Journal of Applied Sciences 9 (4): 731-737, 2009
ISSN 1812-5654

A New Accident Investigation Approach Based on Data Mining Techniques

[1]Sh. Parhizi, [1]J. Shahrabi and [2]M. Pariazar
[1]Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran
[2]Member of Young Research Club, Islamic Azad University, Iran

Abstract: In this study some data mining techniques for accident investigation and risk analysis is proposed. Function of most of accident investigation and risk analysis methodologies have been based on establishment of different scenarios of accident occurrence and simulation of accidents situation and so far no fundamental action for the analysis of remained data from accident has taken place. This study with the approach of data analysis and using different techniques of data mining can eliminate deficiencies of other techniques therewith covers theirs advantages. In this study factor analysis utilized to identify effective factors on occurrence of accidents. Cluster analysis utilized to classify accidents. A case study in a petrochemical company has been done in order to execute and investigate proposed methodology. The results show four different factors effecting on accident's occurrence and ten different clusters of accidents recognized. Also association rules exposed to discover all patterns and rules that cause occurrence of accidents.

Key words: Accident investigation, data mining, factor analysis, cluster analysis, association rules

INTRODUCTION

Iran geographic conditions and availability of huge crude oil reservoirs, Iran also has faced with creation and increasing development of oil, gas and petrochemical industries. As a direct consequence of ever-increasing variety of products and processes in these companies, the distances between such industries and human settlements have become lesser and lesser. Obviously growing these industries, the probability of occurring frightening and terrible accidents, treating life and financial capitals will increase. The damage potential of process industries is a cause of concern in many countries.

Strupczewski (2003) has made a comparison between nuclear power plant and oil, gas and other process industries. It was found that the accidents risk due to oil, gas and petroleum accidents are much larger than the risk due to nuclear-power accidents.

Reviewing published literature during years 1998 till 2007 clears that extended attempts for controlling effective factors on incidents of accidents in different countries of world's have been done. Some methodologies introduced and developed such as the worst case methodology that have been provided by Carter and Hirst (2000), PROTEUS provided by Stam *et al*. (2000), SCAP provided by Khan *et al*. (2002), PRISMA provided by Dye and Schaaf (2002), Haswim provided by Reniers *et al*. (2005) and finally MRRA methodology provided by Hu *et al*. (2007). More over some attempts for standardization of reporting and follow-up systems after accident taking places have been executed that can mention Berentsen and Holmboe (2004) method.

Function of most of accident investigation and risk analysis methodologies have been based on establishment of different scenarios of accident occurrence and simulation of accidents situation and so far no fundamental action for the analysis of remained data from accident has taken place, even though huge volume of data have been gathered and stored in data centers of HSE databases in petrochemical company. In methodologies used from 1998 till 2007, one case of using factor analysis technique for recognition of effective factors in incidents in a construction site by Sawacha (1998), one case of using cluster analysis for classification of cultural factors effective on accidents by Oltedal and Rundmo (2007), one case of using bayes theory by Trucco *et al*. (2008) and one case by Meel *et al*. (2007) have been seen. Although in recent years specially 2007, tendency for using some techniques of data mining such as bayes theory is obvious, but non utilization of data mining techniques and it's execution for risk analysis and accident investigation both for accident's pertinent to other sectors such as nuclear industries, chemical industries, food industries, public transportation and air industry is intensely felt.

The goal of this study is a desired reduction of this gap. This research proposed some data mining techniques for accident investigation and risk analysis. Factor analysis utilized to identify effective factors on occurrence of accidents. Cluster analysis utilized to classify accidents. Also, association rules exposed to discover all patterns and rules that cause occurrence of accidents. A case study in a petrochemical company has been done in order to execute and investigate proposed

Corresponding Author: Shaghayegh Parhizi, No. 99, Akbary Street, Yoosef Abaad, Tehran, Iran Tel: +0098-9125456653

techniques. Results show four different factors effecting on accident's occurrence and ten different clusters of accidents recognized and discovered patterns and rules have been considered.

MATERIALS AND METHODS

The structure of this study is as shown in Fig. 1. Four steps are involved in the methodology. The first step includes identifying effective criteria on accident occurrence. The second step involves reducing effective criteria by using factor analysis. At the third step cluster analysis was utilized for accident clustering. Finally fourth step presents discovering patterns and rules by using association rules algorithm.

A short description of using techniques

Data mining: Also a non-sensitive and non-infectious skin problem that does not affect other body organs. An individual with psoriasis usually has patches of red areas with clear boundaries and layers of data mining is an important part of information management technology. Simply put, it is a method to extract and analyze meaningful patterns and correlations in a large relational database (Frawley *et al*., 1991).

Factor analysis: Factor analysis is a general term that is given to a group of statistical multivariate methods which their primary goal is to define covert structure in data. In general terms by defining a collection of joint covert dimensions that are referred to as factors, it analysis relations structure (correlation) amongst great volume of variables (Lattin *et al*., 2003).

Cluster analysis: Cluster analysis is the name of a group of multivariate techniques that their primary goal is to classify things based on their specifications. Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters. Similarity can be expressed as functions specified by the users or experts. Good clusters show high similarity within a group and low similarity between any two different groups (Lattin *et al*., 2003).

Association rules: The association algorithm is a basic data mining algorithm and nothing more than a correlation counting engine. This method discovers all possible and interesting patterns in a database. Association rules can be used to discover the relationships and potential associations, of attributes among huge amounts of data. These rules can be effective in uncovering unknown relationships and provide results that can be the basis for forecasting and decision making (Tsay and Chiang, 2005; Chae *et al*., 2001).

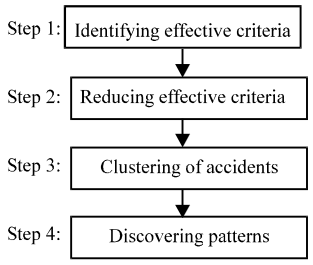

Fig. 1: Structure of methodology

Case study: According to the literature reviewed between 1998 till 2007, although using of data mining techniques for risk analysis increased nowadays, but most of these methodologies focus in limited aspects or study some cases. The methodology represent in this study try to propose a combined techniques for risk analysis. By using this methodology, researchers able to present a comprehensive view of accidents. For executing and investigating the proposed methodology, a case study in an Iranian petroleum company has been applied. The database contains 275 records of accidents that were registered in HSE department of a petroleum company for the period 1993-2007.

The structure of database is as shown in Table 1.

Step 1: Identifying effective criteria: This research commenced by reviewing the relevant literature on accident investigation and analysis published. This was followed by exploratory interviews which took place with operatives, managers and safety officer in a petroleum company.

Contributing factors to accident occurrence are shown in Table 2.

Step 2: Reducing effective criteria: After the exploratory interviews, as a pilot study questionnaire was designed and discussed with 100 personal. The response to each attitudinal question was measured on a five- point, strongly agree, agree, neither agree nor disagree, disagree, strongly disagree. The data was then analyzed utilizing the statistical computing package SPSS. As a result, the factor analysis technique was applied to reduce the large amount of criteria to a small number of factors.

The factor analysis technique was utilized to help identifying the underlying cluster of factors that dominate

Table 1: Database structure

Immediate effect	Report profile
Human deaths	Accident code
Human injuries	Date
Ecological harm	Accident type
National heritage loss	Release
Material loss	Water contamination
Community disruption	Fire
Emergency measures taken	Explosion
On-site services	Substance directly involve
External services	Toxic
Sheltering	Exotoxic
Evacuation	Flammable
Decontamination	Explosive
Restoration	Immediate source of accident
Suspected cause	Storage
Historical factors	Process
Environmental factors	Transfer
Operative factors	Transport
Organization and equipment factors	

Table 2: Different criteria

Conflict	Knowledge and experience
Safety and quality culture	Physical access
Problem solving style	Alertness
Self confidence	Pressure
Visual and audio effect	Procedure availability
Memory of recent diagnose action and result	Tool availability
Human-system interface	Skill
Work process designs	Tool adequacy and quality
Attitude	Procedure adequacy and quality
Harsh environment	Time constrain load

Table 3: Results of factor analysis before rotation

Criteria	Factor 1	Factor 2	Factor 3	Factor 4
Knowledge and experience	-0.5231	0.1581	-0.2224	0.0001
Physical access	0.1592	0.4702	0.0973	0.0758
Alertness	0.1683	0.0222	-0.7250	-0.2361
Pressure	0.0504	-0.0030	0.6859	0.0122
Procedure availability	-0.3252	-0.0010	-0.0788	0.7973
Tool availability	0.0222	-0.1809	0.0305	0.7207
Skill	0.7213	-0.2467	0.3360	0.0534
Tool adequacy and quality	-0.0788	0.2136	-0.0980	0.5062
Procedure adequacy and quality	0.1223	0.1235	0.3245	0.5564
Time constrain load	0.1508	0.2721	0.4890	0.1002
Conflict	-0.0515	-0.1394	-0.5062	-0.3033
Safety and quality culture	0.0758	-0.2018	-0.0328	0.7324
Problem solving style	0.15761	0.21609	0.4648	-0.3329
Self confidence	0.14685	-0.1493	0.5632	0.2411
Visual and audio effect	0.07881	0.5062	0.0204	0.1484
Memory of recent diagnose action and result	-0.4834	0.0234	-0.2128	0.1398
Human-system interface	0.0701	0.2023	-0.3215	0.5170
Work process designs	0.0312	-0.3510	0.0522	0.7280
Attitude	0.0120	-0.0215	0.4482	0.0543
Harsh environment	0.0341	-0.4702	-0.2897	-0.1258

safety performance. The factor analysis on the 20 accident factors and Table 3 shows the factor matrix.

In some criteria, elements of the factor loading matrix for several factors would be close to one another and allocation of those criteria to a special factor would cause loss of some part of data. In this regard, in order to deducing better and better specifying the quantity of factors, factors were rotated based on varimax method.

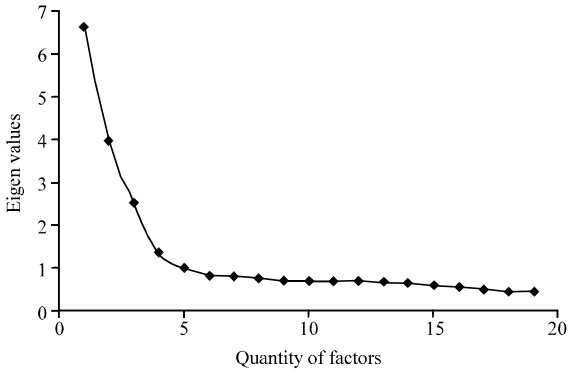

Fig. 2: Screen plot

Table 4: Results of factor after rotation

Criteria	Factor 1	Factor 2	Factor 3	Factor 4
Knowledge and experience	-0.6231	0.0481	-0.0224	0.0001
Physical access	0.1212	0.8702	0.0973	0.0758
Alertness	0.1123	0.0222	-0.8925	-0.0354
Pressure	0.0504	-0.0030	0.7859	0.0122
Procedure availability	-0.0252	-0.0010	-0.0788	0.8973
Tool availability	0.0222	-0.0809	0.0305	0.9207
Skill	0.7213	-0.0021	0.1245	0.0534
Tool adequacy and quality	-0.0788	0.0136	-0.0980	0.7062
Procedure adequacy and quality	0.0123	0.1005	0.0245	0.6698
Time constrain load	0.1002	0.1721	0.6890	0.0002
Conflict	-0.0515	-0.1394	-0.6845	-0.0033
Safety and quality culture	0.0758	-0.0018	-0.0328	0.7324
Problem solving style	0.05061	0.01609	0.8648	-0.0029
Self confidence	0.1205	-0.1493	0.6987	0.0543
Visual and audio effect	0.07881	0.8062	0.0204	0.0148
Memory of recent diagnose action and result	-0.6834	0.0234	-0.0128	0.0001
Human-system interface	0.0701	0.0056	-0.1321	0.8170
Work process designs	0.0312	-0.3510	0.0522	0.8280
Attitude	0.0120	-0.0215	0.7482	0.0543
Harsh environment	0.0341	-0.6702	-0.0897	-0.1258

After factor rotation, allocation of criteria to factor is executed with higher accuracy and less error. We can also understand the number of factor by screen plot (Table 4). This diagram shows number of factors in comparison with eigenvalues. As it is seen in the Fig. 2, elbow point is placed at factor 4.

It is necessary to assign an identifiable name to the group of factors of high correlation coefficients. The researchers named the groups of factors as historical factors, environmental factors, operator behavior, organization and equipment factors. Allocated criteria to factors are shown in Table 5.

Step 3: Accident clustering by using cluster analysis: For attaining the best answer, 15 different structures including 2 step models (cluster-association rules) with consideration of input variables and different prediction variable were devised. Prediction variable with letter P and input variable with letter I has been shown as seen in Table 6.

Table 5: Allocated criteria to factors

Criteria	Historical	Environmental	Operator behavior	Organization
Knowledge and experience	-0.6231			
Skill	0.7213			
Memory of recent diagnose action and result	-0.6834			
Physical access		0.8702		
Visual and audio effect		0.8062		
Harsh environment		-0.6702		
Alertness			-0.8925	
Pressure			0.7859	
Time constrain load			0.6890	
Conflict			-0.6845	
Problem solving style			0.8648	
Self confidence			0.6987	
Attitude			0.7482	
Procedure availability				0.8973
Tool availability				0.9207
Tool adequacy and quality				0.7062
Procedure adequacy and quality				0.6698
Safety and quality culture				0.7324
Human-system interface				0.8170
Work process designs				0.8280

Table 6: Data mining structure with various predict and input variables

Model name	Accident type	Immediate source of accident	Substance directly involve	Suspected cause	Emergency measures taken	Immediate effects
1	P	I	I	I	I	I
2	I	P	I	I	I	I
3	I	I	P	I	I	I
4	I	I	I	P	I	I
5	I	I	I	I	P	I
6	I	I	I	I	I	P
7	P	I				
8	P		I			
9	P			I		
10	P					I
11		I	I		P	
12	I		I			P
13	I	I	I	P		
14			P			I
15				I		P

From amongst below considered structures, as it follows we will focus of analyzing the best made models in cluster and association algorithms. SQL SERVER 2005 is used to execute proposed methodology.

In this step, the accidents were clustered by using cluster analysis. Two hundred records of accident had been used to build training model. Performing cluster analysis in first structure generates 10 clusters. In a cluster analysis, each object is assigned to precisely one of a set of clusters. The cluster diagram view of SQL SERVER 2005 presents each cluster as a single node. These nodes are scattered across a field and allowed to group based on similarities. It is attention worthy that in cluster analysis; the suitable answer is an answer that samples in a cluster are similar furthest possible and samples belonging to different clusters are separate from each other and not similar. There more thickness of cluster connection lines, the more strengths between clusters. As

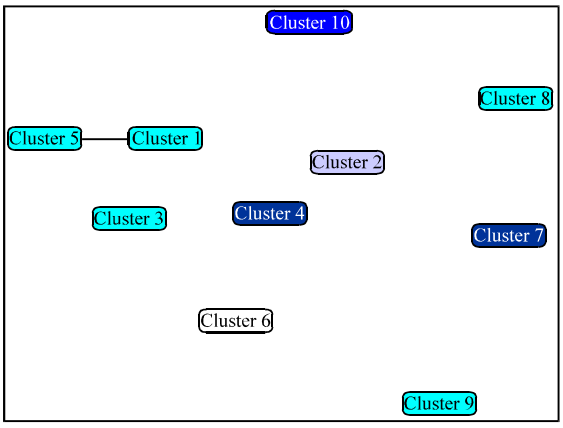

Fig. 3: Similarities between different clusters

it is observed in Fig. 3, only between cluster No. 1 and 5, there is weak connection.

Table 7: Cluster profiles

Cluster name	Cluster population	Accident type	Immediate source of accident	Substance directly involve	Suspected cause	Emergency measures taken	Immediate effects
1	50	Explosion and fire	Process	Flammable and explosive	Equipment	Internal and external services	Human injuries and material loss
2	40	Release	Process	Toxic	Equipment	Internal and external services	Material loss
3	20	Release	Process	Toxic and flammable	Equipment and standards	Internal and external services and sheltering and evacuation	Human injuries and material loss and community disruption
4	20	Release	Process		Economic and organization	Internal and external services	
5	15	Fire	Process	Flammable	Organization and team related factors and operator and standard	Internal services	Material loss
6	15	Release	Process	Toxic	Equipment and standards	Internal services	
7	10	Release	Storage	Toxic	Operator	Internal services	
8	10	Release	Transport	Toxic	Team related factors and operator	Internal services	
9	10	Release and explosion	Process and storage	Flammable and toxic and explosive	Standards and operator	Internal and external services	Material loss and community disruption
10	10	Water contamination and release	Storage	Toxic	Organization and environmental		National heritage

In Table 7, specifications of accidents assign to each clusters and its population is shown.

Seventy records of accident have been utilized to investigate the accuracy of clustering model and the model accuracy chart is shown in Fig. 4.

The blue line shows ideal model. An ideal model is a theoretical model that predicts the result correctly 100% of the time and red line is actual model can compare against the results of the ideal model to estimate accuracy of model. Maximum error of actual model as compared with ideal model is 6%.

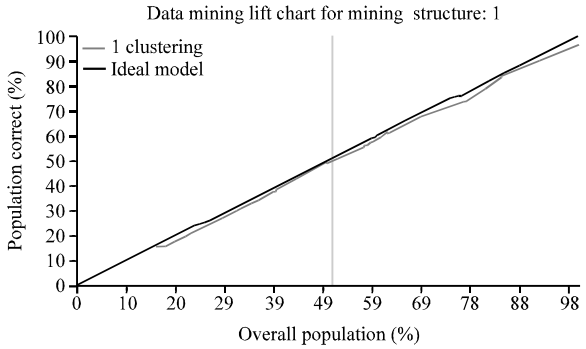

Fig. 4: Accuracy chart of cluster analysis

Step 4: Discovering patterns by using association rules: Association rules show existing patterns in data without considering any special goal. Due to this reason, this algorithm is examples of undirected data mining. Considered structure for analyzing association rules is the model related to structure 12 that from precision point of view, has more accuracy in comparison with other models. Structure of association rules shows probability of incidence of rules and importance of its occurrence. In this research, probability of occurrence 1 and minimum importance has been considered as 0. Below refers to most important rules discovered in above model:

- Incase substance directly involved in accident are not toxic and flammable, at the occurrence of accident, national heritage doesn't occurrence
- Incase occurrence of accident is in transfer or storage, occurrence of accident doesn't culminate in human death, material loss and community disruption
- Incidence of accident in the transfer stage without fire and explosion does not culminate in cutting off of community disruption
- In case contamination by toxic takes place in storage, doesn't culminate in human death
- Incase involve materials in accident are toxic but explosion does not take place, accident does not culminate in human death

Dependency network diagram pertinent to structure 12 association rules has been shown in Fig. 5.

Seventy five records of accident have been utilized to investigate the accuracy of association rules model and the model accuracy chart is shown in Fig. 6.

Maximum error of actual model as compared with ideal model is 14%.

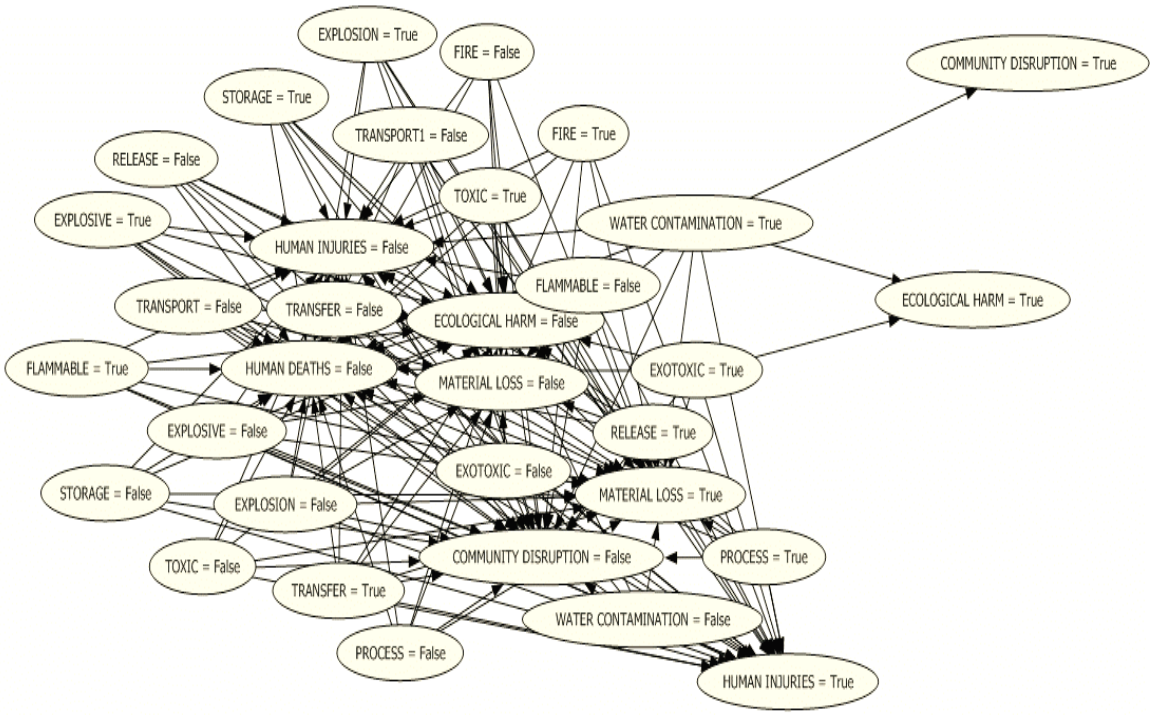

Fig. 5: Dependency network diagram of association rules model

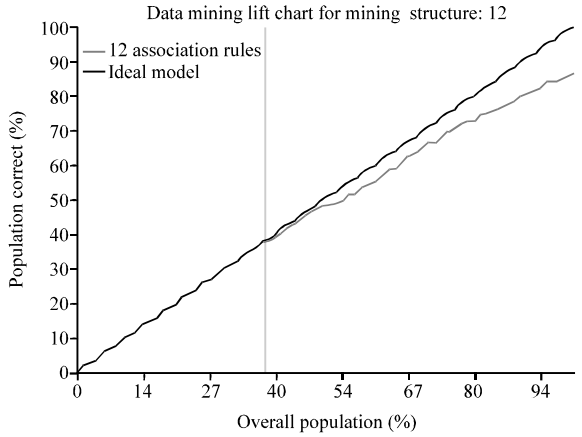

Fig. 6: Accuracy chart of association rules

CONCLUSION

The aim of this study was to show a new application of data mining techniques in the field of accident investigation and risk analysis where, they have never been used and to show results and benefits. This research commenced by reviewing the relevant literature on accident investigation and analysis published. This was followed by exploratory interviews which took place with operatives, managers and safety officer in a petroleum company. In this case, 20 different effective factors have been found. Then factor analysis utilized to reducing effective factors on occurrence of accidents. Therefore as a pilot study questionnaire was designed and discussed with 100 personal. Four different factors recognized and labeled. A case study in a petrochemical company has been done in order to execute and investigate proposed techniques. Cluster analysis utilized to classify accidents. Results show 10 different clusters of accidents recognized. Also association rules exposed to discover all patterns and rules that cause occurrence of accidents.

REFERENCES

Berentsen, R. and R.H. Holmboe, 2004. Incidents/ accidents classification and reporting in Statoil. J. Hazardous Mater., 111: 155-159.

Carter, D.A. and I.L. Hirst, 2000. Worst case methodology for the initial assessment of societal risk from proposed major accident installations. J. Hazardous Mater., 71: 117-128.

Chae, Y.M., S.H. Ho, K.W. Cho, D.H. Lee and S.H. Ji, 2001. Data mining approach to policy analysis in a health insurance domain. Int. J. Med. Inform., 62: 103-111.

Dye, J. and T.V.D. Schaaf, 2002. PRISMA as a quality tool for promoting customer satisfaction in the telecommunications industry. Reliab. Eng. Syst. Safety, 75: 303-311.

Frawley, W.J., G. Paitetsky-Shapiro and C.J. Matheus, 1991. Knowledge Discovery in Databases: 1st Edn., AAAI/MIT Press, California.

Hu, S., F. Quangen, X. Haibo and X. Yongtao, 2007. Formal safety assessment based on relative risks model in ship navigation. Reliab. Eng. Syst. Safety, 92: 369-377.

Khan, F.I., H. Tahir and S.A. Abbasi, 2002. Design and evaluation of safety measures using a newly proposed methodology SCAP. J. Loss Prevent. Proc. Ind., 15: 129-146.

Lattin, J., D. Carroll and P. Green, 2003. Analyzing Multivariate Data. 3rd Edn., Brooks Cole, USA., ISBN: 0534349749, pp: 240.

Meel, A., L.M. O'Neill, J.H. Levin, W.D. Seider, U. Oktem and N. Keren, 2007. Operational risk assessment of chemical industries by exploiting accident databases. J. Loss Prevent. Proc. Ind., 20: 113-127.

Oltedal, S. and T. Rundmo, 2007. Using cluster analysis to test the cultural theory of risk perception. Trans. Res. Part F, 10: 254-262.

Reniers, G.L.L., W. Dullaert, J.M. Aleb and K. Soudan, 2005. Developing an external domino accident prevention framework: Hazwim. J. Loss Prevent. Proc. Ind., 18: 127-138.

Sawacha, E., 1998. Factors affecting safety performance on construction sites. Int. J. Project Manage., 17: 309-315.

Stam, G.J., P.H. Bottelberghs, J.G. Post and H.G. Bos, 2000. PROTEUS, a technical and management model for aquatic risk assessment of industrial spills. J. Hazardous Mater., 71: 439-448.

Strupczewski, A., 2003. Accident risks in nuclear-power plants. Applied Energy, 75: 79-86.

Trucco, P., E. Cango, F. Ruggeri and O. Grande, 2008. A bayesian belief networkmodeling of organizational factors in risk analysis: A case study in maritime transportation. Reliab. Eng. Syst. Safety, 93: 845-856.

Tsay, Y.J. and J.Y. Chiang, 2005. CBAR: An efficient method for mining association rules. Knowledge-Based Syst., 18: 99-105.