT

For rare events with extreme losses, it is important to identify those that exceed a high threshold. EVT is a powerful and fairly robust framework to study the tail behavior of a distribution. Embrechts et al. (1997) provide an overview of EVT as a risk management tool, discussing its potential and limitations. In another study, McNeil (1997) examines the estimation of the tails of the *loss-severity* distributions and the estimation of quantile risk measures for financial time-series using EVT. Herein, EVT, which uses the *generalized Pareto* distribution (GPD), is employed to develop a *loss-severity* distribution for the seven chemical companies. Other methods use the *log-normal*, *generalized extreme value*, *Weibull*, and *Gamma* distributions.

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

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlp.2006.10.003

$$F_u(y) = Pr\{l - u \leqslant y | l > u\} = \frac{F(y+u) - F(u)}{1 - F(u)}, \quad l \in L, \tag{7}$$

which represents the probability that the value of l exceeds the threshold, u, by less than or equal to y, given that l exceeds the threshold, u, where F is the cumulative probability distribution, and L is the set of losses. This is the so-called *loss-severity* distribution. Note that, for the NRC database, l is defined in Section 3.7.1. For sufficiently high threshold, u, the distribution function of the excess may be approximated by the GPD, $G(l)$, and consequently, $F_u(y)$ converges to the GPD as the threshold becomes large. The GPD is

$$G(l) = \left\{ \begin{array}{ll} 1 - \left(1 + \xi \frac{l-u}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - e^{-l/\beta} & \text{if } \xi = 0 \end{array} \right\}, \tag{8}$$

where β is the scale parameter, ξ is the shape parameter, and the tail index is ξ^{-1}. Note that the GPD reduces to 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 the NRC database

Because few incidents have high severity levels, the incidents analyzed for the seven companies are assumed to be i.i.d. Consequently, the incidents for a specific company (internal data) are combined with those for the other companies (external data) to obtain a common *loss-severity* distribution for the seven companies. The loss associated with an incident, l, is calculated as a weighted sum of the numbers of evacuations, N_e; injuries, N_i; hospitalizations, N_h; fatalities, N_f; and damages, N_d:

$$l = w_e N_e + w_i N_i + w_h N_h + w_f N_f + w_d N_d, \tag{9}$$

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. Note the weighting factors were adjusted to align with the company performance histories.

For the NRC database, the threshold value, u, was chosen to be \$10,000. As expected, the NRC database has few incidents that have a sizable loss. Only 157 incidents among those reported had monetary loss ($l>0$), 64 exceeded the threshold, and 108 exceeded or equaled the threshold. A software package, Extreme Value Analysis in MATLAB (EVIM) Gencay et al. (2001), obtained the parameters of the GPD, $\xi = 0.8688$ and $\beta = 1.7183 \times 10^4$, for the NRC database using the maximum likelihood method. Fig. 10 shows the predictions of $F_u(y)$, the cumulative probability of the losses, l, that exceed or equal the threshold, u. Note that while the cumulative distribution of the losses could be improved with data from more

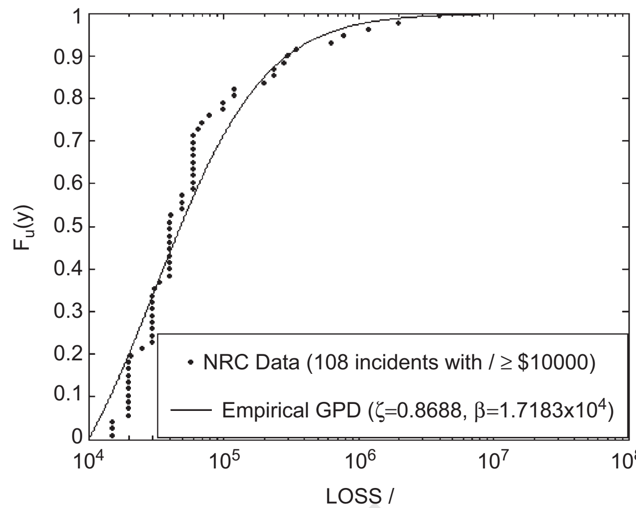

Fig. 10. *Loss-severity* distribution of the NRC database.

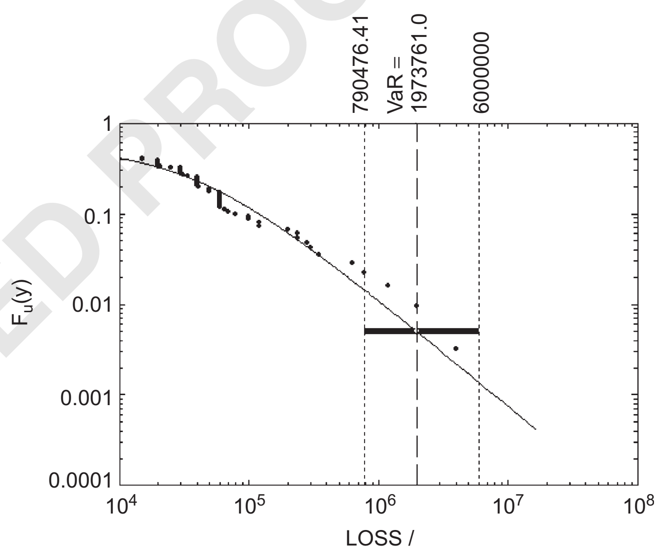

Fig. 11. Tail behavior of the *loss-severity* distribution for companies A–G.

companies in Harris County, the predictions in Fig. 10 are considered to be satisfactory.

By graphing $\log(1{-}F_u(y))$, Fig. 11 emphasizes the tail of the *loss-severity* distribution, with the value at risk (VaR) defined at 99.5% ($1{-}F_u(y) = 0.005$) cumulative probability equal to $\$1.97 \times 10^6$ and the lower and upper bounds on the 95% confidence interval equal to $\$7.9 \times 10^5$ and $\$6.0 \times 10^6$, respectively. The VaR is a forecast of a specified percentile (e.g., 99.5%), usually in the right tail, of the *loss-severity* distribution over some period (e.g., annually).

4. Operational risk

Several types of risks, for example, credit, market, and operational risks are encountered by chemical companies. In this work, the primary focus is on calculating the

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlp.2006.10.003

operational risk associated with a chemical company, which is defined as the risk of direct or indirect losses resulting from inadequate or failed internal resources, people, and systems, or from external events.

Capital charge (that is, CaR) of a company due to operational risk is calculated herein. Capital charge is obtained from the *total loss* distribution (to be defined below) using the VaR. Computation of the *total loss* distribution is a common statistical approach in the actuarial sciences. This paper applies this approach to risk analysis in the chemical industries. There are four methods for obtaining capital charge associated with operational risk: (i) the basic indicator approach (BIA), (ii) the standardized approach (SA), (iii) the internal measurement approach (IMA), and (iv) the loss distribution approach (LDA). The LDA (Klugman, Panjer, & Willmot, 1998) is considered to be the most sophisticated, and is used herein.

In the LDA, the annual frequency distribution of incidents is obtained using internal data, while the *loss-severity* distribution of an incident is obtained using internal and external data, as discussed in Section 3.7.1. By multiplying these two distributions, the *total loss* distribution is obtained.

Fig. 12 shows a schematic of the *total loss* distribution for a chemical company. The *expected* loss corresponds to the mean (expected) value and the *unexpected* loss is the value of the loss for a specified percentile (e.g., 99.5%) minus the *expected* loss. Note that, in some circles, the CaR is defined as the *unexpected* loss. However, herein, in agreement with other institutions, the CaR is the sum of the *expected* and *unexpected* losses, at the 99.5 percentile of the *total loss* distribution.

Highly accurate estimates of the CaR are difficult to compute due to the scarcity of internal data for the extreme events at most companies. Also, internal data are biased towards low-severity losses while external data are biased towards high-severity losses. Consequently, a mix of internal and external data is needed to enhance the statistical significance. Furthermore, it is important to balance the cost of recording very low-severity data and the truncation bias or accuracy loss resulting from the use of unduly high thresholds.

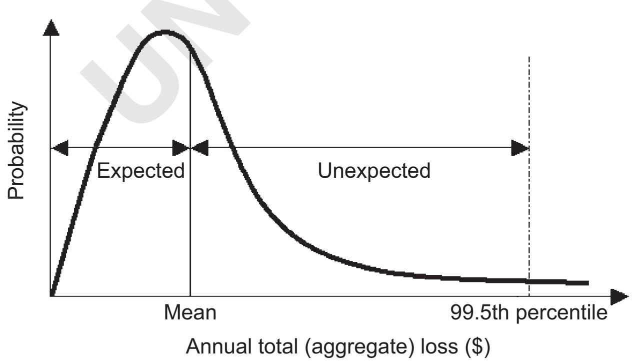

Fig. 12. Schematic of *total loss* distribution for a chemical company.

As when estimating the frequency of incidents (Section 2), a frequency distribution is obtained initially using Bayesian theory for events with losses that exceed a threshold, u. Because operational risks are difficult to estimate shortly after operations begin, conservative estimates of the parameters of the *Poisson* distribution may be obtained. In these cases, the sensitivity of the CaR to the frequency parameter should be examined. After the frequency distribution is obtained, it is multiplied with the *loss-severity* distribution and the FFT is used to calculate the *total loss* distribution.

4.1. FFT algorithm

The algorithm for computing the *total loss* distribution using the FFT is described in this section. Aggregate losses are represented as the sum, Z, of a random number, N, of individual losses, $l_1, l_2, \ldots, l_N$. The characteristic function of the total loss, $\phi_z(t)$, is:

$$\begin{aligned}\phi_z(t) &= E[\mathrm{e}^{it(Z)}] = E_N[E[\mathrm{e}^{it(l_1+l_2+\ldots+l_N)}|N]] \\ &= E_N[\phi_l(t)^N] = P_N(\phi_l(t)),\end{aligned} \tag{10}$$

where P_N is the probability generating function of the frequency of incidents, N, and ϕ_l is the characteristic function of the *loss-severity* distribution. The FFT produces an approximation of ϕ_z and, using ϕ_z, the inverse fast-Fourier transform (IFFT) gives $f_z(Z)$, the discrete probability distribution of the total (aggregate) loss. The details of the FFT, IFFT, and the characteristics function are found elsewhere (Klugman et al., 1998).

First, $n_p = 2^r$ for some integer r is chosen, where n_p is the desired number of points in the distribution of total losses, such that the *total loss* distribution has negligible probability outside the range $[0, n_p]$. Herein, $r = 13$ provides a sufficiently broad range. It can be adjusted according to the number of incidents in a company. The next steps in the algorithm are:

1. The *loss-severity* distribution is transformed from continuous to discrete using the method of rounding (Klugman et al., 1998). The span is assumed to be \$20,000 in line with the threshold for the GPD. The discrete loss-severity vector is represented as $f_l = [f_l(0), f_l(1), \ldots, f_l(n_p\text{-}1)]$.
2. The FFT of the discrete loss-severity vector is carried out to obtain the characteristic function of the *loss-severity* distribution: $\phi_l = \mathrm{FFT}(f_l)$.
3. The probability generating function of the frequency, $P_N(t) = e^{\lambda(t-1)}$, is applied, element-by-element, to the FFT of the discrete loss-severity vector to obtain the characteristic function of the *total loss* distribution: $\phi_z = P_N(\phi_l)$.
4. The IFFT is applied to ϕ_z to recover the discrete distribution of the total losses: $f_z = \mathrm{IFFT}(\phi_z)$.

Please cite this article as: Meel, A., et al. Operational risk assessment of chemical industries by exploiting accident databases. *Journal of Loss Prevention in the Process Industries*, (2006), doi:10.1016/j.jlp.2006.10.003

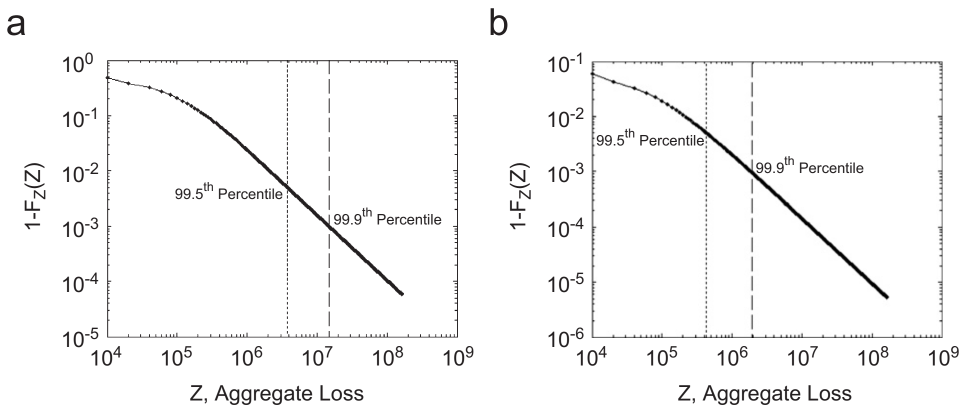

Fig. 13. *Total loss* distribution for: (a) company B, (b) company F.

4.2. Total loss distribution for companies B and F

The *Poisson* frequency parameters for companies B and F, obtained using internal data for each company, are $\lambda_B = 0.8461$ and $\lambda_F = 0.0769$. These are obtained using Bayesian theory for their incident data through the years 1 to $n-1$ (1990–2001) for incidents having losses that exceed or equal the threshold, \$10,000. The low λ_F indicates the low probability of incidents having significant losses in company F. For company B, λ_B indicates that about one event, with $l >$ \$10,000, is anticipated in the next year. Note that the *loss-severity* distributions in Figs. 10 and 11 are obtained using both internal and external data.

Fig. 13(a) shows the tail of the cumulative plot of the *total loss* distribution for company B. The total loss at the 99.5th percentile is \$$3.76 \times 10^6$ and at the 99.9th percentile is \$$14.1 \times 10^6$. When $\lambda_B \gg 1$, a much higher value of CaR is expected. Similarly, Fig. 13(b) shows the tail for company F. The total loss at the 99.5th percentile is \$$0.43 \times 10^6$ and at the 99.9th percentile is \$$1.78 \times 10^6$. As expected, the CaR for company F is lower than for company B by an order of magnitude.

Hence, this method provides plant-specific estimates of the CaR. Such calculations should be performed by chemical companies to provide better estimates for insurance premiums and to add quantitative support for safety audits.

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 higher variations in incidents are observed.
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 beneficial for obtaining the value at risk (VaR) from the *loss-severity* distribution using EVT and the capital at risk (CaR) from the *total loss* distribution.

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.

Acknowledgment

The interactions and advice of Professor Paul Kleindorfer of the Wharton Risk Management and Decision Center, Wharton School, University of Pennsylvania, and Professor Sam Mannan of the Mary Kay O'Connor Process Safety Center, Texas A&M University, are appreciated. Partial support for this research from the National Science Foundation through grant CTS-0553941 is gratefully acknowledged.

References

Anand, S., Keren, N., Tretter, M. J., Wang, Y., O'Connor, T. M., & Mannan, M. S. (2006). Harnessing data mining to explore incident databases. *Journal of Hazardous Material*, *130*, 33–41.

Baumont, G., Menage, F., Schneiter, J. R., Spurgin, A., & Vogel, A. (2000). Quantifying human and organizational factors in accident management using decision trees: The HORAAM method. *Reliability Engineering System Safety*, *70*(2), 113–124.

Bradlow, E. T., Hardie, B. G. S., & Fader, P. S. (2002). Bayesian inference for the negative binomial distribution via polynomial expansions. *Journal of Computational and Graphical Statistics*, *11*(1), 189–201.

CCPS (1995). Process Safety Incident Database (PSID). 〈http://www.aiche.org/CCPS/ActiveProjects/PSID/index.aspx〉.

Chung, P. W. H., & Jefferson, M. (1998). The integration of accident databases with computer tools in the chemical industry. *Computers and Chemical Engineering*, *22*, S729–S732.

Elliott, M. R., Wang, Y., Lowe, R. A., & Kleindorfer, P. R. (2004). Environmental justice: Frequency and severity of US chemical industry accidents and the socioeconomic status of surrounding communities. *Journal of Epidemiology and Community Health*, *58*(1), 24–30.

Embrechts, P., Kluppelberg, C., & Mikosch, T. (1997). *Modelling external events*. Berlin: Springer.

Gencay, R., Selcuk, F., & Ulugulyagci, A. (2001). EVIM: A software package for extreme value analysis in MATLAB. *Studies in Nonlinear Dynamics and Econometrics*, *5*(3), 213–239.

Gentleman, R., Ihaka, R., Bates, D., Chambers, J., Dalgaard, J., & Hornik, K. (2005). The R project for Statistical Computing. 〈http://www.r-project.org/〉.

Goossens, L. H. J., & Cooke, R. M. (1997). Applications of some risk assessment techniques: Formal expert judgement and accident sequence precursors. *Safety Science*, *26*(1–2), 35–47.

Kirchsteiger, C. (1997). Impact of accident precursors on risk estimates from accident databases. *Journal of Loss Prevention in the Process Industries*, *10*(3), 159–167.

Kleindorfer, P. R., Belke, J. C., Elliott, M. R., Lee, K., Lowe, R. A., & Feldman, H. I. (2003). Accident epidemiology and the US chemical industry: Accident history and worst-case data from RMP*Info. *Risk Analysis*, *23*(5), 865–881.

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (1998). *Loss Models: From data to decisions*. Wiley series in probability and statistics. Inc. John Wiley & Sons.

Mannan, M. S., O'Connor, T. M., & West, H. H. (1999). Accident history database: An opportunity. *Environmental Progress*, *18*(1), 1–6.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, *27*, 117–137.

Meel, A., & Seider, W. D. (2006). Plant-specific dynamic failure assessment using Bayesian theory. *Chemical Engineering Science*, *61*, 7036–7056.

NRC (1990). National Response Center. 〈http://www.nrc.uscg.mil/nrchp.html〉.

Phimister, J. R., Oktem, U., Kleindorfer, P. R., & Kunreuther, H. (2003). Near-miss incident management in the chemical process industry. *Risk Analysis*, *23*(3), 445–459.

Rasmussen, K. (1996). The experience with Major Accident Reporting System from 1984 to 1993. *European Commission, Joint Research Center, EUR 16341 EN*.

RMP (2000). 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*.

Robert, C. P. (2001). *The Bayesian choice*. New York: Springer-Verlag.

Sonnemans, P. J. M., & Korvers, P. M. W. (2006). Accidents in the chemical industry: Are they foreseeable? *Journal of Loss Prevention in the Process Industries*, *19*(1), 1–12.

Sonnemans, P. J. M., Korvers, P. M. W., Brombacher, A. C., van Beek, P. C., & Reinders, J. E. A. (2003). Accidents, often the result of an 'uncontrolled business process'—a study in the (Dutch) chemical industry. *Quality and Reliability Engineering International*, *19*(3), 183–196.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). Bayesian inference Using Gibbs Samping (BUGS). 〈http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml〉.

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

Uth, H. J., & Wiese, N. (2004). Central collecting and evaluating of major accidents and near-miss-events in the Federal Republic of Germany—results, experiences, perspectives. *Journal of Hazardous Materials*, *111*(1–3), 139–145.