Pattern extraction for high-risk accidents in the construction industry: a data-mining approach

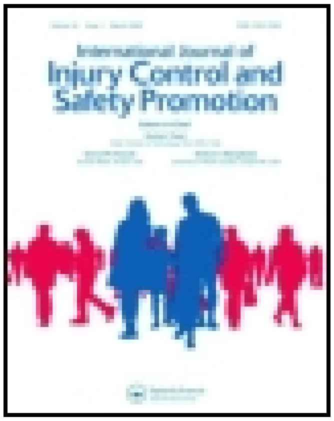
Article *in* International Journal of Injury Control and Safety Promotion · September 2016

DOI: 10.1080/17457300.2015.1032979

4 authors, including:

Mehran Amiri
Amirkabir University of Technology

Elahe Soltanaghaei
University of Virginia



International Journal of Injury Control and Safety Promotion

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nics20

Pattern extraction for high-risk accidents in the construction industry: a data-mining approach

Mehran Amiri[a], Abdollah Ardeshir[a], Mohammad Hossein, Fazel Zarandi[b] & Elahe Soltanaghaei[c]

[a] Civil and Environmental Engineering Department, Amirkabir University of Technology, Tehran, Iran

[b] Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran

[c] Computer Engineering Department, Sharif University of Technology, Tehran, Iran

Published online: 21 May 2015.

International Journal of Injury Control and Safety Promotion, 2015
http://dx.doi.org/10.1080/17457300.2015.1032979

Pattern extraction for high-risk accidents in the construction industry: a data-mining approach

Mehran Amiri[a], Abdollah Ardeshir[a]*, Mohammad Hossein Fazel Zarandi[b] and Elahe Soltanaghaei[c]

[a]*Civil and Environmental Engineering Department, Amirkabir University of Technology, Tehran, Iran;* [b]*Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran;* [c]*Computer Engineering Department, Sharif University of Technology, Tehran, Iran*

(*Received 9 January 2014; accepted 10 December 2014*)

Accidents involving falls and falling objects (group I) are highly frequent accidents in the construction industry. While being hit by a vehicle, electric shock, collapse in the excavation and fire or explosion accidents (group II) are much less frequent, they make up a considerable proportion of severe accidents. In this study, multiple-correspondence analysis, decision tree, ensembles of decision tree and association rules methods are employed to analyse a database of construction accidents throughout Iran between 2007 and 2011. The findings indicate that in group I, there is a significant correspondence among these variables: time of accident, place of accident, body part affected, final consequence of accident and lost workdays. Moreover, the frequency of accidents in the night shift is less than others, and the frequency of injury to the head, back, spine and limbs are more. In group II, the variables time of accident and body part affected are mostly related and the frequency of accidents among married and older workers is more than single and young workers. There was a higher frequency in the evening, night shifts and weekends. The results of this study are totally in line with the previous research.

Keywords: pattern extraction; construction safety; high-risk accidents; multiple correspondence analysis; ensembles of decision tree; association rules

1. Introduction

Occupational accidents are the cause of more than 300,000 mortalities and 300 million injuries around the world each year (International Labour Organization, 2013). This considerable number of cases has led to severe human and financial impacts in different countries (Warch, 2002). Previous studies have shown that workers in various industries are vulnerable to occupational accidents in different ways (Dudarev, Karnachev, & Odland, 2013). Construction is known as one of the most dangerous industries all over the world (Cheng, Leu, Lin, & Fan, 2010).

Occupational safety in the construction industry is studied in different countries and regions around the world (such as Halvani, Jafarinodoushan, Mirmohammadi, & Mehrparvar, 2012; Lin, Chen, & Wang, 2011; López Arquillos, Rubio Romero, & Gibb, 2012; etc.). Here, seeking for accident occurrence patterns in different types of accidents can help involved parties significantly in providing appropriate preventive strategies.

1.1. Data-mining and its application in analysing occupational accident data

It is several years that data-mining techniques have been used to analyse data in various fields (Chang & Wang, 2006). However, a review of the literature shows that these methods have been employed in the occupational safety analysis (particularly occupational accident data) on a limited basis (Bevilacqua, Ciarapica, & Giacchetta, 2008; Parhizi, Shahrabi, & Pariazar, 2009; Persona, Battini, Faccio, Bevilacqua, & Ciarapica, 2006). Few authors have used data-mining techniques in the study of occupational accidents in the construction industry.

Liao and Perng (2008) identified the characteristics of work-related injuries in the construction industry of Taiwan between 1999 and 2004 with the help of the association rules technique (Liao & Perng, 2008).

Liao, Perng, and Chiang (2009) introduced a new measure, namely ‘extracted probability’ in order to improve the effectiveness of the association rules method. Then, this method was applied in the analysis of 1062 fatal occupational injury cases in the construction industry

*Corresponding author. Email: ardeshir@aut.ac.ir

of Taiwan which had occurred between 2000 and 2005 (Liao et al., 2009).

Cheng, Lin, and Leu (2010) employed the association rules technique in the analysis of 1347 accident cases in Taiwan's construction industry, which had occurred between 2000 and 2007 (Cheng et al., 2010).

1.2. *High-risk occupational accidents in the construction industry of Iran*

According to statistics provided by the Iranian Social Security Organization (ISSO), falling object accidents and falls constitute a total of 44 per cent of construction accidents recorded between 2007 and 2011 in the database of the organization. On the other hand, being hit by a vehicle, electric shock, collapse in the excavation and fire or explosion accidents, while being only 7% of the total records, form about 26% of all fatal and totally disabling cases (Social Security Organization of the Islamic Republic of Iran, 2012). A recent study revealed that these six types of accidents are the most risky accidents in the construction industry of Iran (Amiri, Ardeshir, & Fazel Zarandi, 2014). Accidents in group I have high occurrence probability and moderate consequences and accidents in group II have low frequency and severe outcomes. Description of these two groups of accidents is presented in Table 1.

Hence, the aim of this study is to investigate the characteristics of these two groups of accidents using decision tree and association rules methods. Furthermore, it uses ensembles of trees to provide more efficiency and accuracy. In addition, this paper discusses the obtained results in order to identify potential hazards in the construction industry. Based on the literature review performed by the authors, it seems that this study is the first application of data-mining methods on the occupational accident data of Iran. Moreover, the ensembles method has not been applied on occupational accident databases earlier. The findings of this study can assist policy-makers and involved parties in pointing out hazardous conditions and setting preventive measures and strategies.

Table 1. Description of the groups I and II of accidents.

	Group I	Group II
Types of accidents	Falling objects Falls or slips	Hit by vehicle Electric shock Collapse in the excavations Fire or explosion
Total number of cases	9671	1024
Frequency	High	Low
Severity	Moderate	High

2. Materials and methods

The characteristics of the accident database used in this study and the methodology of this research are explained in the following subsections.

2.1. *Accident data*

Since 1975, the ISSO must be notified of all occupational accidents causing injury to the insured workers. In this study, ISSO provided the anonymous data of all occupational accidents among Iranian insured construction workers during the period of 2007–2011. A total number of 25,057 cases were supplied, but after performing data cleansing and preprocessing steps, 21,864 cases were accepted for the analysis. This sample included 4158, 4528, 4059, 4270 and 4849 annual cases for each year from 2007 to 2011, respectively.

Few number of characteristics of insured construction workers (including age and sex) have also been archived by ISSO. Table 2 presents some data of insured construction workers of the year 2011 ($N = 312{,}492$).

2.2. *Data analysis*

When the data-set of construction industry accidents is obtained, the data analysis procedure follows in three steps: (1) data cleansing, (2) preprocessing and (3) data mining (data analysis) (as shown in Figure 1).

In the first and second steps, the Rapid Miner software was employed to perform the data cleansing and preprocessing procedures. The data cleansing method includes removing missing values and duplicate records. Then, in the preprocessing step some appropriate features are generated and the numerical features are discretized. After that, a list of proper features are selected.

Once the features were prepared according to the preprocessing techniques, a descriptive analysis of the variables was conducted to present the frequency distribution

Table 2. Some characteristics of insured construction workers of the year 2011.

Variables and categories	Frequency *n* (%)
Gender of the insured worker	
Male	261,063 (84)
Female	51,429 (16)
Age of the insured worker (years)	
≤19	2389 (1)
20–35	174,007 (55)
36–45	74,154 (24)
46–55	43,026 (14)
≥56	18,916 (6)
In total	312,492 (100)

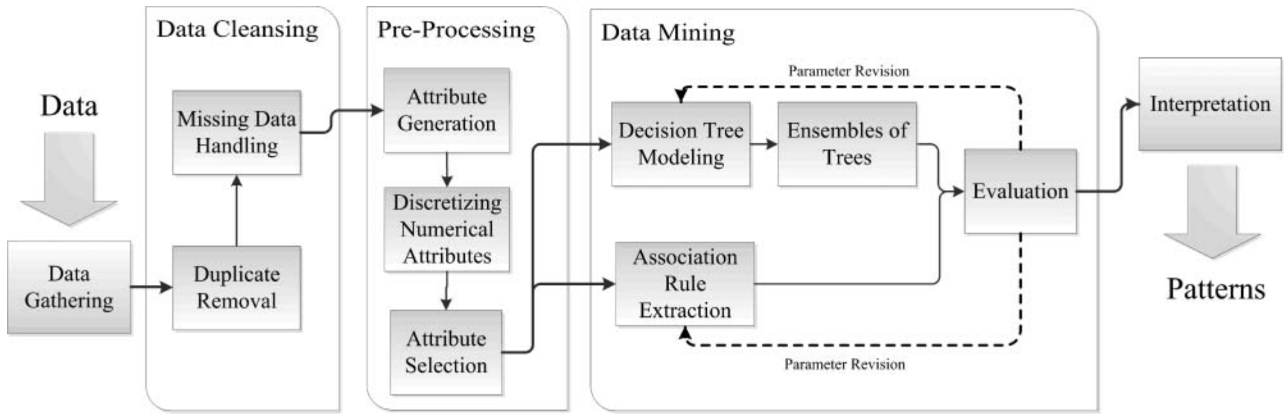


Figure 1. The flowchart of the methodology.

of accidents for each category of each variable. In the next step, a multiple correspondence analysis (MCA) was employed to find out which variables showed correspondences. Finally, three data-mining models: the decision tree, ensembles of trees and association rules were applied. In the third step, the SPSS software and two professional data-mining softwares, namely the Weka freeware as well as the Rapid Miner are used.

The multiple-correspondence analysis is a technique to find patterns of association among qualitative variables of a data-set. In this procedure, the information will be organized along two orthogonal dimensions.

The principal reason for the employment of the above-mentioned data-mining techniques is that the interpretable structure of these techniques as well as the main purpose of this paper for identifying the pattern of high-risk accidents makes them suitable for our study. In addition, based on the literature studies, these two methods of decision tree and association rules were the two most widely used approaches in the previous research works on the analysis of accident data (Cheng et al., 2010; Liao & Perng, 2008; Liao et al., 2009; Nenonen, 2012). Furthermore, we proposed a novel approach by using ensemble methods and tree-based models to provide more prediction accuracy.

The decision tree technique is a classification method that can produce interpretable rules and is represented in a tree-like diagram. In this tree, the nodes are labelled by the attributes, the edges specify the attribute values and the leaves represent the class label of the branches leading to it. Then, the branches construct the rule set of the class labels. Like other classification methods, the decision tree requires a training data-set to be constructed. In each step, an attribute with the highest capability of data separation is chosen. For this aim, a criterion named 'information gain' is defined in order to minimize the entropy of each class label. Two main parameters in the tree construction are 'confidence factor' and 'minimum number of objects'. The smaller the confidence factor is, the more pruned the tree will be. In addition, the minimum number of objects in the leaves can limit the height of the tree. To classify the unknown and test data-set, each instance is routed through the tree based on its values of the attributes in each node, and it leads to a leaf which determines its class label according to the label of the leaf.

According to the previous studies (Mingers, 1989), decision tree has limited accuracy. An efficient way to improve its accuracy is the usage of ensemble methods. An ensemble is a supervised algorithm which needs two sets of training and testing data. Ensembles combine different classifiers to produce a stronger result. The two principal algorithms of this family are Bagging and AdaBoost. One of the best classifiers for these two ensemble techniques is decision tree. Then, we implemented Bagging and AdaBoost methods by combining different decision trees as their classifiers.

The other data-mining technique used in this paper is the association rules. Generally, the association rules are used for discovering the hidden relations and correlations among massive data-sets (Han & Kamber, 2001). The association rule sets consist of the rules in the following structure. The body and head parts contain some attributes with specific values that exist in the data-set and the head part is considered as the result

$$\text{Body} \Rightarrow \text{head}\;\;[\text{support}, \text{confidence}]. \tag{1}$$

The strength of an association rule can be measured with the parameters of support, confidence and lift (Giudici, 2003; Wang, Yeh, Huang, & Chang, 2009). The support value shows how often a rule is applicable to a given data-set, while confidence determines how frequently the items in the head appear in transactions that contain the body

$$\text{Support}\ (A \Rightarrow B) = P(A \cup B), \tag{2}$$

$$\text{Confidence}\ (A \Rightarrow B) = \frac{P(A \cup B)}{P(A)}. \tag{3}$$

In addition to support and confidence parameters, Brin, Motwani, and Silverstein defined a new evaluation metric named lift (Brin, Motwani, & Silverstein, 1997). This metric compares the confidence of a rule with the occurrence of the head part. This is because in some cases, the high value of confidence is the result of the high occurrence of the head part (Wang et al., 2009). Then, the lift measures the probability of simultaneous occurrence of body and head and is calculated by the ratio of confidence to the number of occurrences of the head. Hence, the higher the value of lift, the stronger the relation of the body and the head will be

$$\text{Lift}\,(A \Rightarrow B) = \frac{P(B\,|\,A)}{P(B)} = \frac{\text{confidence}\,(A \Rightarrow B)}{P(B)}. \tag{4}$$

In this paper, first, the decision tree method and its combination with ensemble methods is used to determine the effective factors in the occurrence of high-risk occupational accidents in comparison with the others. In the second step, we use ensemble methods to combine different decision trees with the goal of improving accuracy. We tested Bagging and AdaBoost algorithm implemented in Weka and the precision of Bagging method was better. As the result of single decision tree model and ensembles of decision trees were similar, we selected the accurate output of ensemble method to demonstrate in this paper.

Moreover, we use the association rules method to analyse the circumstances and possible consequences of high-risk occupational accidents. For this aim, we consider the records related to high-risk accidents individually. Minimum values of 20% for support and 70% for confidence are considered in the employed association rules method similar to previous studies (Cheng et al., 2010; Liao & Perng, 2008). Moreover, a number of rules with the lift value of more than one are chosen.

3. Results

Results of investigating accident groups of I and II are presented in three following subsections.

3.1. *Descriptive analysis of data*

The list of variables used in data-mining analysis is shown in Table 3. This table also includes the frequency distribution of accident groups of I and II based on the variables. It is observed that the likelihood of accidents in group I is more than other accidents in some categories, including older workers, the afternoon hours, due to exposed and defective equipment, causing injury in limbs, cranium, brain, back, spine, neck, and lost workdays more than two months. Conversely, it can be seen from Table 1 that the occurrence of these accidents (group I) is less probable than other accidents in some categories including hand injuries and between 4:31 pm and 7:00 am.

In group II, the frequency of accidents is more than other accidents in the categories, including married and

Table 3. Comparing group I and II and other accidents based on variables used in data-mining analysis.

Variables and categories used in the analyses	Frequency *n* (%)			
	Group I of accidents	Other accidents	Group II of accidents	Other accidents
Marital status of the injured worker				
Single	2603 (27)	3388 (28)	202 (20)	5789 (28)
Married	7068 (73)	8805 (72)	822 (80)	15,051 (72)
Age of the injured worker (years)				
≤19	562 (6)	619 (5)	27 (3)	1154 (6)
20–35	5555 (57)	7414 (61)	585 (57)	12,384 (59)
36–45	1972 (20)	2451 (20)	224 (22)	4199 (20)
46–55	1135 (12)	1239 (10)	134 (13)	2240 (11)
≥56	447 (5)	470 (4)	54 (5)	863 (4)
Time of accident				
07:01–10:00	3249 (34)	4068 (33)	240 (24)	7077 (34)
10:01–12:30	1902 (20)	2418 (20)	178 (17)	4142 (20)
12:31–14:00	1084 (11)	1369 (11)	139 (14)	2314 (11)
14:01–16:30	2110 (22)	2472 (20)	178 (17)	4404 (21)
16:31–21:00	802 (8)	1047 (9)	145 (14)	1704 (8)
21:01–07:00	524 (5)	819 (7)	144 (14)	1199 (6)

(*continued*)

Table 3. (*Continued*)

Variables and categories used in the analyses	Frequency n (%)			
	Group I of accidents	Other accidents	Group II of accidents	Other accidents
Day of the week				
Saturday	1547 (16)	1913 (16)	143 (14)	3317 (16)
Sunday	1522 (16)	1937 (16)	147 (14)	3312 (16)
Monday	1470 (15)	1841 (15)	156 (15)	3155 (15)
Tuesday	1499 (16)	1942 (16)	160 (16)	3281 (16)
Wednesday	1442 (15)	1820 (15)	159 (16)	3103 (15)
Thursday	1383 (14)	1690 (14)	153 (15)	2920 (14)
Friday	808 (8)	1050 (8)	106 (10)	1752 (8)
Season of accident occurrence				
Spring	2523 (26)	3204 (26)	259 (25)	5468 (26)
Summer	2862 (30)	3515 (29)	314 (31)	6063 (29)
Fall	2466 (25)	3174 (26)	254 (25)	5386 (26)
Winter	1820 (19)	2300 (19)	197 (19)	3923 (19)
Place of accident				
Inside workshop	9320 (96)	11,442 (94)	752 (73)	20,010 (96)
Outside workshop	340 (4)	662 (5)	206 (20)	796 (4)
During commuting to the workshop	11 (0.1)	89 (1)	66 (7)	34 (0.2)
Cause of accident				
Imprudence	5316 (55)	7126 (58)	484 (47)	11,958 (57)
Improper environmental conditions	74 (1)	116 (1)	10 (1)	180 (1)
Exposed and defective equipment	1356 (14)	1191 (10)	104 (10)	2443 (12)
Lack of awareness	188 (2)	332 (3)	26 (3)	494 (2)
Noncompliance with safety regulations	732 (7)	807 (7)	97 (9)	1442 (7)
Others	2005 (21)	2621 (21)	303 (30)	4323 (21)
Body part affected				
Cranium and brain	507 (5)	292 (2)	66 (6)	733 (4)
Eyes	58 (1)	294 (2)	12 (1)	340 (2)
Face	377 (4)	487 (4)	74 (7)	790 (4)
Neck	161 (2)	127 (1)	25 (3)	263 (1)
Hand	3144 (32)	6246 (51)	248 (24)	9142 (44)
Trunk	307 (3)	213 (2)	27 (3)	493 (2)
Spine and back	919 (10)	471 (4)	62 (6)	1328 (6)
Limbs	3275 (34)	3102 (25)	312 (31)	6065 (29)
Others	923 (9)	961 (7)	198 (19)	1686 (8)
Lost workdays (in calendar days)				
0 days	882 (8)	874 (7)	229 (23)	1467 (7)
1–30 days	3345 (35)	4962 (41)	267 (26)	8040 (38)
31–60 days	2766 (29)	3675 (30)	208 (20)	6233 (30)
61–120 days	2008 (21)	2128 (17)	236 (23)	3900 (19)
>120 days	730 (7)	554 (5)	84 (8)	1200 (6)
Final consequence of accident				
Death	64 (1)	71 (1)	42 (4)	93 (0.4)
Total disability	119 (1)	131 (1)	54 (5)	196 (1)
Disability between 66% and 33%	181 (2)	226 (2)	42 (4)	365 (2)
Disability between 33% and 10%	387 (4)	659 (5)	66 (6)	980 (5)
Complete recovery	8920 (92)	11,106 (91)	820 (81)	19,206 (92)
In total	9671 (100)	12,193 (100)	1024 (100)	20,840 (100)

older workers, during lunch hours and between 4:31 pm and 7:00 am, on weekends and the summer period, while commuting to the workshop, due to noncompliance with safety regulations, causing injury in cranium, brain and neck, with no lost workdays (which probably is a sign of instant death) or more than two months, and resulting in death or disability.

3.2. *Analysis of variables using multiple correspondence analysis*

A multiple correspondence analysis was applied on the 10 variables of groups I and II of accidents to reveal the correlation between variables.

3.2.1. *Group I of accidents*

Two dimensions were presented by the model, so that the first illustrated some 26.66% of the variance and an autovalue of 2.666 with a Cronbach coefficient of 0.802, while the second dimension illustrated 41.24% of the variance and an autovalue of 4.124 with a Cronbach coefficient of 0.642. Hence, for the overall model, the total variance illustrated was 33.95%, the mean autovalue was 3.395 and the mean coefficient of the Cronbach α was 0.722, pointing out proper reliability of the model.

The discrimination measurements of each variable regarding each of the two dimensions are presented in Table 4. The first dimension showed very large discriminations with the variables age of the injured worker (0.713) and marital status of the injured worker (0.696). The second dimension also demonstrated strong discriminations with the variables body part affected (0.890), final consequence of accident (0.780), and lost workdays (0.662).

The result of representing these values in a system of orthogonal axes was a figure of discrimination measurements of the variables in the model (Figure 2). According to this figure, the variable age of the injured worker was the most explicative variable of the homogenizing model. In addition, very explicative variable correlated with the previous one was marital status of the injured; less explicative variables were season of accident occurrence, place of accident, day of the week, cause of accident and time of accident.

The variables age of the injured worker and marital status of the injured worker were most related, because the angle made by the lines that connect the origin of the coordinates with both variables was smaller. Also, the variables time of accident, place of accident, body part affected, final consequence of accident and lost workdays are related, while the variables season of accident occurrence, cause of accident and day of the week were the least related owing to the greatest angle between these lines.

From another point of view, similar discrimination measures of a variable in the two dimensions expresses that assigning the variable to a given dimension is difficult. Hence, the situation of a variable with a high value in a single given dimension and a low one in the other would be ideal, as occurred with age of the injured worker and marital status of the injured worker, such that these variables were more correlated with dimension 1. In the same way, the variables time of accident, place of accident, body part affected, final consequence of accident and lost workdays were more correlated with dimension 2.

3.2.2. *Group II of accidents*

Dimension 1 illustrated 47.33% of the variance and an autovalue of 4.733 with a Cronbach coefficient of 0.810,

Table 4. Discrimination measures of the variables in each dimension.

	Group I of accidents			Group II of accidents		
		Dimension			Dimension	
Variable	Mean	1	2	Mean	1	2
Marital status of the injured worker	0.353	0.696	0.009	0.576	0.208	0.944
Age of the injured worker	0.363	0.713	0.013	0.588	0.334	0.842
Time of accident	0.270	0.060	0.480	0.549	0.762	0.336
Day of the week	0.340	0.400	0.280	0.252	0.288	0.216
Season of accident occurrence	0.263	0.405	0.120	0.210	0.224	0.196
Place of accident	0.240	0.060	0.420	0.427	0.746	0.108
Cause of accident	0.295	0.120	0.470	0.338	0.342	0.333
Body part affected	0.498	0.106	0.890	0.464	0.642	0.285
Lost workdays	0.354	0.046	0.662	0.414	0.524	0.304
Final consequence of accident	0.420	0.060	0.780	0.498	0.663	0.333
Total active	3.395	2.666	4.124	4.315	4.733	3.897
% of variance	33.95%	26.66%	41.24%	43.15%	47.33%	38.97%

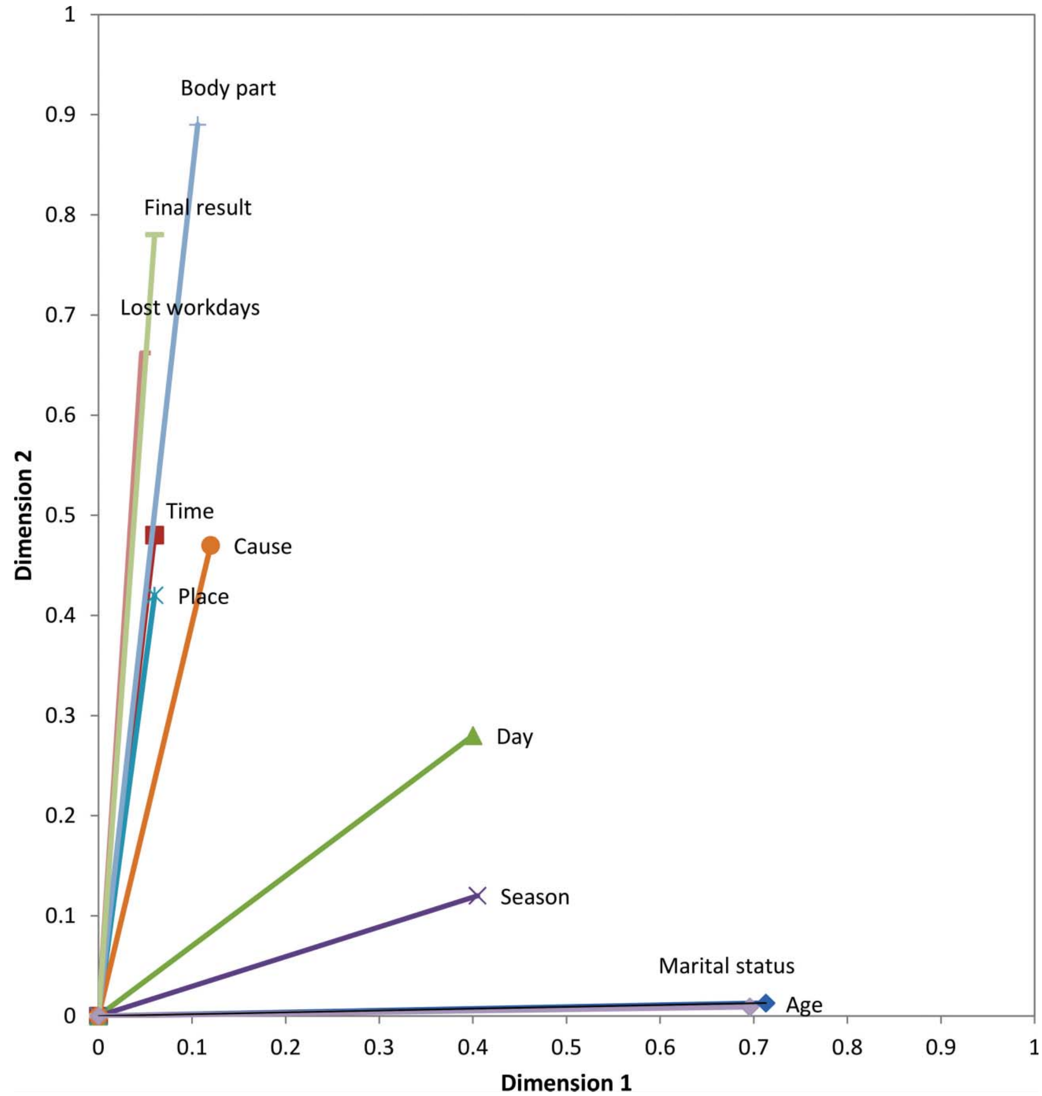


Figure 2. Representation of the discrimination measures of the variables of accidents in group I.

while the second dimension illustrated 38.97% of the variance and an autovalue of 3.897 with a Cronbach coefficient of 0.746. Hence, the total variance illustrated was 43.15% for the overall factorial model, and the mean autovalue was 4.315 with the mean coefficient of the Cronbach α of 0.778, pointing out good reliability of the model.

Table 4 demonstrates the discrimination measurements of each variable with regards to each of the two dimensions of the model. Dimension 1 showed very large discriminations with the variables time of accident (0.762) and place of accident (0.746), while the second dimension demonstrated high discriminations with the variables age of the injured worker (0.944) and marital status of the injured worker (0.842).

Figure 3 depicted in a system of orthogonal axes presenting that the variable marital status of the injured worker was the most explicative variable of the variance of the homogenizing model, while the variables age of the injured worker, place of accident and time of accident were very explicative; less explicative variables were season of accident occurrence, day of the week and cause of accident.

The most related variables were time of accident and body part affected, while the variables age of the injured worker and cause of accident were the least related.

Place of accident was more correlated with dimension 1, and the variable marital status of the injured worker was more correlated with dimension 2.

3.3. *Analysis of factors influencing accidents using ensembles of decision tree*

As explained before, the ensemble method produces different trees for each portion of data-set and finally it combines them to produce a single tree. In this section,

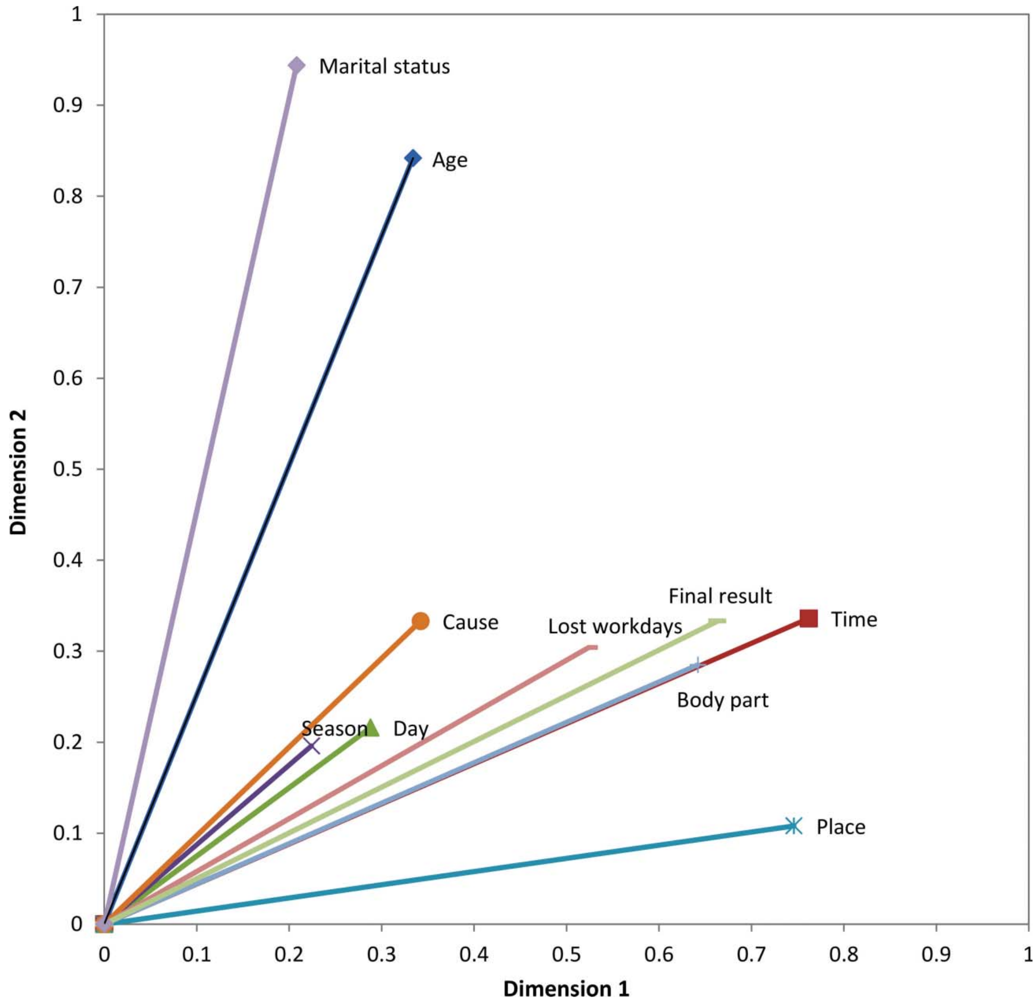


Figure 3. Representation of the discrimination measures of the variables of accidents in group II.

because of lack of space, we just present the final combined trees for accident groups I and II.

3.3.1. *Group I of accidents*

According to the model (Figure 4), the most important variable used in this analysis is the body part affected. Eighty-two per cent of the accidents that led to back and spine injuries are related to group I. Next, neck and brain injury by 74% and 73%, respectively, are related to this group of accidents. However, facial and hand injuries in group I are the least likely.

The next important factor is the cause of the accident. Seventy-six per cent of accidents in which the limbs are affected and the accident is due to exposed and defective equipment are within group I. However, only 32% of accidents involving the limbs and due to improper environmental conditions fall within the accidents of the first group.

In addition, noncompliance with safety regulations (and not using protective equipment) is the second important cause affecting the occurrence of accidents in group I. In this regard, it can be observed that 70% of accidents resulting in limb injury and are due to noncompliance with safety regulations are in group I of the accidents.

The next significant factor is the age of workers. Sixty-seven per cent of accidents causing limb injury and occurred due to imprudence of workers aged 45−55 years are in group I.

3.3.2. *Group II of accidents*

According to this model (Figure 5), the most important variable is the place of the accident. Ninety-one per cent of accidents that occurred while commuting to the workshop are the second group of accidents. However, accidents inside and outside the workshop are least likely to be in group II.

The second important factor is the injured body part. Seventy-six per cent of accidents that occurred outside the workshop and resulted in cranium and brain injury, and

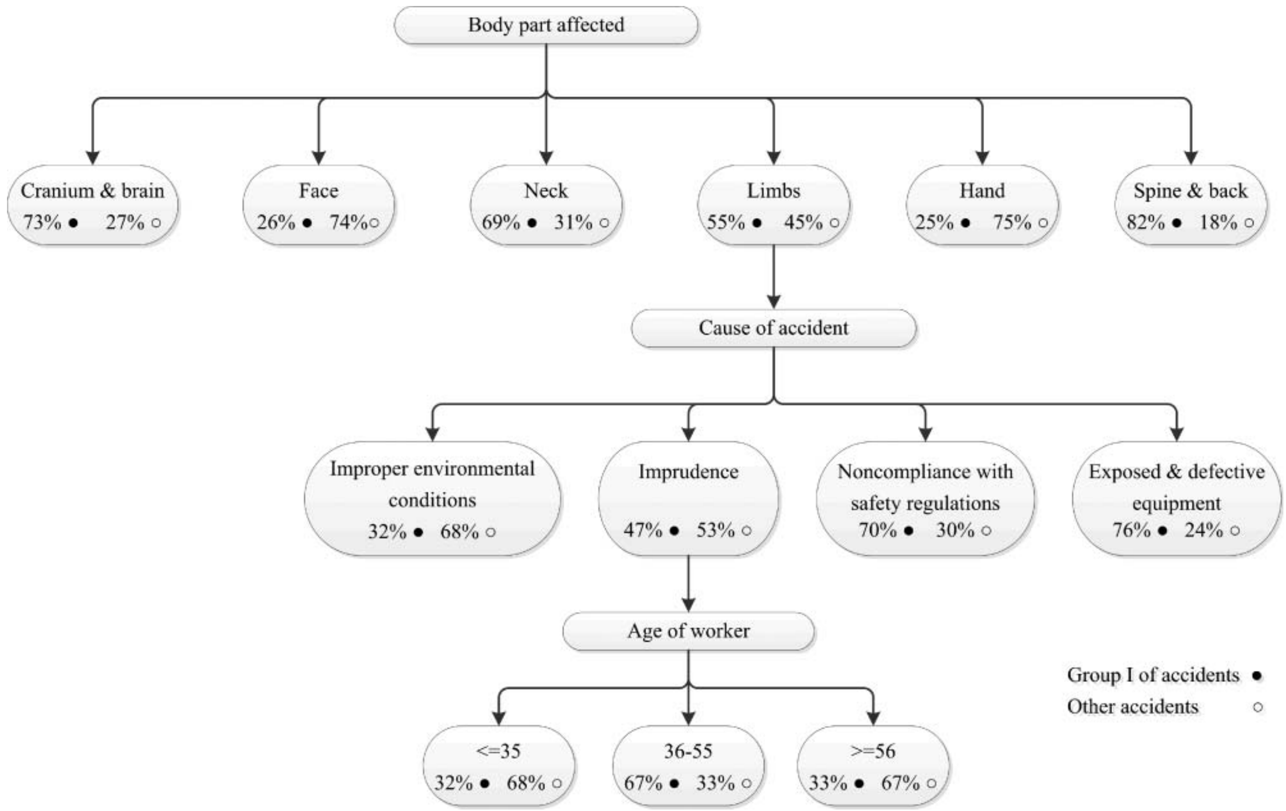


Figure 4. Tree model of accidents in group I.

69% of accidents that occurred outside the workshop with face injuries are in group II. However, only 20% of accidents that occurred outside the workplace and limbs are injured, fall within this group. Moreover, 72% of off-site accidents causing hand injury in which lost time equals zero correspond to group II.

On the other hand, in 86% of accidents that occurred inside the workshop, causing death or above 33% disability, no workdays were lost and simultaneously took place in the night shift (9:01 pm to 7:00 am) were in group II.

3.4. *Analysis of relationships between circumstances and consequences of accidents using the association rules method*

In this section, the data of each group of accidents were isolated, and numerous trial and error experiments by changing attributes affecting the rules were performed. Consequently, the most significant results of these tests are presented in Tables 5 and 6.

In Table 5, rules 1–4 were selected based on the support and confidence parameters, and rules 5–7 were chosen based on the lift value. Selection criteria were described previously in Section 2.2. Support of rules 1–4 varies between 10% and 20% and their confidence value is at least 90%.

The lift value of rules 5–7 are between 1.12 and 1.23. The lift value of more than one means that the body part of the rule has a positive impact on the occurrence of transactions that contain the head part. These rules are related to a small proportion of group I in which the support values are between 7% and 15%, and confidence levels vary between 40% and 80%.

In Table 6, the rules were also selected based on support and confidence parameters. The support values of the rules vary between 7% and 20% and their confidence value is at least 80%. Rules 15–18 are related to a small proportion of group II in which the support value is between 10% and 20% and the confidence is at least 80%.

4. Discussion and conclusion

In general, the findings of this study are in line with the results of previous research. The first group of accidents comprises a significant proportion of occupational accidents in the Iranian construction industry (44%). This ratio is consistent with similar studies (see Ale et al., 2008; Halvani et al., 2012; Im et al., 2009; Müngen & Gürcanli, 2005; Pérez-Alonso, Carreño-Ortega, Vázquez-Cabrera, & Callejón-Ferre, 2012; Tam, Zeng, & Deng, 2004). For example, Tam et al. presented that falls from height was the most frequent type of accidents in the construction industry of China (50%) (Tam et al., 2004). It

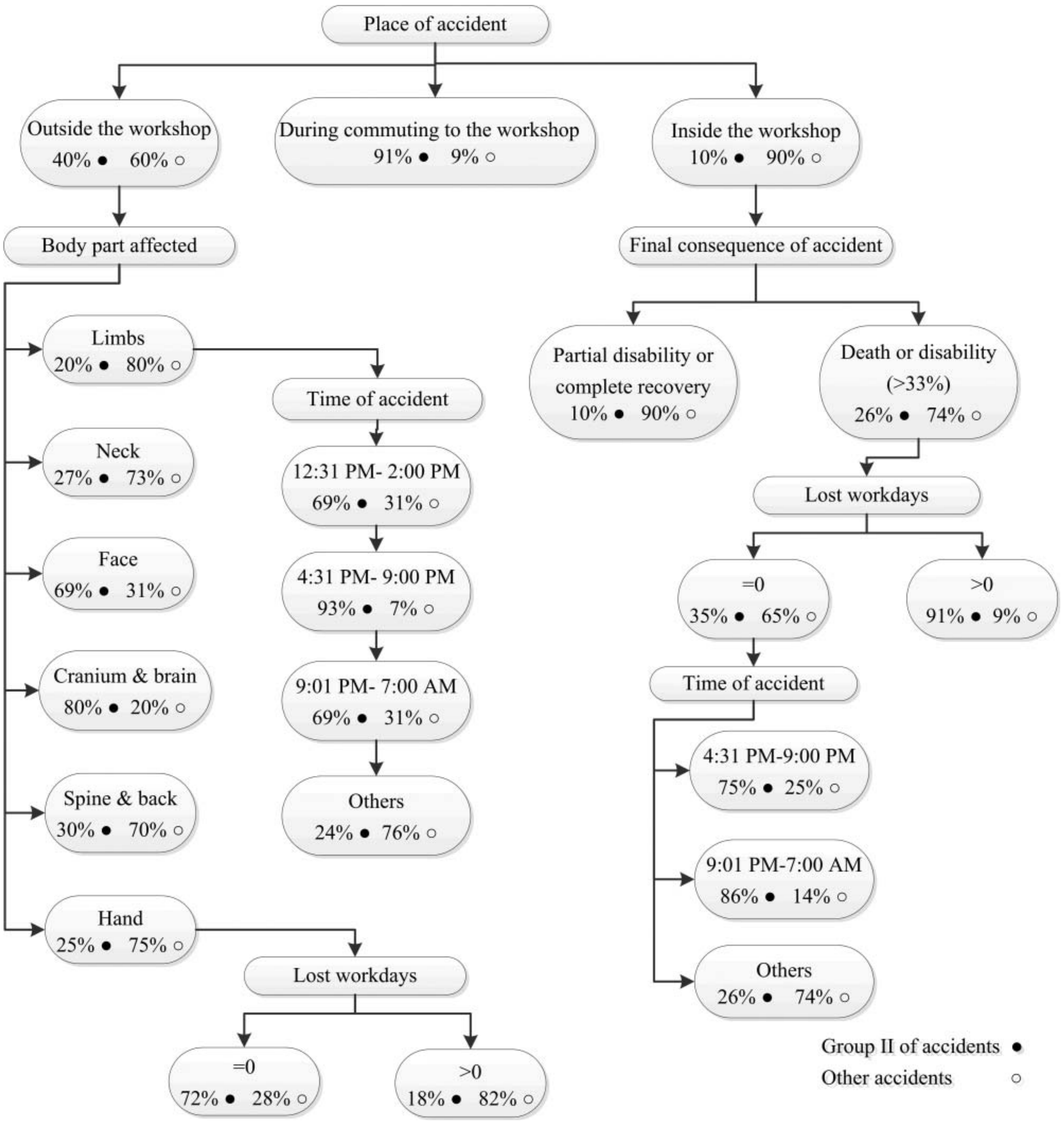


Figure 5. Tree model of accidents in group II.

seems that providing enough light to work at heights at night is harder than providing a prerequisite for other activities; hence, the frequency of occurrence of these accidents in the night shift is less than other accidents. This finding is in line with research conducted by Wojtczak-Jaroszowa and Jarosz. They found that the frequency of falls diminishes between 2:00 am and 4:00 am (Wojtczak-Jaroszowa & Jarosz, 1987). Injuries to the head, back, spine and lower extremities are more than other body parts which is in line with previous research (Huang, & Hinze, 2003; Nenonen, 2012; Scallan, Staines, Fitzpatrick, Laffoy, & Kelly, 2004). Thoracolumbar spine fracture during a fall from height is originated from axial load transferred to the spine because of the sudden deceleration of the buttocks upon impact with the ground integrated with the continued downward momentum of the torso, upper extremities, head and neck (Ivancic, 2013). The final consequence (severity of accidents) is almost

Table 5. Extracted association rules for circumstances and consequences of group I.

ID	Association rule		Confidence (%)	Lift
	Body part	Head part		
1	Lost workdays between 31 and 60 days	Complete recovery	96	–
2	Lost workdays between 31 and 60 days, and imprudence	Complete recovery	96	–
3	Age of worker between 20 and 35	Complete recovery	93	–
4	Lost workdays between 61 and 120 days	Disability between 10% and 33%	97	–
5*	Back or spine injury	Age of worker between 20 and 35	64	1.23
6*	Cranium and brain injury	No lost workdays	57	1.14
7*	Back or spine injury	No lost workdays	44	1.12

*These rules are related to a part of accidents in group I which resulted in death or disability.

identical to the other accidents. The reason may be the lower height of buildings in Iran as compared to more developed countries (based on building permits issued by municipalities in the year 2011, 77% of the permits issued are devoted to one to four-story buildings). It was observed that the frequency of accident occurrence among younger workers of group I of accidents was less than others. This finding is parallel with past research (Camino López, Ritzel, Fontaneda, & González Alcantara, 2008; Huang, & Hinze, 2003). It was also observed that a higher ratio of injured young workers recovered. This finding corresponds well with Salminen's study in which he concluded that the injuries of young workers were less often fatal than those of older workers (Salminen, 2004).

According to the findings of the MCA technique, there is a high correlation between age and marital status of the injured workers of group I of accidents which seems logical. Furthermore, it was found that in group I, variables time of accident, place of accident, body part affected, final consequence of accident and lost workdays are related. This finding is in parallel with research conducted by Courtney et al. who found association among the variables body part affected, severity of accident (i.e. final consequence of accident) and days away from work (Courtney, Sorock, Manning, Collins, & Holbein-Jenny, 2001).

On the other hand, the second group of accidents has also been identified as serious and fatal accidents in

Table 6. Extracted association rules for circumstances and consequences of group II.

ID	Association rule		Confidence (%)
	Body part	Head part	
1	Lost workdays between 1 and 30 days	Complete recovery or disability between 10% and 33%	95
2	Lost workdays between 1 and 30 days, and age of worker between 20 and 35	Complete recovery or disability between 10% and 33%	99
3	Lost workdays between 61 and 120 days	Complete recovery or disability between 10% and 33%	94
4	Lost workdays between 61 and 120 days, and imprudence	Complete recovery or disability between 10% and 33%	96
5	Hand injury	Complete recovery or disability between 10% and 33%	92
6	Limbs injury	Complete recovery or disability between 10% and 33%	91
7	Accident occurrence in spring	Complete recovery or disability between 10% and 33%	90
8	Age of worker between 20 and 35 days, and imprudence	Complete recovery or disability between 10% and 33%	89
9	Imprudence	Complete recovery or disability between 10% and 33%	89
10	Accident occurrence in summer	Complete recovery or disability between 10% and 33%	88
11	Accident time between 7:01 and 10:00 am	Complete recovery or disability between 10% and 33%	88
12	Accident time between 10:01 am and 12:30 pm	Complete recovery or disability between 10% and 33%	90
13	Age of worker between 20 and 35	Complete recovery or disability between 10% and 33%	87
14	Accident occurrence in summer and imprudence	Complete recovery or disability between 10% and 33%	91
15*	Accident occurrence on Monday	No lost workdays	83
16*	During commuting to the workshop	No lost workdays	82
17*	Accident time between 10:01 am and 12:30 pm	Inside the workshop	82
18*	Accident occurrence in summer and inside the workshop	No lost workdays	82

*These rules are related only to fatal accidents in group II.

previous studies (Im et al., 2009; Müngen & Gürcanli, 2005; Scallan et al., 2004; Suárez-Cebador, Rubio-Romero, & López-Arquillos, 2014). In this group, it was observed that the frequency of accidents among married workers is more than single ones. This finding is probably due to the occurrence of these accidents among older workers (which matches to the observations). Ling, Liu, and Woo observed that the frequency of severe accidents in the elderly is higher than in other age groups. They associated the reason to repeating an activity in their work and loss of consciousness in the elderly (Ling, Liu, & Woo, 2009). According to the results, the frequency of this group of accidents during lunch hours is higher than other accidents. This result is in line with past research (Suárez-Cebador et al., 2014). It is also observed that the frequency of occurrence of the second group of accidents in the afternoon and especially night hours, and also on weekends is much more than other accidents at work. This may be due to executing earth-moving activity at night and on weekends in Iran (to observe special urban traffic provisions for soil moving machinery, etc.). Moreover, accidents that occurred outside the workshop or during commuting to the workshop (which are mostly related to being hit by vehicle accidents) are more frequent in this group than others. Injuries to the head, face and neck in this group are more frequent than other accidents that had more severe (fatal and disabling) results. In this regard, the ratio of accidents with no lost workdays (which are probably related to instant death of a worker) and accidents with more than 60 lost workdays (which are probably related to the disability of a worker) are also greater than other accidents in the community studied.

Conducting the MCA technique, it was found that in group II of accidents, there is a high correlation between time of accident and body part injured. This finding is implied in past research. For instance, Loudoun showed that the time of accident has an impact on the severity (Loudoun, 2010). On the other hand, a statistically significant association between the injured body part and severity is also reported by Dumrak, Mostafa, Kamardeen, and Rameezdeen (2013).

4.1. *Limitations of the study*

Archiving the attributes of occupational accident digitally in the ISSO has just started about six years ago and is still not in accordance with comprehensive classifications and formats. In addition, the quality of gathering accident information by work inspectors is not yet satisfactory. Hence, some important variables such as worker occupation could not be considered in this study. Moreover, although according to the Iranian law ISSO must be notified of all occupational accidents causing injury to insured workers, it is possible that some cases remain unreported or misreported. The ISSO does not archive the attributes of near misses yet; therefore, this study is only based on accidents happened. Despite these limitations, this study was defined to be the first application of data-mining techniques on the occupational accident data of Iran.

The results of this study confirm the results of previous studies as a whole; hence, it can be concluded that the application of data-mining techniques has been successful. In this regard, the capabilities of these techniques in modelling large databases and detecting relationships between variables were identified as their advantage. Finally, the identified accident occurrence patterns can assist policy-makers, managers and safety professionals in the design and implementation of preventive measures and strategies.

Investigating other serious types of accidents and also analysing accidents considering weather conditions or their geographical distribution in the country could be considered as suitable subjects for future research.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Ale, B.J., Bellamy, L.J., Baksteen, H., Damen, M., Goossens, L. H., Hale, A.R., ... Whiston, J.Y. (2008). Accidents in the construction industry in the Netherlands: An analysis of accident reports using storybuilder. *Reliability Engineering & System Safety, 93*(10), 1523–1533.

Amiri, M., Ardeshir, A., & Fazel Zarandi, M.H. (2014). Risk-based analysis of construction accidents in Iran during 2007-2011-meta analyze study. *Iranian Journal of Public Health, 43*(4), 507–522.

Bevilacqua, M., Ciarapica, F.E., & Giacchetta, G. (2008). Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach. *Accident Analysis & Prevention, 40*(4), 1468–1479.

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond market baskets: Generalizing association rules to correlations. *ACM SIGMOD Record, 26*(2), 265–276.

Camino López, M.A., Ritzel, D.O., Fontaneda, I., & González Alcantara, O.J. (2008). Construction industry accidents in Spain. *Journal of Safety Research, 39*(5), 497–507.

Chang, L.Y., & Wang, H.W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention, 38*(5), 1019–1027.

Cheng, C.W., Leu, S.S., Lin, C.C., & Fan, C. (2010). Characteristic analysis of occupational accidents at small construction enterprises. *Safety Science, 48*(6), 698–707.

Cheng, C.W., Lin, C.C., & Leu, S.S. (2010). Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. *Safety Science, 48*(4), 436–444.

Courtney, T.K., Sorock, G.S., Manning, D.P., Collins, J.W., & Holbein-Jenny, M.A. (2001). Occupational slip, trip, and fall-related injuries–can the contribution of slipperiness be isolated? *Ergonomics, 44*(13), 1118–1137.

Dudarev, A.A., Karnachev, I.P., & Odland, Ø.J. (2013). Occupational accidents in Russia and the Russian Arctic. *International Journal of Circumpolar Health, 72*, 32–32.

Dumrak, J., Mostafa, S., Kamardeen, I., & Rameezdeen, R. (2013). Factors associated with the severity of construction accidents: The case of South Australia. *Australasian Journal of Construction Economics and Building, 13*(4), 32–49.

Giudici, P. (2003). *Applied data mining: Statistical methods for business and industry*. New York, NY: Wiley.

Halvani, G.H., Jafarinodoushan, R., Mirmohammadi, S.J., & Mehrparvar, A.H. (2012). A survey on occupational accidents among construction industry workers in Yazd city: Applying time series 2006–2011. *Journal of Occupational Health and Epidemiology, 1*(1), 1–8.

Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. *China Machine Press, 8*, 3–6.

Huang, X., & Hinze, J. (2003). Analysis of construction worker fall accidents. *Journal of Construction Engineering and Management, 129*(3), 262–271.

Im, H.J., Kwon, Y.J., Kim, S.G., Kim, Y.K., Ju, Y.S., & Lee, H. P. (2009). The characteristics of fatal occupational injuries in Korea's construction industry, 1997–2004. *Safety Science*, *47*(8), 1159–1162.

International Labour Organization (2013). Retrieved from http://www.ilo.org/safework/events/meetings/WCMS_204594/lang–en/index.htm.

Ivancic, P.C. (2013). Hybrid cadaveric/surrogate model of thoracolumbar spine injury due to simulated fall from height. *Accident Analysis & Prevention*, *59*, 185–191.

Liao, C.W., & Perng, Y.H. (2008). Data mining for occupational injuries in the Taiwan construction industry. *Safety Science, 46*(7), 1091–1102.

Liao, C.W., Perng, Y.H., & Chiang, T.L. (2009). Discovery of unapparent association rules based on extracted probability. *Decision Support Systems, 47*(4), 354–363.

Lin, Y.H., Chen, C.Y., & Wang, T.W. (2011). Fatal occupational falls in the Taiwan construction industry. *Journal of the Chinese Institute of Industrial Engineers, 28*(8), 586–596.

Ling, F.Y.Y., Liu, M., & Woo, Y.C. (2009). Construction fatalities in Singapore. *International Journal of Project Management, 27*(7), 717–726.

López Arquillos, A., Rubio Romero, J.C., & Gibb, A. (2012). Analysis of construction accidents in Spain, 2003–2008. *Journal of Safety Research, 43*(5), 381–388.

Loudoun, R.J. (2010). Injuries sustained by young males in construction during day and night work. *Construction Management and Economics, 28*(12), 1313–1320.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine Learning, 4*(2), 227–243.

Müngen, U., & Gürcanli, G.E. (2005). Fatal traffic accidents in the Turkish construction industry. *Safety Science, 43*(5), 299–322.

Nenonen, N. (2012). Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database. *Applied Ergonomics, 44*, 215–224.

Parhizi, S., Shahrabi, J., & Pariazar, M. (2009). A new accident investigation approach based on data mining techniques. *Journal of Applied Sciences, 9*(4), 731–737.

Pérez-Alonso, J., Carreño-Ortega, Á., Vázquez-Cabrera, F.J., & Callejón-Ferre, Á.J. (2012). Accidents in the greenhouse-construction industry of SE Spain. *Applied Ergonomics, 43* (1), 69–80.

Persona, A., Battini, D., Faccio, M., Bevilacqua, M., & Ciarapica, F.E. (2006). Classification of occupational injury cases using the regression tree approach. *International Journal of Reliability, Quality and Safety Engineering, 13*(02), 171–191.

Salminen, S. (2004). Have young workers more injuries than older ones? An international literature review. *Journal of Safety Research, 35*(5), 513–521.

Scallan, E., Staines, A., Fitzpatrick, P., Laffoy, M., & Kelly, A. (2001). *Injury in Ireland* (Report No. 15042) (Report of the Department of Public Health Medicine and Epidemiology). Dublin: University College Dublin.

Social Security Organization of the Islamic Republic of Iran. (2012). *Statistical report of occupational accidents in the construction industry between 2007–2011*. Tehran: Statistics and Social-economic Calculations Office.

Suárez-Cebador, M., Rubio-Romero, J.C., & López-Arquillos, A. (2014). Severity of electrical accidents in the construction industry in Spain. *Journal of Safety Research, 48*, 63–70.

Tam, C.M., Zeng, S.X., & Deng, Z.M. (2004). Identifying elements of poor construction safety management in China. *Safety Science, 42*(7), 569–586.

Wang, H.S., Yeh, W.C., Huang, P.C., & Chang, W.W. (2009). Using association rules and particle swarm optimization approach for part change. *Expert Systems with Applications, 36*(4), 8178–8184.

Warch, S.L. (2002). *Quantifying the financial impact of occupational injuries and illnesses, and the costs and benefits associated with an ergonomic risk control intervention within the unapprised business segment of UnitedHealth group* (Unpublished doctoral dissertation). Menomonie, WI: University of Wisconsin-Stout.

Wojtczak-Jaroszowa, J., & Jarosz, D. (1987). Time-related distribution of occupational accidents. *Journal of Safety Research, 18*(1), 33–41.