Bayesian network analysis of accident risk in information-deficient scenarios

Análisis mediante redes bayesianas de situaciones de riesgo de accidente en las que existe déficit de información

José Enrique Martín
Departamento de Ingeniería de los Recursos Naturales y Medio Ambiente
Universidad de Vigo, Lagoas Marcosende, 36310 Vigo (Spain)
jmartinsuarez@uvigo.es

Javier Taboada-García (Main Author)
Departamento de Ingeniería de los Recursos Naturales y Medio Ambiente
Universidad de Vigo
Lagoas Marcosende, 36310 Vigo (Spain)
jtaboada87@gmail.com

Saki Gerassis
Departamento de Ingeniería de los Recursos Naturales y Medio Ambiente
Universidad de Vigo
Lagoas Marcosende, 36310 Vigo (Spain)
ou_sk10@hotmail.com

Ángeles Saavedra (Corresponding Author)
Departamento de Estadística e Investigación Operativa, Universidad de Vigo
Lagoas Marcosende, 36310 Vigo (Spain)
saavedra@uvigo.es

Roberto Martínez-Alegría
Departamento de Enseñanzas Técnicas, Universidad Europea Miguel de Cervantes
C/Padre Julio Chevalier, nº 2, 47012 Valladolid (Spain)
rmartinez@uemc.es

Manuscript Code: 788
Date of Acceptance/Reception: 30.11.2017/08.07.2016
DOI: 10.7764/RDLC.16.3.439

Abstract
Analysis of accidents using Bayesian networks links certain predictor factors with other target factors representing types of accidents under study. Databases of real accident reports are typically used for both designing and training networks, which inevitably skews future inferences. Inferences are also limited because such databases do not usually include data on situations where accidents have not occurred. Inferences can thus be made about the occurrence of an accident, but not about specific types of accident. We describe a novel Bayesian network strategy for the field of occupational risk prevention which, extracting data from a database that includes situations where no accident has occurred, quantifies the influence and interactions of factors. It also allows particular accident types to be studied individually, thereby highlighting not only the correlation but also the causal relationship between work setting and accident risk.

Key words: Civil engineering, information deficit, Bayesian networks, workplace accident, model reduction.

Resumen
Cuando se analizan situaciones de accidentes de trabajo mediante el empleo de redes bayesianas, se relacionan determinados factores causales con otros factores finales o de predicción que representan los tipos de accidente objeto de análisis. Tanto para el diseño como para el aprendizaje de este tipo de redes suelen emplearse bases de datos de partes de accidentes que, inevitablemente, sesgará las futuras inferencias. Además, debido a que habitualmente las bases de datos no recogen situaciones en las que no ha acaecido un siniestro, el tipo de inferencias a realizar es también limitado. Así, es posible realizar una inferencia sobre el hecho de que ocurra un siniestro, pero resulta imposible inferir sobre cada tipo de accidente de forma individualizada. En este artículo se presenta una novedosa estrategia en el campo de la prevención de riesgos laborales cuando se trabaja con redes bayesianas. Mediante su implementación se resuelve la problemática de operar con bases de datos en las que no se contemplan casos en los que no ha acontecido un accidente. Esta nueva metodología aquí propuesta permite cuantificar la influencia de los factores causales y la interdependencia existente entre ellos. Además, hace posible el estudio individualizado de un determinado accidente mediante de redes bayesianas. Este último punto demuestra no solo la relación de correlación sino de causalidad entre entorno de trabajo y riesgo de accidente laboral.

Palabras clave: Obra civil, déficit de información, redes bayesianas, accidente de trabajo, técnicas de reducción de modelos.

Introduction

The European Union (EU) has, for decades, provided Spain with a natural framework for political and economic development. One consequence has been the creation of an EU legal acquis on the health and safety at work, which was transposed some 20 years ago into a Spanish law on occupational risk prevention. This law aims to promote worker health and safety through the implementation of procedures aimed at preventing risk in the workplace. Although efforts to date to reduce workplace accidents have been notable (López-Arquillos, & Rubio-Romero, 2015), the accident rate continues to be high. In the EU, and in Spain in particular, the construction sector has worryingly high workplace accident rates, much higher than for other sectors and particularly in terms of fatal accidents (Eurostat, 2016).

A key way to reduce accidents in any sector is to improve companies' management of risk prevention (Bird, Germain, & Clark, 2007). Analytical tools are crucial to achieving this goal. A vast amount of workplace accident data is available from surveys and interviews (of both employers and employees), accident reports, safety inspections, etc. Such data, suitably processed, can potentially provide crucial insights to the circumstances most likely to contribute to accidents. The development of models to numerically quantify accident risk on the basis of this data is a task of paramount importance. Bayesian networks have been used by different research groups (Green, Hjort, & Richardson, 2003) in studies as varied as analyses of accidents (Martín, Rivas, Matías, Taboada, & Argüelles, 2009 and Rivas, Paz, Matín, Matías, García, & Taboada, 2011), the prediction of natural disasters (Li, Wang, Leung, & Jiang, 2010) and accident risk assessment in steel constructions projects (Leu, & Chang, 2013).

In occupational risk settings, Bayesian networks have proven very useful in identifying links between factors when a full set of accidents is analysed jointly. However, in analysing individual accident reports, only the conditions affecting the employee at the time of a particular accident are revealed, and not the conditions in which an accident might have occurred. Consequently, individual inferences about specific types of accidents are not possible.

This article describes a methodology to address this problem based on the design and training of a Bayesian network specific to workplace risk prevention. This network describes situations in which accidents would be both very likely and very unlikely. This strategy, applied to processes of knowledge discovery in databases (KDD), is a novel application in the field of occupational risk prevention.

Description of the Problem

The database, covering six years of research, consisted of 417 workplace accidents corresponding to a Spanish group specializing in earthmoving operations.

Before modelling, and taking into account several starting working hypotheses, a number of factors were identified and defined from the available data.

To define the factors, various parameters of interest were selected for evaluation according to the different problems posed. Data were obtained from two sources: accident reports and face-to-face or telephone interviews with foremen or accident prevention experts involved in investigating accidents. From both these sources, information about n-factors associated with earthmoving operations was compiled, some generic (it could refer to any accident) and other specific to particular accidents. The factors were considered to have different subjectivity weights, depending on the information source. Thus, information from accident reports was considered to be objective, whereas information from personal interviews was considered to be more subjective and therefore implying some uncertainty as to its accuracy.

Although a total of 34 factors were included in this study of accidents associated with earthmoving operations, in the interest of illustrating our methodology, this number was reduced to 8: 5 predictor factors describing typical working environments and 3 target factors describing the probability of occurrence of common types of accident. The C4.5 algorithm, based on the simple divide-and-conquer algorithm for producing decision trees (Witten, Frank, Hall, & Pal 2016), was implemented in order to establish the factors showing a stronger relationship with the accident risk. The fully expanded decision tree was prune to eight factors with the aim of improving the understanding of the subsequent Bayesian network. The factors and their possible states are shown in Table 1.

Table 1. Predictor factors describing typical working environments and their possible states. Source: self-elaboration.

Predictor factors	States	
F02[1]. Shift duration	S1. 8 or fewer hours	S2. More than 8 hours
F05[2]. Conditions of (un)loading areas and haul roads	S1. Satisfactory	S2. Unsatisfactory
F11[3]. Outsoucing in the (un)loading area	S1. No	S2. Yes
F13. Order and cleanliness	S1. Satisfactory	S2. Unsatisfactory
F15. Works completion schedule	S1. Scheduled deadlines met	S2. Scheduled deadlines not met

[1]In both cases, independently of the number of daily shifts.
[2]There should be no overly steep slopes, blind corners, poorly banked surfaces, humps, potholes, loose grit where vehicles move, etc. and there should be adequate provision for maintenance
[3]Operations such as drilling and blasting, soil stabilization, the use of heavy machinery, water tanks, etc.

The implemented network analysed the links between the above 5 predictor factors and 3 possible accidents as follows: (1) Falls from the same or different height. (2) Falls of loose objects through detachment, collapse or handling. (3) Accidents involving vehicles.

Whenever analysing the probability of an accident an end node with two possible states was considered: S1: No. Ideal work conditions (not conducive to an accident situation). S2: Yes. Real accident situation.

Methodology

Bayesian network design and training

Bayesian networks model relationships between factors using directed acyclic graphs (Neapolitan, 2004 and Nielsen, & Jensen, 2009). A Bayesian network is model of probabilistic inference defined by a triplet (X,G,P) where $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_\mathrm{n})$ is the set of factors, G is a directed acyclic graph and P is a joint probability distribution. The nodes for G are labelled with the elements of X and the arcs of G represent the probabilistically conditional dependence relationship between nodes. This dependence relationship can be a probabilistic, causal or influential one, whereas the direction of the arc defines the probabilistic dependences which can be cause-effect or diagnostic ones (Russell, & Norvig, 2010).

Thus, each factor is associated with a conditional probabilistic table that define the probability of each state for that factor:

$$P(X_i = x|X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n) = \frac{P(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n|X_i = x)P(X_i = x)}{P(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)} \quad (1)$$

Using the set of factors described above, three networks were designed in order to analyse the probability of occurrence of each type of accident (falls from a height, falls of objects and accidents involving vehicles). The structure of these theoretical Bayesian networks were designed, with the input of an occupational risk prevention expert, in such a way as to not only reflect working conditions but also how these were interrelated.

To estimate the conditional probabilistic tables, a theoretical database was created including situations in which accidents were both very likely and very unlikely. That is, two types of initial conditions were considered: (1) Ideal conditions, reflecting risk prevention excellence, (2) unfavourable conditions, conducive of an accident situation.

The predictor factors, $(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_{\mathrm{n}-1})$, describe typical working environments and the target factor, X_n, represents a certain type of accident. Then the database is created with fictitious records as follows: if $(X_1, X_2, \cdots, X_{n-1})$ describes a risk environment, then $\mathrm{P}(\mathrm{X}_\mathrm{n} = \mathrm{Yes})$=1. Conversely, if $(X_1, X_2, \cdots, X_{n-1})$ describes a safe environment, then $P(X_n = No) = 1$. These networks reflected a theoretical risk prevention framework for earthmoving operations. They, therefore, yielded mathematical models that allowed us to conduct descriptive and predictive analyses for input data describing specific work settings, $(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_{\mathrm{n}\text{-}1})$. A study of conditional probabilities, $\mathrm{P}(\mathrm{X}_\mathrm{i} =$

$x|X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_n)$, for each predictor node enabled sensitivity studies to be conducted; thus, interactions between factors could be explored in depth in order to obtain a numerical quantification of sensitivity to changes in other factors and to identify the more influential factors in the network. Subsequently, conditional probabilities for the target node, $P(X_n = \text{Yes}|X_1,\cdots,X_{n-1})$, could be computed in order to make diagnoses regarding accident risk situations.

Developing a network of this kind that considers a range of scenarios is impossible using just data on real accidents; such databases are skewed by the fact that they only contain data on real accidents, which means that the networks trained from them are not capable of drawing inferences regarding ideal work conditions.

Inferences drawn from real accidents

In the final design stage, data from real accident reports were input into the network. Evaluation, within a theoretical network, of the actual conditions in which accidents occurred, $X_n = \text{Yes}$, enabled important conclusions to be drawn.

Thus, if an accident probability, $P(X_n = \text{Yes}|X_1,\cdots,X_{n-1})$, is obtained that is close to 1, we can conclude that the conditions indicate real risk and that the relationship between the factors defining the work environment is causal of the fact that an accident may occur. Another important conclusion is that changes in working conditions can be expected to have a significant preventive effect. Conversely, if the accident probability is close to 0 for real accident data included in the database, then the conclusion would be that the accident was random and that changing working conditions would not avoid risk.

As with any mathematical model, estimating accident probabilities is subject to error, possibly due to random causes that cannot be controlled or determined unequivocally. Thus, when using a theoretical network to evaluate the actual conditions of an accident, $X_n = \text{Yes}$, the difference $0 \leq 1\text{-}P(X_n = \text{Yes}|X_1,\cdots,X_{n-1}) \leq 1$ must be interpreted as a mismatch or error between the theoretical model (prevention framework) and the accident setting. In short, obtained for each theoretical network is a relative measure of the agreement between the theoretical prevention framework and actual risk.

Results

Different contexts and environments that verified the potential of the tool are described below. The examples focus on the analysis of 3 particular aspects of managing risk prevention in the workplace: falls from the same or different height, the impact of outsourcing and the impact of the works completion schedule.

Falls from the same or different height

One of the most common accidents in earthmoving operations is falls from the same or different height. Analysed below is how the prediction model can help minimize that risk. Figure 1 shows part of the Bayesian network obtained after including data for this type of accident in the theoretical network. Two factors, conditions of (un)loading areas and haul roads and order and cleanliness, were selected which, according to an occupational risk expert, directly affect the occurrence of such accidents.

The following conclusions can be drawn from the network output:

- Major deficiencies were evident for the conditions of the (un)loading areas and order and cleanliness (88% and 91% of cases, respectively), with ideal conditions only occurring in 12% and 9% of cases, respectively.
- The overall probability of a fall from the same or different height was 91%.

It should be recalled that cases of real accidents were included in the database. The high risk probability calculated by the network for these cases indicates that conditions were ripe for an accident to occur, thereby ruling out the hypothesis that the accident occurred by chance.

Figure. 1. Expert-structured Bayesian network for factors F05 (conditions of (un)loading areas and haul roads) and F13 (order and cleanliness) and falls from the same or different height in an earthmoving operations setting. Source: self-elaboration.

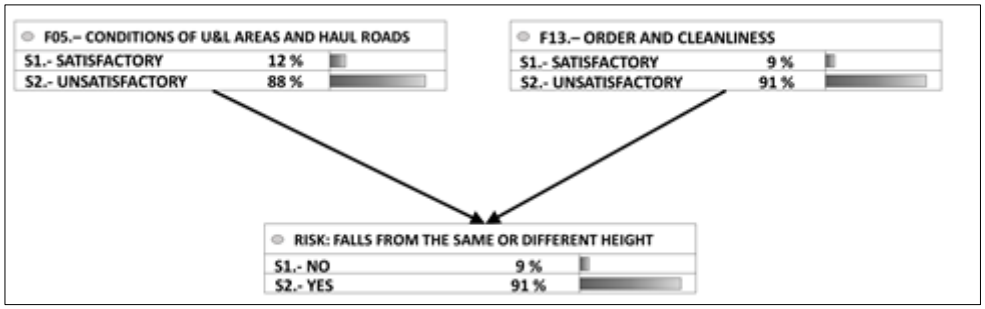
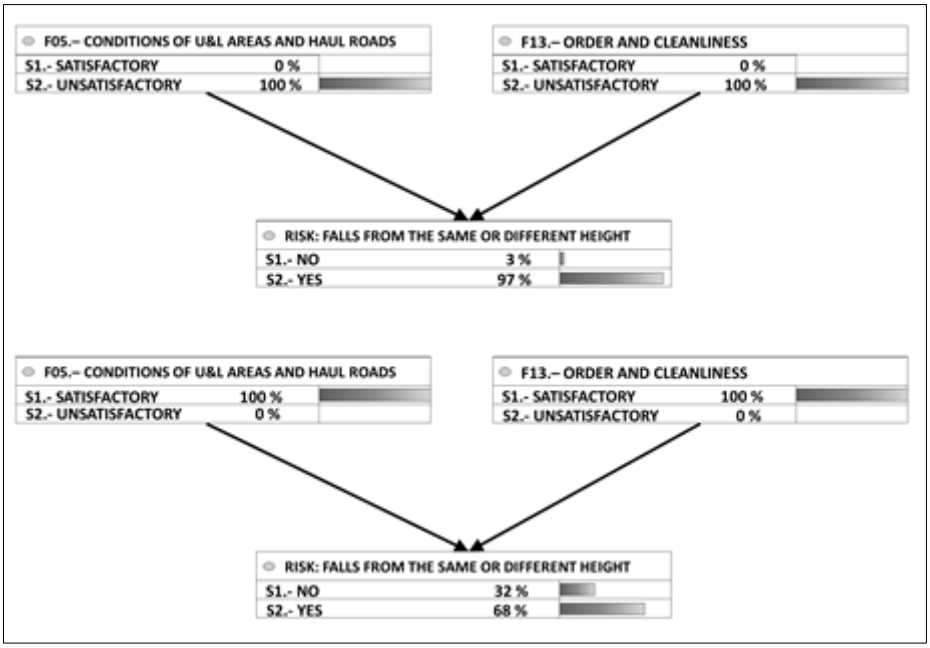


Figure 2 shows the network-predicted risk for situations in which the predictor factors F05 and F13 were absolutely unfavourable (top) and absolutely ideal (bottom). It can be observed that, for unfavourable conditions, the probability of a fall from the same or different height was 97%, an increase of 6% with respect to the situation described in Figure 1. When conditions were ideal, however, the probability of this type of accident fell to 68%.

Figure 2. Inference regarding the likelihood of falls from the same or different height for extreme states of factors F05 (conditions of (un)loading areas and haul roads) and F13 (order and cleanliness). Source: self-elaboration.



In conclusion, we could say that the network predicts that improving just 2 factors in the earthmoving context would lead to a reduction of 29 percentage points (from 97% to 68%) in the probability of falls from the same or different height.

Impact of outsourcing on accident rates

Given that Law 32/2006 (Spanish Government, 2006) limits excessive outsourcing in the construction sector, we analysed the impact of outsourcing on workplace accidents to see if this measure was justified.

Figure 3 (left) depicts a Bayesian network that links the risk of an accident involving a vehicle with 2 predictor factors: F11 (outsourcing) and F02 (shift duration). Our aim was, first, to analyse the impact of outsourcing and, second, to detect whether this affected shift duration at the time of the accident.

Figure 3. Left: expert-structured Bayesian network for F02 (shift duration), F11 (outsourcing in the (un)loading area) and accidents involving vehicles in earthmoving operations. Right: inference assuming a situation in which there is no outsourcing. Source: self-elaboration.

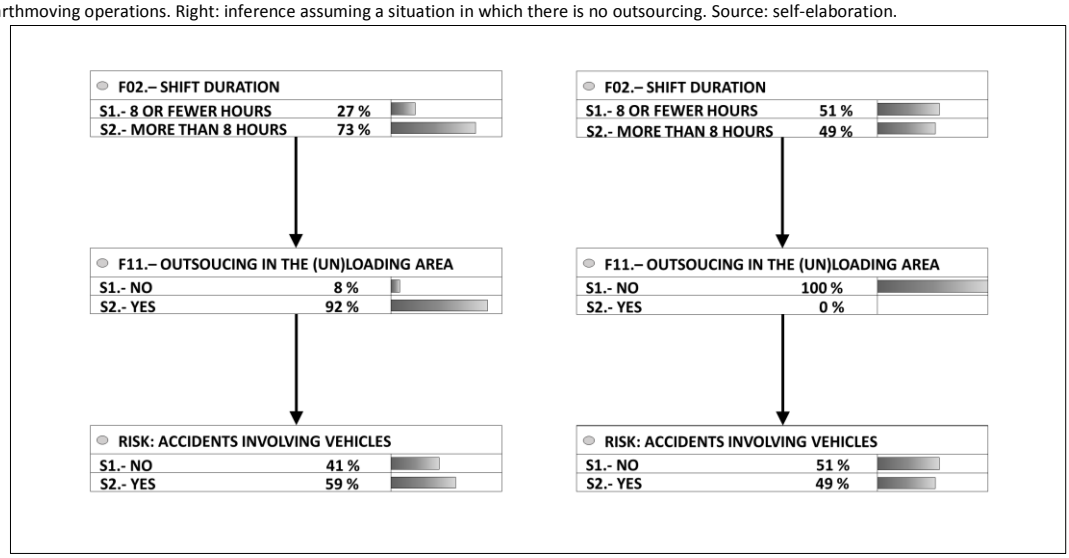


Figure 3 (left) shows that outsourcing in earthmoving operations was a standard procedure, recorded in 92% of cases, and also shows that shift duration was more than 8 hours in 73% of cases. In these conditions accidents involving vehicles occurred in 59% of the cases.

Under the premise of no outsourcing in the (un)loading area, we can infer new probabilities regarding both shift duration and the risk of accidents involving vehicles. Thus, Figure 3 (right) shows that the accident rate would drop from 59% to 49%. Moreover, shift duration would also be significantly affected, with a drop from 73% to 49% for workdays of more than 8 hours.

This simulation and the corresponding conclusions ratify the aforementioned Law 32/2006 which, inter alia, states that excessive outsourcing may result in practices incompatible with health and safety at work.

Impact on works completion schedule

Below we describe a simulation combining the factors analysed above, conditions of (un)loading areas and haul roads and order and cleanliness, with another factor, namely, the works completion schedule, along with a different type of accident accidents caused by falls of loose objects through detachment, collapse or handling.

Interpretation of the network shown in Figure 4 (up) results in the following conclusions:

- 97% of earthmoving operations fell behind schedule.
- Order and cleanliness was deficient in 95% of cases.
- The conditions of the (un)loading area were deficient in 79% of cases.
- Under these conditions, the probability of an accident involving loose objects was 75%.

The inference for an assumption of on-schedule completion is shown in Figure 4 (down). The network shows that F13 and F05 improved by 39 and 23 percentage points, i.e., from 95% and 79%, respectively, to 56%. Furthermore, the accident rate dropped to 55% compared to the behind-schedule scenario, 75%, depicted in Figure 4 (up). Furthermore, it was shown previously that outsourcing had a bearing on accidents involving vehicles.

To see whether outsourcing had any impact on falls of loose objects, we introduced F11 (outsourcing in the (un)loading area) in the network. The inference was performed on F13, F05 and falls of loose objects under the premise that the works were performed according to schedule and that no tasks were outsourced (Figure 5).

Figure 4. Left: expert-structured Bayesian network for F15 (works completion schedule), F13 (order and cleanliness), F05 (conditions of (un)loading areas and haul roads) along with the risk of falls of loose objects in an earthmoving operations setting. Down: inference for a situation in which works were on schedule. Source: self-elaboration.

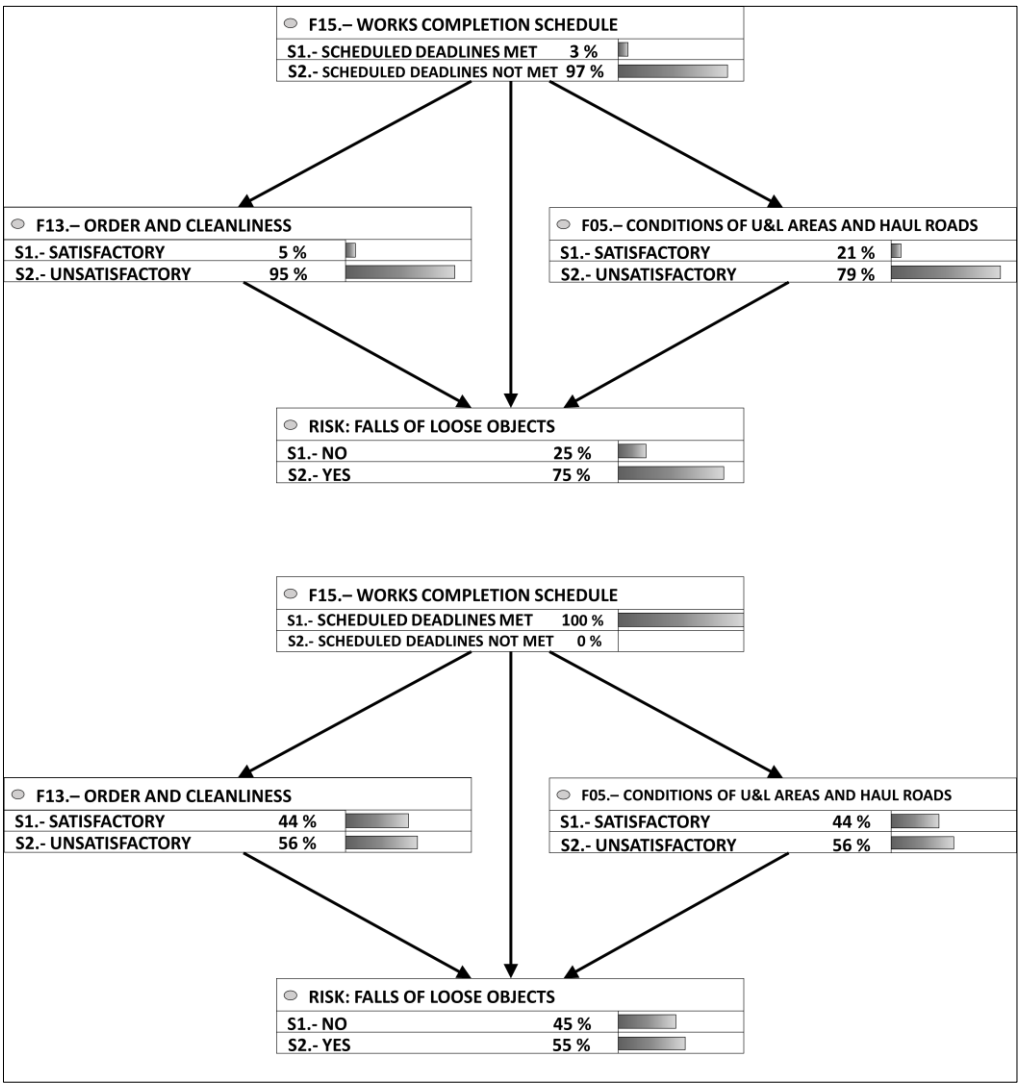


Figure 5. Inference F13 (order and cleanliness), F05 (conditions of (un)loading areas and haul roads) and falls of loose objects for a situation in which works were on schedule and there was no outsourcing. Source: self-elaboration.

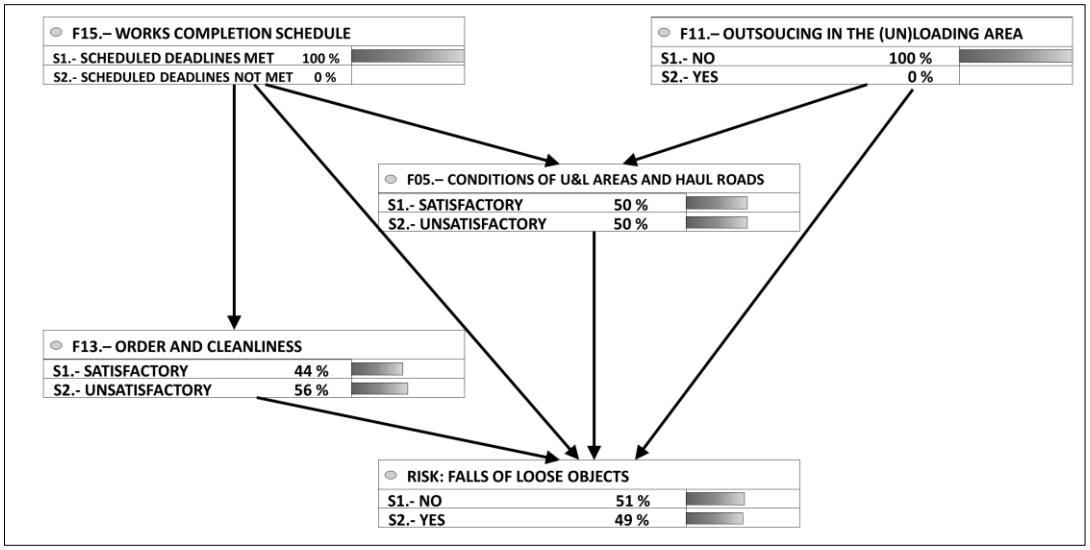


Figure 5 shows how the network predicts that accidents would decrease by a further 6 percentage points, i.e., from 55% (Figure 4 Right) to 49% (Figure 5), although with practically no impact on the other factors.

It can be observed that, depending on the work situation and working with different combinations, some factors have more weight than others in the ultimate accident risk. This fact is important because it enables the prevention-production binomial to be adjusted without reducing safety standards.

Conclusions

This article describes a study of occupational risk prevention in the construction industry, where accidents with serious consequences are frequent. We developed a theoretical Bayesian network using an occupational accidents database and deploying the knowledge of an occupational risk prevention expert. This theoretical Bayesian network was constructed considering both ideal and unfavourable work conditions in order to ensure a network trained for all kinds of situations.

Whereas research to date in this field has indisputably verified the correlation between work setting and accident risk, this methodology demonstrates the actual causal link between these two elements. Each time the theoretical Bayesian network assessed real situations as included in the accidents database, it invariably predicted accident risk probabilities above 90%. This high percentage is indicative not only of the causality implied by the work setting but also of the unlikelihood that an accident occurred due to uncontrollable factors or due to chance.

The network can also be used to predict the probability of different types of accident, i.e., it can show the influence of various factors on the risk of a certain type of accident and show how factors are interrelated. Another straightforward application of this methodology could be to a small or medium enterprise. Bayesian networks work with databases and, although they work optimally with large databases, this would not be a significant limitation. Of course, some adjustments should have to be made to adapt the network to the particular characteristics of the enterprise.

These results would suggest that this type of network can be reliably used as an analytical tool with potential for applications in other areas. Future applications could include evaluating changing environments in the construction or mining sectors, where it is difficult to assess work posts and conditions using traditional techniques.

References

Bird, F. E., Germain, G. L., & Clark, M. D. (2007). *Practical Loss Control Leadership*, Duluth, Georgia: Det Norske Veritas.

Eurostat. (2016). *Accidents at work statistics. Statistics Explained.* Retrieved July 8, 2016, from http://ec.europa.eu/eurostat/statistics-explained/index.php/Accidents_at_work_statistics.

Green, P. J., Hjort, N. L., & Richardson, S. (2003). *Highly Structured Stochastic Systems*. Oxford, England: Oxford University Press.

Leu, S-S., & Chang, C-M. (2013). Bayesian-network-based safety risk assessment for steel construction projects. *Accident Analysis & Prevention*, *54*, 122–133.

Li L., Wang J., Leung H., & Jiang C. (2010). Assessment of catastrophic risk using Bayesian network constructed from domain knowledge and spatial data. *Risk Analysis: An International Journal, 30*(7), 1157-75.

López-Arquillos A., & Rubio-Romero, J.C. (2015). Proposed indicators of prevention through design in construction projects. *Revista de la Construcción*, *14*(2), 58-64.

Martín, J. E., Rivas, T., Matías, J. M., Taboada, J., & Argüelles, A. (2009). A Bayesian network analysis of workplace accidents caused by falls from a height. *Safety Science*, *47*, 206-214.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Nielsen, T. H., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. New York, NY: Springer.

Rivas T., Paz M., Matín J. E., Matías J. M., García J. F., & Taboada J. (2011). Explaining and predicting workplace accidents using data-mining techniques. *Reliability Engineering and System Safety*, *96*,739-747.

Russell S., & Norvig P. (2010). *Artificial intelligence: a modern approach*. Englewood Cliffs, New Jersey: Prentice Hall.

Spanish Government. (2006). *Law 32/2006, of 18 October, regulating outsourcing in the construction sector*. Retrieved July 8, 2016, from, https://www.boe.es/boe/dias/2006/10/19/pdfs/A36317-36323.pdf.

Witten, I., Frank, E., Hall, M., & Pal, C. (2016*). Data Mining. Practical Machine Learning Tools and Techniques*. Burlington, MA, USA: Elsevier.