Reliability Engineering and System Safety 108 (2012) 108–122

Contents lists available at SciVerse ScienceDirect

Reliability Engineering and System Safety

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

Finding occupational accident patterns in the extractive industry using a systematic data mining approach

Joaquim F. Silva [a], Celeste Jacinto [b,c,*]

[a] Centro de Investigação e Tecnologias de Informação (CITI). Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Campus de Caparica 2829-516, Caparica, Portugal
[b] Mechanical and Industrial Engineering-UNIDEMI research unit. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Campus de Caparica 2829-516, Caparica, Portugal
[c] Safety, Reliability and Maintenance Group of CENTEC, Technical University of Lisbon, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 4 November 2011
Received in revised form
26 June 2012
Accepted 22 July 2012
Available online 31 July 2012

Keywords:
Extractive Industry
Mining and quarrying
Occupational accident
Accident pattern
Eurostat data
Data mining

ABSTRACT

This paper deals with occupational accident patterns of in the Portuguese Extractive Industry. It constitutes a significant advance with relation to a previous study made in 2008, both in terms of methodology and extended knowledge on the patterns' details. This work uses more recent data (2005–2007) and this time the identification of the "typical accident" shifts from a bivariate, to a multivariate pattern, for characterising more accurately the accident mechanisms. Instead of crossing only two variables (Deviation x Contact), the new methodology developed here uses data mining techniques to associate nine variables, through their categories, and to quantify the statistical cohesion of each pattern. The results confirmed the "typical accident" of the 2008 study, but went much further: it reveals three statistically significant patterns (the top-3 categories in frequency); moreover, each pattern includes now more variables (4–5 categories) and indicates their statistical cohesion. This approach allowed a more accurate vision of the reality, which is fundamental for risk management. The methodology is best suited for large groups, such as national Authorities, Insurers or Corporate Groups, to assist them planning target-oriented safety strategies. Not least importantly, researchers can apply the same algorithm to other study areas, as it is not restricted to accidents, neither to safety.

1. Introduction

Since the establishment of the International Labour Organisation (ILO), in the early decades of the 1900 s, the collection of accident data and production of statistical analysis has always been a privileged source of accident information, from where to derive prevention and (international) resolutions concerning safety at work.

The aim of this work is to typify patterns of non-fatal accidents in the Portuguese mineral extractive industry, hereafter called simply Extractive Industry. This study continues and extends substantially a previous one by Jacinto & Guedes Soares in 2008 [1]. The novelty in 2008 was the ability to identify *accident patterns*, especially in the case of the so-called "typical accident"; such patterns were quite accurate and, above all, their relevance in terms of statistical association were clearly quantified at the level of each modality (or category) of the main variable, rather than simply associating the main variables themselves. This ability and the overall methodology applied at that time offered some novelty.

In the present work, the authors intend to go further and establish accident patterns, which are even more accurate and also more complete, as they now encompass multiple variables; or best said: include specific categories of multiple variables (i.e., modalities from the main categorical variables). The first study covered the triennium 2001–2003 of Economic Activity – Sector C (Mining & Quarrying; also referred to as Extractive Industry), whilst this one covers 2005–2007. Again, all the data was supplied fist-hand (raw data) to the authors directly by the competent authority, i.e., the *Office of Strategy & Planning* (GEP), which is the national agency responsible for collecting and coding all data on accidents at work.

The motivation for this second study was driven by the fact that availability of accident data is continuously increasing, partly because more countries are implementing the ESAW system (European Statistics of Accidents at Work), defined by the Eurostat in 2001 [2] and the 1998 ILO Resolution [3]. As stressed by Jørgensen in 2008 [4], who was for many years a leading member of the ESAW task-force, the statistical analysis of accidents at

* Corresponding author at: Department of Mechanical and Industrial Engineering, Faculty of Science and Technology, Universidade Nova de Lisboa, Campus de Caparica 2829-516, Caparica, Portugal. Tel.: +351 212 948 567; fax: +351 212 948 531.
E-mail addresses: jfs@fct.unl.pt (J.F. Silva), mcj@fct.unl.pt, mcjacinto@mar.ist.utl.pt (C. Jacinto).

0951-8320/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ress.2012.07.001

work constitutes an essential source of information to support the development of new prevention strategies. Hence, the higher is the availability of data, the higher is the need to explore new techniques and statistical tools for mining "hidden" details, which might help a better understanding of the phenomena; the main assumption is that understanding a phenomena is an essential condition to be able to control it.

The novelty in this work lies on two aspects: the multivariate facet of the findings (i.e., much more informative accident patterns) and, equally important, on the data mining approach developed to find such patterns. The referred technique not only permitted more information about patterns, but also enabled measuring (quantitatively) their cohesion.

2. Background review

Within the mining sector alone, there are abundant studies in the specialised literature concerning statistical analysis of accidents at work [5–19] some focus on accident causes – and sometimes causal patterns – whereas others report forecasting techniques for prediction of accident rates. However, the development of data-mining approaches to find causal patterns and to establish the statistical cohesion and significance of such patterns has not been seen that much.

Research on occupational accidents often uses generic descriptive statistics, which is adequate for describing and characterising accident data, but such approach holds limited usefulness for prevention. By contrast, inferential statistics provides a deeper insight of the phenomena, resulting in a stronger pillar for prevention. However, inferential studies are less common in the literature, at least when compared with descriptive ones. Inferential statistics may be used to find relevant associations between variables for explaining accident mechanisms, e.g.: [1,20–22], or for developing prediction models to forecast future trends, e.g.: [17,19,23].

The present work uses inferential statistics to find accident patterns based on multivariate associations. As mentioned, it extends the boundaries of a previous study [1], which already includes a review of the relevant safety literature in this economic activity (Extractive Industry). Nonetheless, this section outlines some pertinent new additions to the review, aimed at disclosing recent trends that are relevant to this work, especially the statistical modelling, which is a strong argument in this case.

The referred 2008 study draw attention to the usefulness of the ESAW variables and firstly established the term "typical accident", which was defined then, and now, as the *markedly most frequent modality* (or category) of accident. However, the statistically relevant "cause-effect" relationships had been quantified solely for the pair of variables Contact and Deviation; the first describes the type of accident (also called mode of injury), whereas the latter describes the immediate cause(s) of it. This quantified assessment was attained through a *R-coefficient*, which derives from the chi-square (χ^2) test and the previous paper explains in detail the mathematical reasoning behind the calculation. The focus was on the interrelationship between the several modalities of Contact and the modalities of Deviation, i.e., between the categories of both main variables.

The results obtained by this method revealed certain details of the cause-effect mechanism that were unknown before; Table 1 illustrates the characterisation of the "typical accident" for the non-fatal accidents.

From the table, one can verify the level of detail given by the approach used, but also its main shortfall: the pattern is restricted to the categories (or modalities) of only two main variables (Contact and Deviation). Despite the limitation, one can still argue that these two particular variables constitute the central part – or the nucleus – of any accident pattern.

The same approach, using the *R-coefficient*, was also successfully applied in a study [22] of fishing accidents to examine the dependency relationship between two modalities of two nominal variables. An alternative approach is the *Phi* (φ) *coefficient*, which is also based on a modification of the chi-square (χ^2). It also provides a metric for assessing the strength of pairwise associations and is explained either in a number of studies on accidents, e.g.: [20,21,24,25], or in general statistical textbooks, e.g.: [26–28]. The *Phi* (φ) *coefficient* also assesses both the strength and the direction [scale -1; $+1$] of significant associations between pairs of nominal variables; it holds the same advantages and limitations of *R-coefficient* but appears to be more difficult to apply.

These types of bivariate approaches, i.e., that measure associations of only two variables at a time, may constitute an "invitation" to infer associations by transitivity. It seems that this was the case with a study by Paul & Maiti [10], who claimed using multivariate analysis to establish the role of behavioural factors on safety management in underground mines; they have indeed used four variables in their work, but apparently they have measured associations of two at a time and then drawn conclusions for the whole set. In this context, transitivity means that if A&B have a statistically strong association, and if the same happens with the pair B&C, one may be tempted to assume that A&C are also strongly associated, which may not always be true. More importantly, yet, is that this type of inference does not allow one to calculate the strength of the apparent relationship between A and C. In other words: this kind of inference may be somehow abusive from a scientific point of view.

More recently, new tendencies are emerging in the safety literature; apparently the focus is shifting from the classical univariate or even bivariate approaches to multivariate analysis. PCA (Principal Components Analysis) and MCA (Multiple Correspondence Analysis), for instance, are becoming increasingly popular in the study of accidents, especially MCA, which deals with nominal (categorical) variables so common in accident data. Among the new attempts to apply MCA is the work of Rivière & Marlair in 2010 [29], who used this approach to identify incident typologies in the biofuel industry, or the findings of Factor et al. [30], also in 2010, concerning road/car accidents. Both studies were successful to disclose relevant associations of multiple variables, which in turn are so useful to establish accident patterns. Despite the novelty in the field, they also have their own limitations.

Table 1
The non-fatal "typical accident" and its cause-effect relationships (M&Q Sector, Portugal, 2001–03); adapted from a previous study [1].

Contact (category of accident)	Deviation (categories of probable causes)
Struck by object in motion, collision with (code C40) (~34% relative frequency in the period 2001–2003)	In 32% of the cases caused by D20 (overflow, overturn, leak, flow, vaporisation, emission); **R=2.6** (= very strong association) In 22% of the cases caused by D30 (breakage, bursting, splitting, slipping, fall, collapse of material agent); **R=2.2** (= very strong association) In 40% of the cases caused by D40 (loss of control); **R=1.3** (= strong association)