Reliability Engineering and System Safety 96 (2011) 739–747

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Explaining and predicting workplace accidents using data-mining techniques

T. Rivas [a,*], M. Paz [a], J.E. Martín [b], J.M. Matías [c], J.F. García [b], J. Taboada [a]

[a] Dpto. Ingeniería de los Recursos Naturales y Medio Ambiente, E.T.S.I. Minas, University of Vigo, Campus Lagoas, 36310 Vigo, Spain
[b] CIPP International, S.L. Parque Tecnológico de Asturias, Parcela 43, Oficina 11, 33428 Llanera, Spain
[c] Dpto. Estadística e Investigación Operativa, E.T.S.I. Minas, University of Vigo, Campus Lagoas, 36310 Vigo, Spain

ARTICLE INFO

Article history:
Received 19 January 2010
Received in revised form
28 February 2011
Accepted 2 March 2011
Available online 10 March 2011

Keywords:
Workplace accidents
Classification trees
Data mining
Bayesian networks
Support vector machines
Mine and construction safety

ABSTRACT

Current research into workplace risk is mainly conducted using conventional descriptive statistics, which, however, fail to properly identify cause-effect relationships and are unable to construct models that could predict accidents. The authors of the present study modelled incidents and accidents in two companies in the mining and construction sectors in order to identify the most important causes of accidents and develop predictive models. Data-mining techniques (decision rules, Bayesian networks, support vector machines and classification trees) were used to model accident and incident data compiled from the mining and construction sectors and obtained in interviews conducted soon after an incident/accident occurred. The results were compared with those for a classical statistical techniques (logistic regression), revealing the superiority of decision rules, classification trees and Bayesian networks in predicting and identifying the factors underlying accidents/incidents.

1. Introduction

A total of 922,253 workplace accidents resulting in lost workdays occurred in Spain in 2008; 194,248 occurred in the construction sector and 3255 occurred in the mineral extraction sector, representing falls of 23.3% and 14%, respectively, over 2007 [1]. The fall in 2008 compared to 2007 in terms of incidence (i.e., number of workplace accidents/population with social security accident insurance × 100,000) was 10.3%. This fall may be explained by the greater efforts of companies to ensure the safety of their workers by implementing preventive measures, whether on their own initiative or in response to legal obligations. Indeed, the fall may respond to Law 32/2006 governing construction sector subcontracting [2] and Royal Decree 1109/2007 deploying this law [3]. Subcontracting is a normal practice in Spain for construction and earth movement activities (in both the civil engineering and mining sectors), as it generally results in greater business efficiency. The new legislation was approved to ensure compliance with worker health and safety standards throughout the subcontracting chain. Royal Decree 1109/2007, passed in mid-2007 (and therefore in force when collecting data for this research) requires guarantees that ensure that any loss of control in the subcontracting regime does not result in health and safety risks for workers. The new legislation may be partly responsible for the fall in industrial accidents in the second half of 2007 and early 2008. The most recent statistics available, referring to the period April 2009–March 2010, reveal the same falling trend, both in the total number of accidents and in incidence: 4130.7 in 2009, a 18.5% lower than 2008.

Despite this fall, the issue of workplace safety continues to be a priority in social and economic policies and so requires in-depth studies that enable the causes of accidents to be accurately identified so that more effective measures and standards can be implemented.

Several statistical methods have been used in the workplace accident prevention field to process data. Analyses are usually descriptive, resulting in data in the form of historical summaries, percentages and indexes [4,5], or are based on linear models that evaluate the association between accidents and potential causes identified a priori ([6] and references therein). A priori, however, linear models excessively restrict how complex relationships between accidents and possible causes are modelled and may, in fact, fail to detect factors with a bearing on accidents if the relationship is non-linear.

The possibility for explaining and interpreting workplace accidents in terms of the entire range of possible causes is thus limited; this, in turn, conditions any working hypothesis aimed at predicting and reducing accidents. More sophisticated tools are thus needed that would enable full use to be made of information on accidents in terms of assessing dependence relationships between all the variables under consideration.

* Corresponding author. Tel.: +34 986811922; fax: +34 986811924.
E-mail addresses: trivas@uvigo.es (T. Rivas), mpaz.minas@gmail.com (M. Paz), jmartin@cippinternacional.com (J.E. Martín), jmmatias@uvigo.es (J.M. Matías), jgarcia@cippinternacional.com (J.F. García), jtaboada@uvigo.es (J. Taboada).

0951-8320/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2011.03.006

Data mining, which is an important discipline in fields such as medicine, engineering and finance, offers very positive results (see [7–9]). As for workplace risk management, studies have been conducted to assess the usefulness of such techniques in terms of their predictive power [10] and explanatory capacity [11]. Decision rules, for example, have been found to be particularly useful in identifying working conditions in the construction field that are associated with greater accident risk [12,13]. These studies, based on data from accident reports and interviews with workers and employers, confirm the advantages of data mining over conventional statistics in terms of the predictive function and the possibility of identifying interactions between variables with a bearing on accidents.

Our research aims to identify, from among a preselected group of methodologies, the data-mining techniques that extract the most useful information on workplace accidents from a database created from a survey of incidents/accidents in mining and civil construction companies, two sectors which, in Spain, head the list in terms of accidents involving lost workdays (data for 2008 from the Ministry of Labour and Immigration).

Our work is in a pilot phase aimed at evaluating the techniques available and contributing to the gradual development of a structured methodology for analysing workplace accidents that can eventually be safely applied to the design and planning of large-scale and far more costly studies. The potential of data-mining techniques not only derives from the possibility for processing large quantities of data but also from the following:

(1) their capacity to deal with large-dimension problems, which is necessary when endeavouring to identify relevant variables among a large number of potential factors;
(2) their flexibility in reproducing the data-generation structure, irrespective of complexity, thanks to a non-linear structure that is adaptable to the data (non-parametric philosophy);
(3) their great predictive and, in some cases, interpretative, potential.

In our research we evaluated Bayesian networks, decision rules, classification trees, logistic regression and support vector machines, with a view to ultimately reducing workplace accident rates by enabling preventive measures to be concentrated in an effective way in areas of greatest risk.

2. Materials and methods

2.1. Information sources and data description

The information used in this study was obtained from a survey carried out among workers employed in two companies—one in the mining sector and the other in the civil engineering sector—belonging to the same group and with operations in the Aragón, Asturias and Valencia regions of Spain.

In the mining company, workers from four operations were interviewed: an opencast coalmine employing 48 workers, two opencast ceramic-quality clay pits employing 72 workers and a quarry employing 13 workers and producing quartzite for use as a filler for public works and aggregates. The public works company surveyed, which removes earth in road, motorway and railroad (including high-speed train railroad) construction works, normally employs about 50 workers when working at full capacity.

Delivered and circulated in these companies, between September 2007 and March 2008, was a questionnaire with 20 questions to be completed whenever an accident/incident occurred. The questions covered issues related to the circumstances of the accident/incident, the worker, the kind of activity, work conditions and compliance with regulations. An incident was defined as any unexpected deviation from work procedures that might have caused an accident, and an accident was defined as a deviation from working standards or procedures affecting the health or safety of a worker [14].

A total of 62 completed questionnaires, each corresponding to a single accident/incident at different work stations and in different operations, were returned; of these, 18 referred to accidents and 44 to incidents. The severity of the accidents was not recorded. Each of the fields in the survey represented a study variable and, in total, information was obtained on 17 variables that were categorized in three groups referring to the event, the worker and the company (in terms of overall job and risk management). The variables for each category were as follows (see Table 1):

2.1.1. Event (three variables)

- Time of day (HRD): first thing (FH), after lunch (AL), last 2 h of work (LH), overtime hours (OT), or other (OTH).
- Day of the week (DAY): Monday (M), Tuesday (TU), Wednesday (W), Thursday (TH), Friday (F), or Saturday (S).
- Month (MTH): September (S), October (O), November (N), December (D), January (J), February (F), or March (M).

2.1.2. Worker (nine variables)

- *Worker age* (*AGE*): Below 27 (s1_below_27), 27–32.5 years (s2_27_32c5), 32.5–40.5 years (s3_32c5_40c5), or over 40.5 years (s4_40c5_up).
- *Worker nationality* (*NAT*): Native-born Spanish (SP), Eastern European (Armenia, Bulgaria, Poland and Romania) (EE), African (Libya and Morocco) (AF), or Latin American (Chile, Colombia and Ecuador) (LA).
- *Job type* (*JOB*): Job that involved handling vehicles (VH), job that involved handling heavy machinery (MC), office job involving no use of machines (OF), other job (electricians, mechanics, etc.) (OTH), or unspecified job (NS).
- *Length of time in the company (TCO)*: Less than 6 months (s1_below_0c5), 6–12 months (s2_0c5_1), 1–1.5 years (s3_1_1c5), 1.5–2 years (s4_1c5_2), or more than 2 years (s5_2_up).
- *Length of time doing the specific job associated with the accident* (*TJB*): Less than 1 week (s1_below_1W), 1–4 weeks (s2_1W_1M), or more than 4 weeks (s3_1M_up).
- *Accident risk training* (*ATR*) *received*: General training and training specific to the post (applied and practical on-the-job training, in accordance with Law 31/1995 [15]) (GS), general theoretical training (GT), or no training (NT).
- *Type of employment contract* (*ECT*): Temporary employment for a specific purpose (SP), temporary employment (TE), or permanent employment (PE).
- *Job-associated risk awareness* (*RAW*): The worker perceived the risk (YES), or the worker did not perceive the risk (NO).
- *Personal factors* (*PFA*): Personal factors contributed to the accident (YES), or personal factors did not contribute to the accident (NO).

Note that these last two variables are subjective, reflecting, in turn, whether the worker was aware of any risk prior to the accident/incident, and the undeniable influence on accidents/incidents of factors such as worker diligence, stress, boredom, etc.