Accident Analysis and Prevention 40 (2008) 1468–1479

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

# Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach

M. Bevilacqua[a,1], F.E. Ciarapica[b,*], G. Giacchetta[b,2]

[a] DIEM Forlì, University di Bologna, Via Fontanelle 40, 47100 Forlì, Italy
[b] Energy Department, University Politecnica delle Marche, Via Brecce Bianche, Ancona, Italy

ARTICLE INFO

*Article history:*
Received 21 June 2007
Received in revised form 18 March 2008
Accepted 21 March 2008

*Keywords:*
Occupational injury
Ergonomics
Safety management
Data mining
Classification and regression tree
Rules induction

ABSTRACT

This work is an attempt to apply classification tree methods to data regarding accidents in a medium-sized refinery, so as to identify the important relationships between the variables, which can be considered as decision-making rules when adopting any measures for improvement.

The results obtained using the CART (*Classification And Regression Trees*) method proved to be the most precise and, in general, they are encouraging concerning the use of tree diagrams as preliminary explorative techniques for the assessment of the ergonomic, management and operational parameters which influence high accident risk situations. The Occupational Injury analysis carried out in this paper was planned as a dynamic process and can be repeated systematically. The CART technique, which considers a very wide set of objective and predictive variables, shows new cause–effect correlations in occupational safety which had never been previously described, highlighting possible injury risk groups and supporting decision-making in these areas.

The use of classification trees must not, however, be seen as an attempt to supplant other techniques, but as a complementary method which can be integrated into traditional types of analysis.



## 1. Introduction

The application of ergonomics to the planning, organization and set up of equipment in the workplace is not only a legal obligation, but also an indispensable requirement for any firm involved in an entrepreneurial approach to value creation.

Together with the optimization of worker productivity, worker safety is also of the utmost importance from the moral, legal and economic point of view. Ergonomically based design intervention targeted at controlling the occurrence of Occupational Injuries and Illnesses is essential component of any loss management program (Amell et al., 2001).

The goal of these programs is a desired reduction in the frequency and severity of accidents which result in Occupational Injury and Illness.

An ergonomics program does not consider only the anthropometric parameters of the human body in order to manufacture machines, devices and office furniture but also involves other aspects such as work organization, the content of the tasks and the work environment. Thinking and acting in an ergonomic way implies a global assessment of the relationship between man and work on the basis of the following factors:

- Minimum risk of injury and illness.
- Maximum work satisfaction.
- Maximum economic returns (Hoyos and Zimolong, 1988; Schweigert et al., 1999).

In a processing industry like a refinery the number of annual failures is likely to be very high, partly as a result of the normal wear of the components which are often subject to intensive working conditions. Within this context this study attempts to identify the ergonomic, management and operational factors which influence high accident risk situations. With specific reference to a medium-sized refinery (the API refinery in Falconara Marittima), all the information useful for identifying the factors which lead to accidents in the workplace was collected and analyzed. The analyses carried out indicated important relationships between the variables, providing useful decision-making rules which can be followed when adopting measures for improvement.

* Corresponding author at: Dipartimento di Energetica, Università Politecnica delle Marche, Via Brecce Bianche, 60100 Ancona, Italy. Tel.: +39 071 2204435; fax: +39 071 2204770.

*E-mail addresses:* maurizio.bevilacqua@unibo.it (M. Bevilacqua), f.ciarapica@univpm.it (F.E. Ciarapica), g.giacchetta@univpm.it (G. Giacchetta).

[1] Tel.: +39 071 2204874; fax +39 071 2204770.
[2] Tel.: +39 0712204763; fax: +390712204770.

0001-4575/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2008.03.012

A thorough understanding of the variables which influence a particular problem is in fact essential for finding increasingly efficient solutions.

Nevertheless, a great deal of data is often collected which is difficult to understand, considering the number of variables involved. In particular, before starting a complex and expensive analysis of the workplace situation, it would be better to understand the data structure thoroughly in order to identify the intrinsic relationships between the different variables involved and thereby to extract useful decision-making rules. To this end, the process of Data Mining uses data to find new and potentially useful knowledge. Data Mining finds relationships in the data set: it does not make decisions, but gives the decision maker the necessary information to do so.

The principal result of any classification process is that interpretation can improve problem-solving performance in the area of interest (Vorko and Jovic, 2000). Therefore, classifying and finding relationships among a set of variables is a complex and common problem even in Occupational Safety. Studies and classifications of occupational injuries have largely been descriptive in nature. To date, descriptive and analytical parametric modeling procedures such as Descriptive Statistics methods and Regression Analysis have often been used. A large number of studies usually describe the distribution of injuries (numbers, rates, and frequency index) in terms of person, place, and workplace characteristics, being useful for identifying hazardous industries, occupations, and work situations (Sorock et al., 2001; Amell et al., 2002; Biddle and Marsh, 2002; Larsson and Field, 2002; Vredenburgh, 2002; Siu et al., 2003). Another primary statistical tool often adopted in occupational injury studies is Regression Analysis (Khan and Jansson, 2001). This can be used to evaluate the relationship between the injury frequency index and one or more covariate or predictor variables. Researchers have often preferred logistic regression when analyzing data from a case–control study since they can easily quantify the results in the form of odds ratio (OR) and associated confidence intervals. Lindell (1997) and Roudsari and Ghodsi (2005) adopted Logistic Regression. On the other hand, Waller et al. (1995), Siu et al. (2004), Fabiano et al. (2004), Simo (2005) and Smith et al. (2005) used the Statistical Correlation and Significance Study using evaluation of P-value, Odds Ratio, Likelihood Ratio Test to support the results. Poisson Distribution is widely employed as an analytical tool in Safety Analysis, Reliability Engineering and in epidemiological medical studies. Examples of its application in occupational safety studies can be found in Clemens (2003), Yau et al. (2004) and Wang et al. (2003). In contrast with the traditional descriptive and parametric approaches, some studies involve analytical epidemiological methods such as the case-crossover study proposed by Sorock and Courtney (1997), and the case–control study by Hertz and Emmett (1986).

As a Data Mining technique, Classification and Regression Tree methodology (CART) is promising because of its advantages over standard statistical techniques. The differences between parametric and non-parametric approaches can be quite significant. The use of statistical techniques, such as linear regression, is based on general assumptions regarding the data set, which are normally difficult to satisfy. In the case examined here, the intrinsic structure and complexity of the data collected might jeopardize the use of traditional tools for analysis since the variables present the following critical characteristics:

- *high dimensionality*: a high number of predictive variables is a problem of considerable importance for standard statistical analysis in general;
- *interaction and dependency*: the relations between the independent variables can be a problem particularly in parametric analyses, which typically adopt the independence hypothesis;
- *non-homogeneity and non-linearity*: different relationships may exist between the variables in several parts of the measurement space.

From this point of view classification trees are a valid alternative and complementary tool to parametric methods, guiding the researcher towards a more thorough understanding of the data, without the need to formulate limiting *a priori* hypotheses (Mingers, 1987).

Compared with the simultaneity with which traditional methods consider the predictive variables in order to determine the impact on the objective variable, decision trees offer a hierarchical survey and a recursive approach. Therefore, note that in this approach, predictors chosen for subdivision of the first parent and child nodes are the most relevant. The final results of using tree methods for classification or regression can be summarized in a series of (usually few) logical if-then conditions (tree nodes). Therefore, there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variables are either linear, following some specific non-linear link function, or that they are even monotonic in nature. As a result, tree methods are particularly well suited for data mining tasks where there is often little *a priori* knowledge and no coherent set of theories or predictions regarding which variables are related and how they are related.

Some authors have tackled classification and prediction studies by comparing different methods. Chang and Chen (2005) analyzed vehicle accident frequency using Negative Binomial Regression models and CART. By comparing the prediction performance between the two methods this study demonstrates that CART (overall prediction accuracy 58.2% for training data and 52.6% for testing data) is a good alternative method for the Negative Binomial Regression model (overall prediction accuracy 52.9% for training data and 52.3% for testing data) for analyzing freeway accident frequencies. Generally, discriminant analysis and logistic regression are the two most commonly used Data Mining techniques. However, these methods have often been criticized because of their assumption about the categorical nature of the data (Lee et al., 2006). Brown et al. (2002) compared Logistic Regression and CART methods to study how well each model predicts whether or not an accident is a resultant of human error or not. They found that the CART decision tree model outperforms the LR model for cases correctly predicted (79.7% vs. 72.2%) and for the calculation of the “Sensitivity” and “Specificity” indices that detail the percentage of missed cases and false alarms respectively. Neural networks provide an alternative to Linear Discriminant Analysis and logistic regression, particularly in situations where the dependent and independent variables exhibit complex non-linear relationships. Neural Networks are also criticized for their long training process in designing the optimal network topology and their inability to identify the relative importance of potential input variables (Piramuthu, 1999). The CART method has proven to be an effective tool in handling classification problems. Lee et al. (2006) demonstrated that CART outperforms traditional non-parametric methods and neural networks in terms of accuracy to conduct credit scoring tasks. Ture et al. (2005) compared performance of three decision trees, four statistical algorithms and two neural networks to predict the risk of essential hypertension disease. They found that CART and the neural network models performed better than other techniques in terms of sensitivity, specificity and predictive rate.

For the case studied in this article we decided to use classification and regression tree methods in order to consider two aspects: (1) many categories, in turn divided into sub-categories, have been considered; (2) all the variables are of a categorical nature. For this reason the use of parametric or neural network methods would have required many assumptions regarding the starting data. These assumptions may compromise the reliability and validity of the results.

Moreover, in order to identify and analyze the important groups of cases, four of the most widely used classification tree algorithms were tested: CHAID (*Chi-Squared Automatic Interaction Detection*) (Kass, 1980); Exhaustive CHAID (Biggs et al., 1991); CART (*Classification And Regression Trees*) (Breiman et al., 1984); QUEST (*Quick, Unbiased, Efficient Statistical Tree*) (Loh and Shih, 1997). Each of them classifies the cases of the data sample in relation to a certain objective variable, in function of several predictive variables, which may be numerical (continuous) or categorical, and generates the above-mentioned tree diagrams, which highlight the correlations and the level of dependence of the objective variable on the predictors.

The rest of the paper is organized as follows. Section 2 is divided into two parts. The first paragraphs discuss the organization of the workgroup and the development of the methodology for analyzing the cases of injury. Section 2.1 discusses the procedures for risk assessment. In Section 3, the decision trees, developed using the CART method and based on a data set of 206 work-related injuries, have been analyzed. Section 4 highlights the possibility to use the trees both to recognize the most important operational variables which influence occupational injuries, and to represent a sound and easy to use decision-making system support tool. Finally, conclusions are presented in Section 5.

## 2. Materials and method

Data concerning more than 200 cases of accidents which occurred in the decade 1994–2004 were collected at the API refinery. The API refinery in Ancona can currently count on a processing capacity amounting to 3,900,000 ton/year of crude oil, a storage capacity for more than 1,500,000 $m^3$ and the ability to receive tankers and super-tankers up to a tonnage of 400,000 tons. Processing at the refinery is based essentially on a topping, catalytic reforming, isomerization, vacuum, visbreaking, and thermal cracking cycle, which is organized into a series of operational sections that form interconnected functional units. A work group made up of academics and refinery staff reorganized in detail all the data for each injury contained in the data set. For each injury, and in particular on the basis of the "description of the event", it was necessary to elaborate the information required for the analysis and the classification of the injuries.

There is seldom one cause of an accident, and most events are the result of a combination of causes. Each accident must be investigated thoroughly and action taken to reduce the risk of recurrence. According to Wilson (1998) managerial solicitation of worker feedback and input into solutions and risk management is one of the most effective strategies for controlling root causes.

The refinery has a "Safety and Reliability Department" which is legally bound to register various data for each accident that occurs: type of accident, date of the event, the part of the body injured and the description of the event. The main results extrapolated from the analysis of the accident reports are described below:

- The **TYPE OF ACCIDENT** indicates the work situation or the way in which the person had the accident: fall, accident with dangerous liquid, accident with a valve, incorrect support, collision, sudden movement, jump, mechanical damage, vehicle accident, crushing and non-classified accident.
- The **PERIOD**, obtained from the information concerning the date of the accident, The initial data set for the decade 1994–2004 was divided into 3 periods: **1st PERIOD**: 1994–1997 inclusive; **2nd PERIOD**: 1998–2001 inclusive; **3rd PERIOD** 2002–2004 inclusive. The first period follows the introduction in Italy of several decree laws issued on the subject of safety (Italian Law 626/94); the second period spans a period when some important plant reconstruction was carried out by the refinery; finally, the third refers to the period after the industry obtained OHSAS 18001 and ISO 14001 certification. The idea of differentiating the analysis by periods originated from the need to verify any possible changes in the factors which influence high accident risk situations over the years.
- **INJURED PART OF THE BODY**: mouth, arm, ankle, neck, leg, knee, hand, not specified, eyes, sacrum, foot, back, shoulder, head and chest.

The "Safety and Reliability Department" also carries out a re-examination of the event to investigate in detail the following aspects of each accident:

- **IMMEDIATE CAUSES of the accident**, that is to say what caused the accident. The questions asked in order to define the immediate causes are: Why did it happen? What actions or below standard conditions led to the event? The immediate causes were divided into six categories, which were further divided into sub-categories (see note[3] and Table 1);
- **ROOT CAUSES of the accident**, that is to say which element of the operational systems did not work correctly, creating the circumstances which led to the immediate causes? Why was that procedure or condition present? The root causes were divided into 6 categories, which were further divided into sub-categories (Table 2).
- **CORRECTIVE ACTION** which was considered necessary in order to avoid repetition of the same accident in the future. Corrective action was divided into four categories, each of which was further divided into sub-categories (Table 3).
- A **RISK INDEX** was also calculated for each injury. This index, of a numerical type, was transformed into a nominal value by dividing it into three categories: **LOW RISK**, **MEDIUM RISK** and **HIGH RISK**. LOW RISK value was assigned if the risk index was less than or equal to 4; MEDIUM RISK for a risk index between 4 and 10 and finally HIGH RISK if the risk index was greater than 10. The methodology used for calculating the index is described in Section 2.1.

In total 206 cases were collected. The variables considered are all of the categorical and nominal type and could all be used as objective variables. Consequently, the most interesting variables chosen were connected with legislative and economic problems, each of them being considered as an objective variable. The relative classification tree was further developed, using the remaining variables as predictors.

In order to understand the correlation with other variables, especially the immediate causes, the root causes and the corrective action, the objective variables studied were the following: (1) RISK CATEGORY, (2) TYPE OF ACCIDENT and finally (3) PERIOD.

Table 4 shows all the objectives and predictor variables considered.

An accurate analysis defined "the best" classification tree. A "progressive" analysis strategy was developed by considering the large number of potentially interesting variables. Several trees were expanded for each objective variable, characterized by an increasing number of predictors so that the impact of the entry of new variables on classifier precision could be evaluated.

[3] This type of sub-division of the data was made in accordance with an internal procedure applied by the Shell Company (S-RCM, 1999) which was specifically created for the refinery sector.

**Table 1**
Immediate causes of the accident

| Categories | Sub-categories |
|---|---|
| Ergonomics conditions below standard | Overcrowded area/limited movement possible<br>Lack of hygiene and tidiness/presence of obstacles<br>Inadequate or excessive lighting |
| Lack of communication | Inadequate knowledge of regulations and procedures<br>Warning given incorrectly/insufficiently/to the wrong person |
| Planning procedures below standard | Lack of assessment/incorrect or insufficient planning<br>Missing or inadequate procedure<br>Breakage/wear |
| Safety procedures below standard | Unsafe appliances/equipment<br>Dangerous transit/means of transport<br>Missing/wrong/inadequate safety conditions<br>Danger of fire/explosions |
| Operating procedures below standard | Incorrect/inappropriate use of equipment/appliances<br>Procedure not carried out<br>Operation carried out without authorization<br>Lack of precision/inappropriate speed of performance/haste<br>Necessary operation was forgotten<br>PPE (Personal Protective Equipment) used badly/faulty<br>Incorrect loading/lifting/substitution of equipment |
| Work conditions below standard | Equipment/machinery/materials not suitable for use |

**Table 2**
Root causes of the accident

| Categories | Sub-categories |
|---|---|
| Human factors | Physical/mental unsuitability<br>Physical/mental fatigue<br>Lack of motivation<br>Bad habits |
| Planning | Incorrect planning/construction<br>Inadequate instrumentation/information technology/electronic parts |
| Safety management | Inadequate documentation/procedures<br>Assessment/revision of safety, health, environment |
| Factors connected with the task | Problems with processes/operations<br>Inadequate guidance/control/organization/planning<br>Purchases/quality of materials/spare parts |
| Maintenance | Inspections/wear/incorrect use<br>Inadequate maintenance/contractors |
| Training | Lack of specific technical knowledge<br>Lack of general knowledge<br>Training, communication |

**Table 3**
Corrective action

| Categories | Sub-Categories |
|---|---|
| Training and information | Interview with the injured person<br>Safety talk<br>Training on the use of PPE<br>Intensify the training programme |
| Management | Revise documentation and procedures<br>Internal testing or assessment of modifications<br>Identify extra PPE for the workers<br>Plan new maintenance tasks, inspections<br>Control of contractors |
| Planning | Structural modification<br>Work order for component replacement |
| Legal requirements | Verbal warning<br>Written warning to the employee<br>Disciplinary sanction for the employee<br>Written warning to the contractor firm/possible disciplinary action<br>Penal sanctions for the external contractor firms |

**Table 4**
Objective and predictor variables considered during occupational injury analysis

| Objective variables | Potential predictor (independent) variable |
|---|---|
| RISK CATEGORY | Part of body injured |
| TYPE OF ACCIDENT | |
| PERIOD | *Immediate Cause*: Ergonomics |
| | *Immediate Cause*: Lack of communication |
| | *Immediate Cause*: Planning |
| | *Immediate Cause*: Safety |
| | *Immediate Cause*: Operating procedures below standard |
| | *Immediate Cause*: Working conditions below standard |
| | *Root Cause:* Human Factors |
| | *Root Cause*: Planning |
| | *Root Cause*: Safety management |
| | *Root Cause*: Factors connected with the task |
| | *Root Cause*: Maintenance |
| | *Root Cause*: Training |
| | *Corrective Action*: Training and information |
| | *Corrective Action*: Management |
| | *Corrective Action*: Planning |
| | *Corrective Action*: Legal requirements |

As mentioned in Section 1, in this study, four of the most widely used classification algorithms were tested: CHAID, Exhaustive CHAID, CART and QUEST. The level of accuracy of each tree was obtained by referring to *Risk value* with the Gini impurity measure (see Appendix A): the lower this value is, the better the tree obtained. Due to the large number of tests carried out, for the sake of brevity only the most interesting results are reported in Section 3 of this paper.

The results obtained using the CART method proved to be the most precise, showing high levels of accuracy (*Risk value)*, less than 0.25. The QUEST method provided slightly higher levels of accuracy, in most cases of between 0.2 and 0.3. Considerably worse levels were found using the other two methods. Given its greater accuracy compared with QUEST, attention was focused on CART (see Appendix A).

SPSS Answer Tree was used as a software tool to aid the analysis. This is a computerized learning system which can be used to create classification systems visualized as decision-making trees. Using "what–if" analysis this software adapts the model to specific requirements, and at the same time allows time-saving, because a model already tested on one data set can be instantly applied to new data with similar characteristics.

### 2.1. *Risk assessment*

The management of occupational injury is of strategic importance in a refinery from the organizational, engineering and economic point of view. The determination of an algorithm, that allows a methodical and as far as possible automatic approach to management of injury data, can lead to substantial improvements in the organization of work and in the decision-making processes.

Safety in a refinery relies, among other things, on the adopted management criteria. In particular, the Safety Management System (SMS) phase "Identification and evaluation of major hazards" takes into consideration the "adoption and implementation of procedures for systematically identifying major hazards arising from normal and abnormal operation and the assessment of their likelihood and severity" (Demichela et al., 2004). The selection of risk analysis methods and their results, in terms of frequency of occurrence and severity of consequences, is the focal point of the whole SMS and its procedures. In the European Union the Seveso II directive requires companies that store certain amounts of dangerous liquids or gases to introduce appropriate measures to reduce the potential risks (Van Heel et al., 1999). Safety management strategies for critical systems involve multiple dimensions including design philosophy, maintenance policies, and procedures for personnel hiring, training, and evaluation (Cowing et al., 2004).

In this work on the basis of S-RBI (1999) (see note[4]) procedures a risk index for each case of injury was calculated as the product of 2 factors: Probability $\times$ Consequence:

**Probability**: The likelihood of the considered accident occurring. Quantify the probability that an event will occur, assigning a class, from 1 to 5, to the possible scenario. In the case of the refinery a team of experts defined some criteria, which are illustrated in Table 5.
**Consequence**: The outcome of the accident. Failure of equipment in a refinery can be extremely dangerous for health and/or safety owing to toxic products, high pressures and high temperatures during the processes. The severity of the accident can be decided with reference to Table 6.

The risk index, of the numerical type, has been transformed into nominal by dividing it into the three categories: **LOW RISK, MEDIUM RISK** and **HIGH RISK**, as already explained in the first part of Section 2

## 3. Results

The results of the CART analysis for the three objective variables defined in Table 4, immediate causes, root causes and corrective action, are shown in Tables 7–9, respectively. The results only relate to those trees characterized by the best *risk values*. As can be seen, all trees show a low *risk value* (under 25%).

These results confirm the validity of the predictor variable choice. In fact, error can be partially ascribed to the quantity and quality of the data in the analyzed sample (i.e., some possible inaccuracy in data recording, some data not complete, etc.). It is also interesting to note how the most important predictor variable used to split the first root node changes depending on the type of objective variable considered.

The pruned CART trees show a simplified decision-making structure with a reduced number of nodes, improved readability, and easier operation. CART produces robust results by generating

[4] The calculation of the probability (Table 5) and severity (Table 6) are based on another Shell procedure (S-RBI manual, 1999. Shell-Risk Based Inspection. Shell Global Solution International).

**Table 5**
Assigning probability

| Class | Key word | Description of probability | 1 time every x. years | Absolute value | Examples |
|---|---|---|---|---|---|
| 1 | Very rare | Virtually impossible | >20 | 0.001 | -Death in aircraft accident<br>-Death caused by lightening<br>-Double Emergency<br>-Death in petrochemical plant<br>-Fire caused by pump with severe damage to the plant |
| 2 | Rare | Unlikely | 3–20 | 0.05 | -Theft of vehicle with anti-theft system<br>-Breakage of a pipe in a heat exchanger<br>-Fire caused by pump with slight damage to the plant<br>-Gas cloud explosion<br>-Death in house fire<br>-Plant fire with severe economic loss |
| 3 | Occasional | May happen a few times, at least once | 1–3 | 0.3 | -Blocked safety valve<br>-Fire in a furnace<br>-Death from natural causes |
| 4 | Probable | May happen several times | Once every 6 months | 0.5 | -Plant fire with slight economic loss<br>-Theft of vehicle without anti-theft system |
| 5 | Frequent | May happen often | >Once every 6 months | 1 | -Unexpected shutdown of a pump<br>-Non-functional appliance |

**Table 6**
Severity indices

| Severity | Key word | Consequences | Description | Examples |
|---|---|---|---|---|
| 1 | MINOR | Medications/no injury or slight injury/discomfort | Medication or injury with possible absence from work of between 1 and 3 days/discomfort in carrying out work | Grazes, first degree burns of limited extent |
| 2 | MODERATE | Minor injury/brief duration/poor health | Injury or poor health with loss of ability to work and possible absence from work for up to 10 days | Sprains, dislocation, skin irritation |
| 3 | SEVERE | Greater injury/long duration/occupational disease with reversible consequences | Injury or occupational disease with loss of ability to work for an extended period of up to 30 days | Torn muscle, second degree burns, fracture |
| 4 | VERY SEVERE | Long term injury/Injury to more than one person/Damage to health with permanent consequences | Injury with loss of ability to work for a period of more than 30 days and with possible multiple injuries caused by the same initial event; it includes any permanent physical or health impediments | Multiple fracture, third degree burns |
| 5 | CATASTROPHIC | Death/Lethal exposure | It includes the extreme possibility of death to individuals or groups due to the same initial event | Deaths |

**Table 7**
Results of CART analysis for immediate causes

| Objective variable | RISK CATEGORY | TYPE OF ACCIDENT | PERIOD |
|---|---|---|---|
| *Predictor variables* | Operating procedures below standard<br>Ergonomics<br>Lack of communication<br>Safety<br>Part of body injured<br>Type of accident | Lack of communication<br>Operational procedures below standard<br>Part of body injured<br>Work conditions below standard<br>Safety<br>Ergonomics | Operating procedures below standard<br>Ergonomics<br>Planning<br>Safety |
| *Final tree* | | | |
| Total number of nodes | 37 | 41 | 43 |
| Total number of levels | 5 | 5 | 5 |
| Total number of final nodes | 19 | 21 | 23 |
| *Risk analysis* | | | |
| *Risk value* | 0.09221 | 0.24043 | 0.21512 |
| *Risk value* (root node) | 0.50081 | 0.53102 | 0.51251 |
| Predictor variable splitting root node | Operating procedures below standard | Lack of communication | Operating procedures below standard |
| *Pruned tree* | | | |
| Total number of nodes | 24 | 26 | 29 |
| Total number of levels | 5 | 5 | 5 |
| Total number of final nodes | 13 | 15 | 15 |
| *Risk value* (performance) | 0.09270 | 0.24152 | 0.23717 |
| *Risk value* (root node) | 0.50081 | 0.53102 | 0.52051 |

**Table 8**
Results of CART analysis for root causes

| Objective variable | RISK CATEGORY | TYPE OF ACCIDENT | PERIOD |
|---|---|---|---|
| *Predictor variables* | Human factors | Factors connected with the task | Safety management |
| | Training | Training | Planning |
| | Safety management | Maintenance | Type of accident |
| | Factors connected with the task | Part of body injured | Training |
| | Type of accident | Human factors | Maintenance |
| | | Risk category | |
| *Final tree* | | | |
| Total number of nodes | 39 | 40 | 40 |
| Total number of levels | 5 | 5 | 5 |
| Total number of final nodes | 22 | 21 | 21 |
| *Risk analysis* | | | |
| *Risk value* | 0.09856 | 0.22743 | 0.20518 |
| *Risk value* (root node) | 0.530811 | 0.52102 | 0.50251 |
| Predictor variable splitting root node | Human factors | Factors connected with the task | Safety management |
| *Pruned tree* | | | |
| Total number of nodes | 21 | 23 | 23 |
| Total number of levels | 5 | 5 | 5 |
| Total number of final nodes | 11 | 15 | 15 |
| *Risk value* (performance) | 0.10270 | 0.23152 | 0.23717 |
| *Risk value* (root node) | 0.52081 | 0.55102 | 0.52251 |

**Table 9**
Results of CART analysis for corrective action

| Objective variable | RISK CATEGORY | TYPE OF ACCIDENT | PERIOD |
|---|---|---|---|
| *Predictor variables* | Management | Training and information | Management |
| | Training and information | Management | Legal requirements |
| | Type of Accident | Risk category | Planning |
| | Planning | Legal requirements | Type of accident |
| | Part of body injured | | |
| *Final tree* | | | |
| Total number of nodes | 23 | 23 | 31 |
| Total number of levels | 5 | 5 | 5 |
| Total number of final nodes | 12 | 12 | 16 |
| *Risk analysis* | | | |
| *Risk value* | 0.10621 | 0.19743 | 0.20125 |
| *Risk value* (root node) | 0.52081 | 0.44102 | 0.49251 |
| Predictor variable splitting root node | Management | Training and information | Management |
| *Pruned tree* | | | |
| Total number of nodes | 13 | 13 | 15 |
| Total number of levels | 4 | 4 | 4 |
| Total number of final nodes | 7 | 7 | 8 |
| *Risk value* (performance) | 0.10703 | 0.20158 | 0.20719 |
| *Risk value* (root node) | 0.50085 | 0.49106 | 0.49535 |

what is called a “maximal tree” and then examining smaller trees which are obtained by pruning away branches of the maximal tree. The important point is that CART does not stop in the middle of the tree-growing process because there might still be information to be discovered by breaking the tree down into several more levels.

As a result, the trees frequently grow larger than necessary, and so they must be selectively pruned back to improve the quality (i.e., the ability to classify) of the final decision diagram. The selective pruning of the trees will be obtained automatically from the full trees by adopting the 1-SE rule (one standard error). This choice was made because this rule generally provides the best results (Bevilacqua et al., 2003). The high degree of heterogeneity in the data collected indicates the need for careful analysis of the results obtained: the trees are made up of a large number of nodes, which are numerous even at the final level. The parameters set in the development of all the trees are illustrated in Table 10.

Analyzing these numerous nodes one by one would be very dispersive and would not lead to significant results. Therefore, once the tree is expanded, the most interesting objective categories must be selected, and then attention must be focused on those nodes characterized by a 100% gain. The gain represents the percentage of total cases of the node correlated with the objective category. A gain equal to 100% means that all the cases of the node examined have been assigned to the same category, indicating that the node is pure and the misclassification risk is void. Furthermore, even though lacking a 100% gain, the nodes that contain the greatest number of cases belonging to the objective variables also need to be consid-

**Table 10**
Characteristics of the algorithm

| Impurity measure: | Gini |
|---|---|
| *Stop criteria*: | |
| Maximum number of levels | 5 |
| Minimum number of cases for parent node | 3 |
| Minimum number of cases for child node | 1 |
| Minimum change of impurity level | 0.0001 |
| *Pruning rules*: | |
| Sub-tree selection | 1-SE Rule |
| Multiplier | 1.0 standard error |

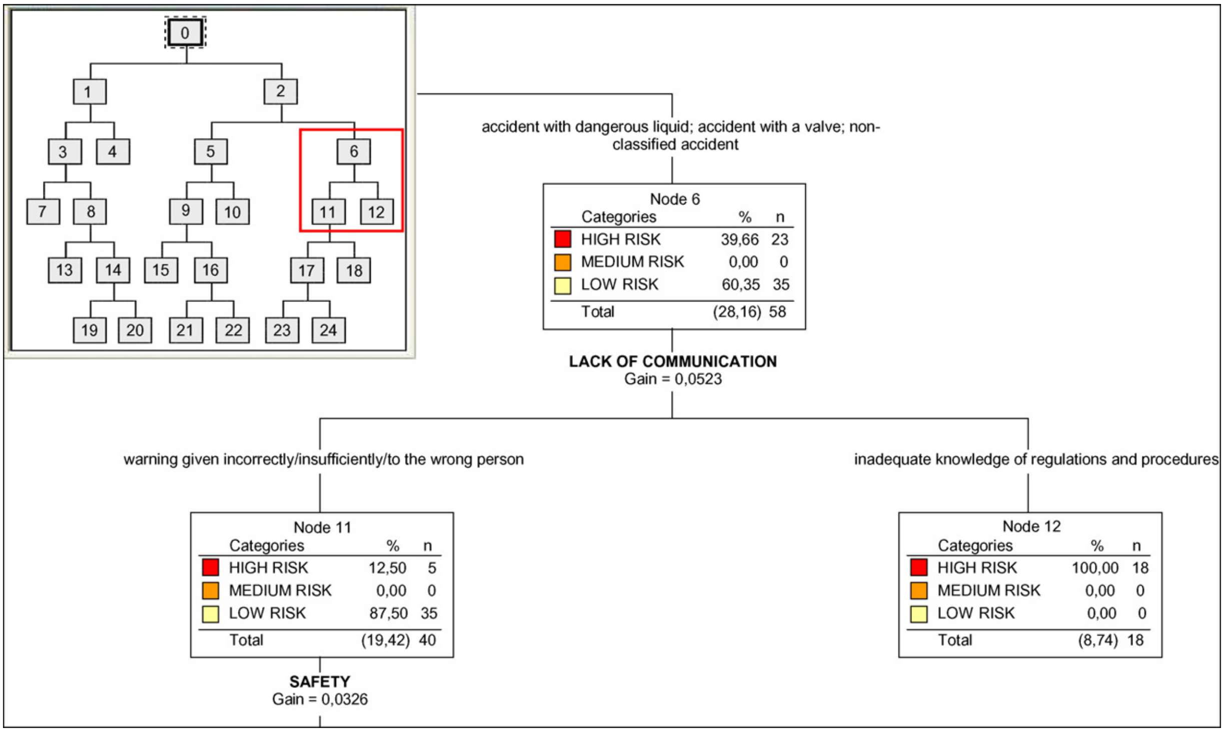


**Fig. 1.** Structure of the pruned tree for the “Risk Category” variable and illustration of nodes 6, 11 and 12.

ered. As an example, Fig. 1 shows the branch tree with the objective variable “Risk Category” as already described in Table 7. In this tree there are 2 nodes belonging to the category “Risk Category = HIGH” and with 100% gain: nodes 4 and 12, including 62.5% of total injuries with a high risk (25 cases out of 40). Particular attention must be addressed to node 19, containing 22.5% of the examined cases (9 cases of the total 40) with a gain of 69.23%. Once the most critical nodes have been selected, the decisional roles, which led to this classification, can be described and the results processed in order to identify the most dangerous work situations and their causes (Table 11).

Other information that can be derived from the decision trees is the hierarchy of the predictor variables. In general, the variable selected to split the first root node is of primary importance and should be carefully taken into account when ergonomic and safety management policies are defined. A complete overview of the results obtained from the trees in the example illustrates that:

- The injuries with a “LOW RISK” happen for every “part of body injured” and the immediate causes that are most important are below standard operating procedures such as “procedure not carried out”, “necessary operation was forgotten”, “PPE used badly/faulty”, or ergonomic causes such as “overcrowded area/limited movement possible” and “inadequate or excessive lighting”.

**Table 11**
Principal decision roles for the category “Risk Category = HIGH”

| |
|---|
| Node 4<br>IF (**Operating procedures below standard** is “incorrect/inappropriate use of equipment/appliances” OR “PPE used badly/faulty”)<br>AND (**Ergonomics** is “overcrowded area/limited movement possible”)<br>THEN<br>Gain = 100%<br>(7 cases) |
| Node 12<br>IF (**Operational procedures below standard** is “procedure not carried out” OR “necessary operation was forgotten”)<br>AND (**Type of accident** is “accident with dangerous liquid” OR “accident with a valve” OR non-classified accident)<br>AND (**Lack of communication** is “inadequate knowledge of regulations and procedures”)<br>THEN<br>Gain = 100%<br>(18 cases) |
| Node 19<br>IF (**Operating procedures below standard** is “incorrect/inappropriate use of equipment/appliances” OR PPE used badly/faulty)<br>AND (**Ergonomics** is lack of hygiene and tidiness/presence of obstacles)<br>AND (**Lack of communication** is “inadequate knowledge of regulations and procedures”)<br>AND (**Safety** is “danger of fire/explosions” OR “unsafe appliances/equipment”)<br>AND (**Type of accident** is “fall” OR “sudden movement” OR “mechanical damage” OR “collision” OR “crushing”)<br>THEN<br>Gain = 69.23%<br>(9 cases) |

- Even for the MEDIUM RISK objective category the most important cause is below standard operating procedures which, in addition to the previously mentioned procedures, include the extra heading: “incorrect/inappropriate use of equipment/appliances”. Immediate causes connected with safety also play a part in this class: “missing/wrong/inadequate safety conditions” and “dangerous transit/means of transport”.
- More serious injuries, which have a HIGH RISK, happen in every type of accident and in association with all kinds of task performed by the worker, but there are usually two different causes connected with these injuries: Lack of communication (Inadequate knowledge of regulations and procedures) and Non-use of Personal Protective Equipment. In a situation where no safety regulation violations are found, the human component plays a primary role: in fact, as has been seen from the analysis of the root causes (Table 8), accidents are mainly caused by negligence or misconduct of the workers themselves. However, a more careful analysis reveals another interesting element. In all these cases the “Ergonomics” variable is very important and one of the following categories is always present: “overcrowding of the area/limited possibility to move”, “lack of hygiene and tidiness/presence of obstacles”. The obvious conclusion which can be drawn is that even if the principal cause of the accident is the human component, “ergonomic” elements as well as respect for regulations and the correct training of the worker, all play an important role that must be controlled.

## 4. Discussion

The Occupational Injury analysis carried out in this paper was planned as a dynamic process and can be repeated systematically. The most relevant elements related to the objective variable considered in each case could be recognized from the diagrams obtained. As a result, control of the selected factors must improve and a great effort should be made to reduce the risk of injury.

To sum up the discussion of the RISK CATEGORY objective variable it is possible to state that:

1. As with the immediate causes, even with the root ones, the objective category which includes most cases is LOW RISK (109 cases out of 206, 52.9%) and the most important causes are “Human Factors” (*Bad habits, Physical/mental fatigue*) and Training *(Lack of general knowledge*).
2. HIGH RISK conditions appear to be dependent on the behavior of the workers and on their awareness, although the factor of safety in the workplace also plays a part since this sometimes not excellent. It is, therefore, necessary to focus more on respect for these regulations so as to reduce the risk factors as much as possible. Two factors are specifically relevant in this context and must be guaranteed: (1) respect for the safety legislation in force; (2) training and information for workers. The data often lack information and training activities to inform the workers about all Italian safety laws (i.e., Law 626/94) and about the correct working procedures which guarantee safer conditions.
3. The cause “Maintenance” does not appear in any of the 3 risk categories, indicating that this is not a decisive factor for the definition of the RISK.
4. The corrective action taken is mainly of the “Management” type: *internal assessment, evaluation of modifications* and *identification of extra PPE for the workers*. Legal action is not taken partly because, as already underlined, the accidents are limited to the medium–low RISK level and the workers or the contractor firms are not held to be the main parties responsible.

From an overview of the objective variable TYPE OF ACCIDENT the following conclusions can be drawn:

1. The objective category with the highest number of cases is *fall* with a total of 58 cases out of 206 (28.2%) followed by *accident with dangerous liquid* (32 cases). In both categories the immediate cause which proves to be most important is “Lack of communication”: *Inadequate knowledge of regulations and procedures* and *Warning given incorrectly/insufficiently/to the wrong person*. The parts of the body injured in these two categories are: “hand”, “arm”, “head”, “leg” and “chest”.
2. Of the root causes, “Factors connected with the task” are those most frequently found when studying the objective function TYPE OF ACCIDENT, and it is, therefore, necessary to intervene at a managerial level both to try to resolve the problems connected with processes and to modify the organization or the control of departments.
3. The predictor “Training” is present only in the objective categories “*accident with dangerous liquids*” and “*vehicle accident*” indicating that specific and general technical knowledge, training and communication influence these two types of accidents but not the others.

An overall analysis of the objective function PERIOD leads to these conclusions:

1. Different corrective action was applied in the decade analyzed: depending on the situation, in the 1st PERIOD all the types of corrective action were used with great attention paid to the management of procedures, for example by assessing the modifications and also the measures taken by the firm towards external firms or workers. In the 2nd PERIOD, for the nodes analyzed which have maximum gain, the predictors “planning” and “legal requirements” do not appear. This is connected with the types of accident which happened over the years, the responsibility for which was not to be attributed to badly functioning structures, and therefore, sanctions to external firms or warnings to workers were not necessary.
2. On the contrary, it is worth noting that two particularly significant causes disappear in the 10 year analysis: “Planning” (*lack of assessment/incorrect/insufficient planning*) and “Safety” (*missing/wrong/insufficient safety conditions*). This is a sign of improved management in the firm. Moreover some comforting results also emerge: the number of annual injuries has dropped, and the average index both of severity and risk is lower, a sign of less serious injuries on average. The fact that the API industry has obtained quality and safety certification OHSAS 18001 and ISO 14001 was undoubtedly an important step in the right direction. It clearly implies that any further action must involve a constant focus on various activities such as: activity planning, organization, documentation of procedures, staff training, application of the procedures, validation and re-examination of the procedures. Action will improve according to how much management is prepared to invest in the quality of the processes, in safety and in involving the workers in these processes.

The analysis carried out in this study was of great use to the refinery for providing feedback on the accident data collected over several years. If the information obtained from reports, drawn up after an accident and subsequently developed by the “Safety and Reliability Department”, is not periodically correlated and classified it risks becoming a mere legal requirement which does not provide possible guidelines for the future.

The decision diagrams obtained in this study allowed the Refinery to identify corrective action to prevent events with "Risk Category = HIGH". For example, the company tried to improve communication with workers, giving them adequate knowledge of regulations and procedures. Moreover, to mitigate the consequence of negligence or misconduct by the workers the most important action taken was a constant monitoring of the appropriate use of Personal Protective Equipment (PPE).

Concerning root causes, the refinery has to pay more attention to the "human factor". Human error caused by "lack of motivation" or "bad habits" is generally attributed directly to operator responsibility. During this study it was possible to notice that operator errors in many cases were related to managerial errors due to "lack of safety culture". Operator motivation for a positive approach to safety is the result of the personal involvement of general management in safety-related activities, in training programs, sponsoring the company internal objectives about safety and guiding the accident investigation towards problem-solving rather than towards finding the "guilty" party.

In the data collection phase it became clear that there is limited documentation about "Near-accidents". This heading refers to those events which have been a source of risk or danger, potentially provoking injuries and accidents; damage to the health of the workers or the population; damage to the environment; damage to company or third party property. There is a common tendency to avoid the near-accident procedure, so as to avoid tedious formalities, possible meetings, etc. On the contrary, together with a real introduction to these procedures, it is essential to encourage operators to report errors so that this type of communication becomes part of their normal routine.

As regards the problems related to the application of the proposed technique, the panel of experts created for this study worked for 3–4 months essentially to overcome difficulties in the data collection phases and in order to find those trees which are characterized by the best *risk values*. It should be mentioned that despite the initial wariness of the refinery staff the multidisciplinary approach was greatly appreciated by all the members of the panel. The definition of rules that can help "decision-makers" was possible only thanks to the combination of multidisciplinary skills. This solution allowed the panel: (1) to combine academics' theoretical knowledge with the day-to-day problems of a high risk company; (2) to investigate the events in depth so as to arrive at "all" possible causes; (3) to identify solutions that are really applicable to the production reality analyzed.

## 5. Conclusions

The aim of this study was to illustrate, by means of a practical example, a new application of tree classification methods in the field of occupational injury, where they have never been used, and to show the results and benefits. The application of these methods to accident risks in the API refinery has provided encouraging results: in particular the model generated using the CART method, related to several variables of interest, shows a classification accuracy which is on average greater than 75% (the *risk value* in the trees developed is always less than 25%), while for the objective variable RISK CATEGORY the results obtained are even better (accuracy 90%), proving that the most significant variables have been chosen and that the conclusions are sufficiently reliable. The residual percentage of error risk can be attributed to the not very high quality of the data sample and to possible errors in sample collection. The CART technique, which considers a very wide set of objective and predictive variables, shows new cause–effect correlations in occupational safety which had never been previously described, highlighting possible injury risk groups and supporting decision-making in these areas.

CART proved to be the most suitable and the most precise of the 4 algorithms tested for classifying the accident sample cases; the pruning algorithm and the standard error rule were particularly useful, allowing more compact and generalizable trees to be obtained with a minimum loss of information. However, some problems arose due to the method CART uses for selecting the variables for sub-dividing the nodes, because of a well-known distortion which tends to privilege those nominal predictors which are characterized by numerous categories. (Huang et al., 2006) For this reason a discussion of the results, in order to determine the variables of real interest was necessary.

In conclusion, this paper demonstrates that CART is a powerful alternative to the frequently used traditional parametric techniques. The large number of possible uses for this non-parametric approach is stressed, including all types of inquiry in any industrial field. CART is appropriate whenever the aim is to understand the problem being examined as thoroughly as possible in order to make the best decisions and choose the best ways of making improvements. The use of a classification and regression tree may also be considered a preliminary exploratory technique for evaluating the parameters influencing the most dangerous situations. However, it must not be seen as an attempt to supplant traditional statistical techniques, but as a complementary method to be integrated into this type of analysis.

## Acknowledgment

The authors would like to thank API refinery for collaboration to collect data concerning more than 200 cases of accidents which occurred in the decade 1994–2004.

## Appendix A

Four basic steps characterize the process of computing classification and regression trees: (1) Specifying the criteria for predictive accuracy; (2) Selecting splits; (3) Determining when to stop splitting; (4) Selecting the "right-sized" tree (Breiman et al., 1984).

The choices made in order to carry out each of these four steps are briefly reported below. The details of these decisions are beyond the scope of this paper but are explained at length in the standard reference for CART. CART starts by dividing an initial training data set into two sub-sets, so the cases in each sub-set are more homogeneous that in the original (single group) set. CART repeats the division for each child node, continuing recursively until the homogeneous level in the generic node required is obtained or a given stopping criterion is verified.

The following conditions will cause the algorithm to terminate:

(a) the maximum tree depth has been reached;
(b) no more splits can be made, because all terminal nodes meet one or more of the following three conditions: (b1) there is no significant predictor variable left to split the node; (b2) the number of cases in the terminal node is less than the minimum number of cases for parent nodes; (b3) if the nodes were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes.

The basic idea of splitting of the tree into successive levels is that each child node must be more "pure" than the original parent, where "pure" is a concept linked to the values of a given variable. As a consequence, in a completely pure node, all cases have the same value as the "splitting" variable. Several algorithms are available to measure the levels of 'impurity'. Examples of different measures of 'impurity' are the Gini and Twoing indices (for categorical objective variables), the ordered Twoing index (for ordinal categorical variables) and the LSD (for continuous objective variables). Since, in this application, the objective variables are all categorical, the Gini method is applied as the impurity measurement technique. Supposing that a branch $s$ subdivides one node $t$ in the two nodes $t_L$ and $t_R$ each having a proportion of cases, with respect to $t$, equal to $p_L$ e $p_R$, respectively, the *soundness of the subdivision* $\Phi(s;\ t)$ is defined as follows:

$$\begin{aligned}\Phi(s;t) &= g(t) - p_L g(t) - p_R g(t)\\ &= \sum_{j\neq i} p(i|t)\cdot p(j|t) - p_L\cdot\sum_{j\neq i} p(i|t_L)\cdot p(j|t_L)\\ &\quad - p_R\cdot\sum_{j\neq i} p(i|t_R)\cdot p(j|t_R)\end{aligned}$$

The Gini measurement is the measurement of impurity of a node and is commonly used when the dependent variable is a categorical variable, defined as:

$$g(t) = \sum_{j\neq i} p(j|t)\cdot p(i|t)$$

where $i$ and $j$ are categories of the objective variable.

The Gini Index is suitable for an interesting interpretation: to classify the objects contained in a node $t$, it uses a role which assigns a casually selected object in the node to a class $i$ with probability $p(i|t)$. The probability that such an object really belongs to class $j$ is $p(j|t)$. Thus, the probability estimation of misclassification depending on this rule is really represented by the Gini measurement of diversity $g(t)$. The best split (and, as a result, the corresponding dependent variable) is the one that maximizes the $\Phi(s;\ t)$ function which indicates the reduction of 'impurity' of the tree due to the split. CART represents a versatile method through which it is possible to consider the risk of classification error. In the analysis carried out misclassification costs have also been specified and *Risk values* have been corrected according to these costs. For the costs of misclassification variable the Gini measurement of impurity, according to the following formula is used:

$$g(t) = \sum_{j,i} C(i|j)\cdot p(j|t)\cdot p(i|t)$$

where the sum extends over all $k$ categories, $p(j|t)$ is the probability of category $j$ at the node $t$ and $C(i|j)$ is the probability of misclassifying a category $j$ case as category $i$.

Once the rules for the development of the tree have been decided it is necessary to deal with the problem of fixing the **right sizes** so as to obtain a more precise estimate of the real probability of classification error, that is to say the real risk of classification error. In general a greater number of sub-divisions leads to lower values of tree classification error risk, although a tree which is too large will be characterized by a real classification error risk which is greater than the risk found for a tree of the right size. On the other hand, a tree which is too small does not use some of the classification information available in the learning sample and will again produce a real error classification risk higher than that of a tree which has the right size. Excessive or uncontrolled growth of the tree, therefore, leads to a series of problems which can be summarized as follows:

- the final divisions may be superfluous, not providing useful information;
- excessive integration may result, whereby the tree integrates not only the real models present in the data, but also some errors present in the sample;
- the *Risk value* becomes progressively less accurate.

The *minimum cost-complexity criterion*, developed by Breiman et al. (1984), is used to try to create the smallest possible tree with a classification error risk which is not too much greater than that of the largest possible tree: it eliminates a branch if the cost associated with having a more complex tree exceeds the profit associated with having another level of nodes.

## References

Amell, T.K., Kumar, S., Rosser, B.W.J., 2001. Ergonomics, loss management, and occupational injury and illness surveillance. Part 1: elements of loss management and surveillance. A review. International Journal of Industrial Ergonomics 28, 69–84.

Amell, T.K., Kumar, S., Rosser, B.W.J., 2002. Ergonomics, loss management, and occupational injury and illness surveillance. Part 2: injury and illness incident profile. Sample data. International Journal of Industrial Ergonomics 29, 199–210.

Bevilacqua, M., Braglia, M., Montanari, R., 2003. The classification and regression tree approach to pump failure rate analysis. Reliability Engineering and System Safety 79, 59–67.

Biddle, E.A., Marsh, S.M., 2002. Comparison of two fatal occupational injury surveillance systems in the United States. Journal of Safety Research 33, 337.

Biggs, D.B., de Ville, B., Suen, E.A., 1991. Method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics 18, 49–62.

Breiman, L., Friedman, J., Olshen, R., Stone, C.J., 1984. Classification and Regression Trees. Belmont, CA, Wadsworth.

Brown, D., Stile, J.R., Gunderson, L., Giras, T.C., 2002. Mining human failure dynamics from accident data using logistic regression and decision tree. *Systems, man and cybernetics*. IEEE International Conference 6, 6–9.

Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research 36, 365–375.

Clemens, P.L., 2003. When is it time to worry about clustering? A Poisson distribution tutorial for SH&E professionals. Professional Safety (May).

Cowing, M.M., Cornell, M.E.P., Glynn, P.W., 2004. Dynamic modelling of the trade off between productivity and safety in critical engineering systems. Reliability Engineering and System Safety 86, 269–284.

Demichela, M., Piccinini, N., Romano, A., 2004. Risk analysis as a basis for safety management system. Journal of Loss Prevention in the Process Industries 17, 179–185.

Fabiano, B., Curro, F., Pastorino, R., 2004. A study of the relationship between occupational injuries and firm size and type in the Italian industry. Safety Science 42, 587.

Hertz, R.P., Emmett, E.A., 1986. Risk factor of occupational hand injury. Journal of Occupational Medicine 28, 36.

Hoyos, C.G., Zimolong, B., 1988. Occupational safety and accident prevention: behavioral strategies and methods. Advances in Human Factors/Ergonomics, 11. Elsevier, Amsterdam, pp. 218.

Huang, D.Y., Lee, R.F., Panchapakesan, S., 2006. On some variable selection procedures based on data for regression models. Journal of Statistical Planning and Inference 136, 2020–2034.

Kass, G., 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29 (2), 119–127.

Khan, J., Jansson, B., 2001. Risk level assessment and occupational health insurance expenditure: a gender imbalance. Journal of Socio Economics 30, 539–547.

Larsson, T.J., Field, B., 2002. The distribution of occupational injury risk in the State of Victoria. Safety Science 40, 419.

Lee, T.S., Chiu, C.C., Chou, Y.C., Lu, C.J., 2006. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis 50 (4), 1113–1130.

Lindell, M.K., 1997. Occupational safety and health inspection scores predict rates of workers' lost-time injuries. Accident Analysis & Prevention 29 (5), 563–571.

Loh, W.Y., Shih, Y.S., 1997. Split selection methods for classification trees. Statistica Sinica 7, 815–840.

Mingers, J., 1987. Rule induction with statistical data: a comparison with multiple regression. Journal of the Operational Research Society 38 (4), 347–351.

Piramuthu, S., 1999. Financial credit-risk evaluation with neural and neurofuzzy systems. European Journal of the Operational Research 112, 310–321.

Roudsari, B.S., Ghodsi, M., 2005. Occupational injuries in Tehran. International Journal of the Care of the Injured 36, 33.

S-RBI manual, 1999. Shell-Risk Based Inspection. Shell Global Solution International.

S-RCM Training guide, (1999). Shell-Reliability Centered Maintenance. Shell Global Solution International.

Schweigert, M.K., House, R.A., Holness, D.L., 1999. Occupational health and safety management systems in the Canadian pulp and paper industry: methods of

auditing. Journal of Occupational and Environmental Medicine 41 (10), 857–862.
Simo, S., 2005. Relationship between injuries at work and during leisure time. Accident Analysis & Prevention 37, 373.
Siu, O., Phillips, D.R., Leung, T., 2003. Age differences in safety attitudes and safety performance in Hong Kong construction workers. Journal of Safety Research 34, 199–205.
Siu, O., Phillips, D.R., Leung, T., 2004. Safety climate and safety performance among construction workers in Hong Kong. The role of psychological strains as mediators. Accident Analysis & Prevention 36, 359.
Smith, G.S., Veazie, M.A., Benjamin, K.L., 2005. The use of sentinel injury deaths to evaluate the quality of multiple source reporting for occupational injuries. AEP 5, 219.
Sorock, G.S., Courtney, T.K., 1997. Advancing analytic epidemiologic studies of occupational injuries. Safety Science 25, 29.
Sorock, G.S., Lombardi, D.A., Courtney, T.K., Cotnam, J.P., Mittleman, M.A., 2001. Epidemiology of occupational acute traumatic hand injuries: a literature review. Safety Science 38, 241.
Ture, M., Kurt, I., Kurum, A.T., Ozdamar, K., 2005. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications 29, 583–588.
Van Heel, K.A.L., Knegtering, B., Brombacher, A.C., 1999. Safety lifecycle management. A flowchart presentation of the IEC 61508 overall safety lifecycle model. Quality and Reliability Engineering International 15, 493–500.
Vorko, A., Jovic, F., 2000. Multiple attribute entropy classification of school-age injuries. Accident Analysis & Prevention 32, 445.
Vredenburgh, A.G., 2002. Organizational safety: which management practices are most effective in reducing employee injury rates? Journal of Safety Research 2, 259–276.
Waller, J.A., Skelly, J.M., Davis, H.J., 1995. Treated injuries in northern Vermont. Accident Analysis & Prevention 27 (6), 819–828.
Wang, K., Lee, A.H., Yau, K.K.W., Carrivick, P.J.W., 2003. A bivariate zero-inflated poisson regression model to analyze occupational injuries. Accident Analysis & Prevention 35, 625.
Wilson, L., 1998. Basic Learnings in Industrial Safety and Loss Management. Association of Professional Engineers, Geologists and Geophysicists of Alberta (APEGGA) and the Industrial Safety and Loss Management Program (ISLMP), Faculty of Engineering, University of Alberta, Edmonton, Canada, 72.
Yau, K.K.W., Lee, A.H., Carrivick, P.J.W., 2004. Modelling zero-inflated count series with application to occupational health. Computer Methods and Programs in Biomedicine 74, 47.