APEM journal
Advances in Production Engineering & Management
Volume 16 | Number 4 | December 2021 | pp 485–499
https://doi.org/10.14743/apem2021.4.415
ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Latent class analysis for identification of occupational accident casualty profiles in the selected Polish manufacturing sector

Nowakowska, M.[a,*], Pajęcki, M.[b]

[a]Faculty of Management and Computer Modelling, Kielce University of Technology, Poland
[b]Faculty of Management and Computer Modelling, Kielce University of Technology, Poland

A B S T R A C T

The objective of the analysis is identifying profiles of occupational accident casualties as regards production companies to provide the necessary knowledge to facilitate the preparation and management of a safe work environment. Qualitative data characterizing employees injured in accidents registered in Polish wood processing plants over a period of 10 years were the subject of the research. The latent class analysis (LCA) method was employed in the investigation. This statistical modelling technique, based on the values of selected indicators (observed variables) divides the data set into separate groups, called latent classes, which enable the definition of patterns. A procedure which supports the decision as regards the number of classes was presented. The procedure considers the quality of the LCA model and the distinguishability of the classes. Moreover, a method of assessing the importance of indicators in the patterns description was proposed. Seven latent classes were obtained and illustrated by the heat map, which enabled the profiles identification. They were labelled as follows: very serious, serious, moderate, minor (three latent classes), slight. Some recommendations were made regarding the circumstances of occupational accidents with the most severe consequences for the casualties.

A R T I C L E   I N F O

Keywords:
Manufacturing industry;
Occupational accidents;
Accident profiles identification;
Modelling;
Latent class analysis (LCA);
Cluster analysis;
Model selection

**Corresponding author:*
spimn@tu.kielce.pl
(Nowakowska, M.)

Article history:
Received 24 June 2021
Revised 5 December 2021
Accepted 6 December 2021

1. Introduction

Occupational accidents play an important role in the functioning of a society. They influence the costs of social security as well as the operating costs of both organizations and individuals. The accidents also affect the productivity, competitiveness and image of enterprises, and contribute to material and moral losses of casualties and their families. Work safety is a crucial issue tackled in various aspects. In general, it can concern ergonomics, in which the increase of effectiveness is indicated by designing workplaces or adapting work along with eliminating (among others) threats that increase the risk of developing employees' illnesses and injuries [1]. It is also included as an element of occupational health and safety in risk assessment in enterprises [2]. In a more detailed aspect, work safety is concentrated on the analysis of occupational accidents and their casualties, trying to discover the circumstances and causes of the events. To reduce accidents at work, mitigate their consequences, and shape a safe work environment, not only systemic actions and countermeasures resulting from the regulations but also various scientific studies are undertaken.

In the scientific literature, the subject has been considered for many years and it includes the analysis of accidents: (1) for specific cases (such as investigations of individual accidents or accidents at specific workplaces in a company, (2) in a variety of spatial aspects (region, country or group of countries), (3) having a certain profile, for example, of the same type or severity, (4) in connection with a specific sector of the economy, for example, construction or mining.

A significant part of the publication is devoted to the analyses of occupational accidents in the construction sector, and both the subject of the research and applied methods are very diverse. Müngen and Gürcanli [3], based on road construction fatal accidents data in Turkey from 1969-1999, using mainly the frequency analysis, indicated a significant share of fatal accidents in road accidents occurring in the road works zones. The authors drew attention to the fact that it was difficult to access the national level data. Rivas *et al*. [4] found that using traditional statistical methods in the investigation does not always allow the identification of actual cause-and-effect relationships as regards accidents at work, recommending the use of advanced analytical techniques, such as: decision rules, Bayesian networks, support vector machines and classification trees. They identified the most important causes of accidents and developed certain predictive models for accidents at work in construction (and additionally mining) companies. Cheng *et al*. [5], investigated the data set of 1,546 observations from 2000-2007 and indicated which factors are particularly responsible for the occurrence of occupational accidents in small construction companies in Taiwan. They used descriptive statistics, correlation analysis and ANOVA. Alizadehi *et al*. [6], in the set of 6722 records, identified 48 groups of workers suffering from accidents at construction workplaces in Iran. Using the Bayes' theorem, they estimated the posterior probability of the severity of the events under consideration. From a literature review, the authors concluded that in similar works on the construction industry, methods such as descriptive statistics, fuzzy inference and fuzzy logic, fuzzy sets, Bayesian analysis or ANOVA were used. Szóstak [7] obtained from the National Labour Inspectorate, 361 individual data records on accidents at work in five voivodeships in the Polish construction sector, for 2008-2014. Using non-hierarchical *k*-means cluster analysis, he defined profiles of occupational accident casualties characterized by qualitative and quantitative variables. Drozd [8], on the basis of one country region data from 2014-2016, analysed accidents in the Polish construction sector using the market basket analysis and defined typical association rules for such events. In turn, Berglund *et al*. [9] analysed and characterized accidents in the Swedish construction industry in 2016, calculating selected accident rates for individual trades. Ayhan and Tokdemir [10] undertook a task of developing a method for forecasting the consequences of accidents at work in the construction industry. They proposed a hybrid model combining the LCA cluster analysis and a supervised artificial neural network. The study was based on the collection of 4109 cases anonymously provided by construction companies located in the Euro-Asia region. Lee *et al*. [11] pointed out that various data mining methods are very often used in the analysis of data on occupational accidents in the construction industry. The authors proposed a research procedure for data pre-processing as well as supervised and unsupervised modelling of relationships between the characteristics of occupational accidents in the discussed economic sector. They used, among others: cluster analysis, chi-square test, V-Cramer test, support vector machine, and decision-tree-based ensembles. The elaborated tool was applied to analyse the set of 963 records from a large construction company in Korea. The possibility of using text mining techniques to analyse reports on occupational accidents in the construction sector was noted by Zhang *et al*. [12]. They examined and labelled (by the causes of the accident) 1954 documents describing events from 2003-2010. Keywords were extracted from the documents, which, together with the label, defined the data structure. This data set was the basis for building the accident cause classification models. A variety of mining techniques were used: support vector machine, *k*-nearest neighbours method, decision trees, logistic regression, and naive Bayesian classifier.

Similar methods of analyses are used in works devoted to scientific studies of occupational accidents in sectors other than construction. Palamara *et al*. [13] used a two-level approach involving Kohonen's SOM maps and *k*-means algorithm to identify the most common accident sequences leading to accidents in the Italian wood processing industry; data from 2002-2004 were analysed. The same SKM method (SOM and *k*-means) in the research of accidents in the same sector (based on the set of 1247 records) was used by Comberti *et al.* [14], however, focusing on the dynamics of accidents. The authors continued their work for the following years, for a larger data set [15] and in the aspect of a sensitivity analysis [16]. Moura *et al.* [17, 18] used unsupervised neural networks (including Kohonen's SOM) for the analysis of 238 serious accidents from technologically advanced industries (such as: aviation, chemical industry, refineries, petrochemical industry). In their opinion, the use of the proposed tools made it possible to reveal typical patterns and present them in the form of a graphic map, even though accidents are quite complex events, difficult to predict, and in which many different interactions take place. Verma *et al.* [19] analysed 843 occupational accidents registered in steel mills in India and identified patterns of such incidents using the association analysis. Ghousi [20], operating on the set of 1954 records from 2003-2010 and using *k*-means clustering method and the association analysis, investigated occupational accidents in production and industrial units in Iran to create a decision supporting system for managers. Sanmiquel *et al.* [21] identified the main factors influencing accidents at work in the Spanish mining industry. The authors analysed the set of 56,034 accident data records from 2005-2015, identifying association rules that define the causes of accidents based on a specific set of input variables. Farina *et al.* [22] used "learning by mistakes" and latent class analysis (LCA) methods to detect patterns of fatal occupational accidents in the injury dynamics aspect. They used data describing 354 events in enterprises of various industries in one of the regions of Italy (Piedmont) in 2005-2014. The LCA method was also used by Davoudi Kakhki *et al.* [23]. They identified occupational risk groups on the basis of the analysis of 1,031 serious injury events recorded in the Midwest US agribusiness industry in 2008-2016.

The researchers' focus on the construction sector may arise from the opinion that it has the most dangerous workplaces, although this judgment may be determined by a country specificity. However, work on a production line in wood processing companies also generates significant threats to the life and health of employees, since it involves the performing many dangerous actions, such as sawing, mechanical working, planning, cutting, laminating, whittling, and gluing. Nevertheless, work safety issues in the wood processing sector are discussed occasionally. In Poland, it is characterized by one of the highest risk indicators related to a work process. Hence, in the presented research, profiles of casualties of occupational accidents in production plants of the sector in question were identified. A qualitative data analysis tool was used, which is the LCA method. The presented work is the continuation and development of the authors' pilot study discussed as regards a chosen Polish region [24].

The work is an added value to a comparatively limited knowledge about occupational accidents in the wood processing industry. The most important elements for a scientific contribution are:

- determining the specificity of threats in manufacturing companies in the industry on a national scale, taking into account data from a long time horizon (10 years),
- proposing measures that support the assessment of the LCA model quality and proposing a method for the LCA model selection considering these measures,
- development of the discriminating ability index for the observed variables of the LCA model.

The article consists of five parts. After reviewing the selected literature in this chapter, the theoretical foundations of the latent class method are presented. Then, the data for the analyses are characterized and the preparation process for data modelling is described. The next chapter discusses the preliminary results, which are a set of models differing in the number of classes (clusters), and proposes a method of selecting the best one taking into account various criteria. In the next part, the latent classes of the selected model are described according to the observed variables characterizing occupational accident casualties, considering the importance of these variables. Finally, a summary and conclusions are presented.

2. Theoretical fundamentals of latent class analysis

Latent class analysis (LCA) is one of the cluster analysis methods used in the investigation of categorical variables in multivariate data. In the LCA model, a certain abstract qualitative variable LV, called a construct or a latent variable, is not directly observed (is hidden) but it reveals (manifests) its presence and intensity through other qualitative variables X_j, j = 1, ..., J, whose values can be observed. These observed variables are symptoms or indicators of the construct [25]. The purpose of the method is to identify disjoint homogeneous subsets (groups, clusters) in the data set. They are called latent classes and represent the values of the LV latent variable.

In each latent class K_c, c = 1, ..., C, each observation z in the Z multivariate data set has a value for the j-th observed variable. Assuming the independence of the observed variables within the classes, the probability of the product of the corresponding events is the product of the probabilities of these events. Therefore, the probability of occurrence in the class K_c of observation z, for which the vector of values of the observed variables $r(z)$ takes the value q, is the product of the probabilities of the occurrence of individual components of the vector q. Taking into consideration the above statement, the following form of the LCA model can be defined, which estimates the probability that in the z observation the $r(z)$ vector, representing the combination of values of the indicators $X_1, \dots, X_J$, has the value equal to q [25]:

$$\mathrm{P}(r(z) = q) = \sum_{c=1}^{C} \gamma_c \cdot \prod_{j=1}^{J} \mathrm{P}(r_j(z) = q_j |\, z \in K_c) = \sum_{c=1}^{C} \gamma_c \cdot \prod_{j=1}^{J} \rho_{q_j|c} \tag{1}$$

where: K_c is c-th latent class, c = 1, ..., C; C is the number of latent classes; γ_c is the probability of the c-th latent class, which is the probability of belongings of an observation to the K_c latent class: $\gamma_c = \mathrm{P}(K_c) = \mathrm{P}(z \in K_c)$, c = 1, ..., C $\left(\sum_{c=1}^{C} \gamma_c = 1\right)$; $\rho_{q_j|c}$ is the conditional probability, that j-th observed variable has the value of q_j in the K_c latent class: $\rho_{q_j|c} = \mathrm{P}\big(q_j|c\big) = \mathrm{P}\big(r_j(z) = q_j \big| z \in K_c\big)$; q_j is the value of j-th observed variable, $q_j \in R_j$; R_j is the set of values of j-th observed variable, j = 1, ..., J.

Non-zero probabilities $\mathrm{P}(r(z) = q)$ are usually estimated by the optimization of the $V(Z)$ negative log likelihood. The optimization for the model Eq. 1 is finding such values of γ_c and $\mathrm{P}(r_j(z) = q_j |\, z \in K_c)$ (= $\rho_{q_j|c}$) estimators that the $V(Z)$ function of the form:

$$V(Z) = -\sum_{q \in R} N(q) \cdot \ln\left(\sum_{c=1}^{C} \gamma_c \cdot \mathrm{P}(r(z) = q |\, z \in K_c)\right) \tag{2}$$

takes the smallest value with the following constraints:

$$\sum_{c=1}^{C} \gamma_c = 1 \tag{3}$$

$$\bigwedge_{c=1,\dots,C} 0 < \gamma_c \leq 1 \tag{4}$$

$$\bigwedge_{j=1,\dots,J} \bigwedge_{c=1,\dots,C} \sum_{q_j \in R_j} \mathrm{P}\big(r_j(z) = q_j \big| z \in K_c\big) = 1 \tag{5}$$

$$\bigwedge_{j=1,\dots,J} \bigwedge_{c=1,\dots,C} \bigwedge_{q_j \in R_j} \mathrm{P}\big(r_j(z) = q_j \big| z \in K_c\big) \in [0, 1] \tag{6}$$

where: $R = R_1 \times \dots \times R_J$; $q = (q_1, \dots, q_J)$; $N(q)$ is the empirical frequency in a contingency table cell defined by the q vector of values for the observed variables (in other words: the number of observations in the Z data set for which the vector of the observed variables $X = (X_1, ..., X_J)$ has values of the q vector).

The G^2 statistic is used to assess the quality of the estimated latent classes model [25]:

$$G^2 = 2\sum_{q \in R} N(q) \cdot \ln\left(\frac{N(q)}{\widehat{N}(q)}\right) \tag{7}$$

where: N is the Z data set size (number of observations); $\widehat{N}(q)$ is theoretical (expected) frequency in a contingency table cell defined by the q vector of values: $\widehat{N}(q) = N \cdot P(r(z) = q)$, $P(r(z) = q)$ is given by the relationship Eq. 1.

To compare the models differing in the number of latent classes, information criteria are used, in particular considering the size of the data set (N) and the complexity of the model (M) [26, 27, 28]:

- $CAIC = G^2 + M \cdot (\ln(N) + 1)$ – Consistent Akaike Information Criterion,
- $BIC = G^2 + M \cdot \ln(N)$ – Bayes Information Criterion,
- $ABIC = G^2 + M \cdot \ln((N + 2)/24)$ – Adjusted Bayes Information Criterion.

In the above formulas, M is the number of the estimated parameters of the model Eq. 1 equal to:

$$M = (C - 1) + C \cdot \sum_{j=1}^{J} (|R_j| - 1) \tag{8}$$

where: $|R_j|$ is the number of values of the X_j variable.

The smaller the values of the measures are, the better the assessment of the model is. With a small sample size, the importance of complex models (with a large number of latent classes and a large contingency table) is greatly reduced.

To build the LCA model in this work, the authors used the *Proc LCA* procedure of the SAS system [29] and their own calculations elaborated in the MS Excel environment.

3. Data for analysis

In Poland, data on occupational accidents for all professional activities, systematized according to the PKD section – Polish Classification of Activities (economic activity), are collected by the Central Statistical Office (GUS). Due to the function of supervising and controlling regulations as regards work health and safety, the National Labour Inspectorate (PIP) also collects information on occupational accidents investigated by PIP inspectors. This refers obligatorily to severe, fatal and group accidents. The GUS registers are digital and their structure complies with the structure of the national statistical accident card (defined by the regulation of the Minister of Labour and Social Policy of January 7, 2009 [30], amended in 2019 by the regulation of the Minister of Family, Labour and Social Policy of June 4, 2019 [31]). Each card contains details on a person injured in an accident, described by 29 features. Most of these features are descriptive, which implies the need to use qualitative data analysis methods.

The subject of the presented research are individual data records on accidents at work for 2008-2017, registered in production plants in Poland, that is in enterprises included in the C section according to the Polish Classification of Activities, obtained from the Central Statistical Office. Out of 24 branches of the C section, branch 16 was selected for the investigation. This is *Manufacture of wood and cork products, excluding furniture; manufacture of articles of straw and plaiting materials*. The companies in this branch produce such elements as: plywood, sawmill products, floor coverings, wooden packaging, veneer, and other carpentry articles. The nature of work of people employed in the production process generates very high accident risks [32]. In comparison with other branches of the C section, the accident rate (number of injured persons per 1000 employees per year), equal to 15.24, placed branch 16 above the 0.8 quantile. The accident casualty severity rate (the number of people injured in serious and fatal accidents per 10,000 employees per year), equal to 36.09, was the highest for this branch. Taking into account the above, the research was focused on the production process. In the collection of the data on accidents at work for branch 16, a subset that met the following criteria (the wordings in italics are taken from the statistical accident card) was selected:

- people injured in accidents at work: *Industrial workers and craftsmen*, *Operators and assemblers of machines and devices*, and *Employees doing simple works*,
- accident location: *Industrial production sites*,
- work process*: Production*, *processing*, *storage*.

For the need of calculations in the work, features characterizing the accidents casualties are marked with the symbol *Pxx*, where *xx* stands for the number of an item from the statistical accident card (for example, *P*02 means the age of the injured person). In a data pre-processing step, variables (indicators) were selected for the LCA model creation and outliers and observations that did not provide significant information were removed (for example, *P*09 – *Injured body part* = *Unknown or undefined*, *P*21 – *Activity performed at accident time = No information available*). Data transformation was proposed, mainly values or variables aggregation, which helped to solve the problem of rare categories. The resulting data set consisted of nine indicators and 13,750 observations. The characteristics of the set are presented in Table 1. Modifications made to the data are marked in italics. The numerical codes of the values (consecutive positive integers) and distributions for each observed variable are given.

Table 1 Characteristics of the research data

Indicators and their descriptive values	Value code	(%)
***P*02 – Casualty age**		
Up to 24 years old	1	17.30
25-34 years old	2	30.23
35-44 years old	3	25.43
45-54 years old	4	18.65
Over 54 years *Aggregation of original values*: (*55-59 years*) + (*over 59 years*)	5	8.38
***P*05 – Casualty occupation**		
Industrial workers, craftsmen, and employees doing simple works *Aggregation of original values*: (*Industrial workers and craftsmen*), (*Employees doing simple works*)	1	67.83
Operators and assemblers of machines and devices	2	32.17
***P*06 – Enterprise job seniority**		
Up to 5 years	1	67.64
6-10 years	2	15.97
Over 10 years *Aggregation of original values*: (*11-15*), (*16-20*), (*21-30*), (*Over 30 years*)	3	16.39
P*08* – Injury type		
Wounds and superficial injuries	1	55.96
Bone fractures	2	16.23
Displacements, dislocations, sprains and strains	3	11.19
Traumatic amputations (loss of body parts)	4	7.24
Various other injuries *Aggregation of original values*: (*Unknown or undefined*), (*Internal injuries*), (*Burns, frostbites*), (*Poisoning, infections*), (*Drowning, suffocating from lack of oxygen*), (*Effects of sounds, vibrations and pressure*), (*Effects of extreme temperatures, lighting and radiation*), (*Shocks*), (*Multiple injuries*), (*Another injury*)	5	9.39
***P*09 – Injured body part**		
Head, neck *Aggregation of original values:* (*Head*), (*Neck with cervical spine*)	1	7.01
Body *Aggregation of original values*: (*Thoracic and lumbar spine*), (*Torso and internal organs*), (*Whole body and its various parts*), (*Other body part*)	2	4.84
Upper limbs	3	67.31
Lower limbs	4	20.84

Table 1 Characteristics of the research data (continuation)

***P*21 – Activity performed at accident time**		(%)
Operating machinery	1	46.92
Working with tools and objects *Aggregation of original values*: (*Working with hand tools*), (*Handling objects*)	2	29.20
Transport at workplace *Aggregation of original values*: *(Driving means of transport / operation of moving machines and other devices), (Manual transporting)*	3	14.63
Being at accident scene *Aggregation of original values*: (*Moving about*), (*Presence)*	4	9.26
***P*26 – Material factor as injury source**		
Buildings, structures, surfaces *Objects as above and their elements including positions*: (*At ground level*), (*Below ground level*), (*Above ground level*)	1	6.83
Another factor *Aggregation of original values*: (*There is no material factor*), (*Supply, distribution and discharge systems for gases, liquids and solids, pipe networks, installations*), (*Equipment for the generation, processing, storage, transmission and distribution of energy*), (*Road vehicles*), (*Other transport vehicles*), (*Chemical, radioactive, explosive, biological substances*), (*Safety related devices and equipment*), (*Office equipment, personal equipment, sports equipment, weapons*), (*People and other living organisms*), (*Waste), (Physical phenomena and elements of the natural environment*), (*Another factor*)	2	9.43
Hand tools *Aggregation of original values*: (*Non-powered hand tools*), (*Hand-held or hand guided mechanized tools*)	3	9.32
Machines and devices *Aggregation of original values*: (*Portable or mobile machines and equipment*), (*Stationary machines, devices and equipment*), (*Machines, devices and equipment for lifting, carrying and storage*)	4	39.99
Materials, objects, products, machine parts	5	34.43
***P*27 – Main accident cause**		
Defect of material factor *Aggregation of original values*: (*Design defects or inappropriate technical and ergonomic solutions of material factor*), (*Improper manufacturing of material factor*), (*Material defects of material factor)*	1	16.38
Misuse of material factor *Aggregation of original values*: (*Inappropriate exploitation of material factor*), (*Employee's non-use or inappropriate handling of material factor*)	2	13.98
Inappropriate work organization *Aggregation of original*: (*Inadequate overall organization of work*), (*Inappropriate organization of a workplace*), (*Employee's failure to use protective equipment*)	3	11.67
Safety neglect *Aggregation of original values*: (*Employee's psychophysical state, not ensuring safe work performance*), (*Employee's inappropriate arbitrary behaviour*), (*Employee's misconduct*)	4	57.97
***P*289 – Casualty injury severity; a new variable, defined on the basis of the variables: *P*28 (accident consequence) and *P*29 (inability to work)**		
Minor accident resulting in inability to work for 0-29 days (up to a month) *Aggregation of original values*: (*Minor accident resulting in inability to work for 0-13 days*), (*Minor accident resulting in inability to work for 14-29 days*)	1	46.41
Minor accident resulting in inability to work for 30-89 days (from one to three months)	2	37.97
Serious accident *Aggregation of original values*: (*Severe or fatal accident*), (*Minor accident causing inability to work for more than 90 days*)	3	15.61

4. Selection of the LCA model

As in every cluster analysis method, also in the case of the LCA model, the decision on the number of latent classes is the key element. The following aspects were considered:

- relatively low values of information criteria, such as: *CAIC*, *BIC*, and *ABIC*,

- acceptable probability values of latent classes (which are weights determining classes support) – these values cannot be too low,
- the number of classes must not be too small to prevent the loss of information relevant to the definition of patterns,
- the number of classes must not be too large for the identified patterns to be distinguishable.

The decision on all the criteria is subjective and requires balancing, which means that a trade-off between fit and practical usability of the model should be considered. The latter aspect is related to the interpretability of the model – its suitability for distinguishing between classes. To facilitate the decision-making process, two measures were defined: the model balance ratio and the measure of the discriminating ability of the observed variable.

If, according to the assumption of the LCA model, the latent classes K_c and their probabilities γ_c define the distribution of the LV latent random variable that has C values (latent classes), then the entropy H of this variable is given by the relationship [33]:

$$H = -\sum_{c=1}^{C} \gamma_c \cdot \ln(\gamma_c) \tag{9}$$

Entropy has the greatest value when all probabilities are the same. In this case, H equals $\ln(C)$ and means that the importance of all latent classes is the same, which is considered to balance the model. The HB balance ratio is proposed as the quotient of the entropy H of the estimated LCA model and the entropy of the corresponding balanced model, taking into account the number of classes:

$$HB = \frac{H}{\ln(C)} \tag{10}$$

The closer to unity the value of HB is, the smaller the number of trivial latent classes (with small values of probabilities) are. The HB indicator, when combined with other measures, may support the selection of the LCA model.

The paper also proposes an index of the discriminating ability AR of the X_j observed variable. It is defined on the basis of the range of conditional probabilities of the X_j variable:

$$AR(X_j) = \frac{1}{|R_j|} \sum_{q_j \in R_j} Range(q_j|\{c\}) \tag{11}$$

where: $Range(q_j|\{c\}) = \max_{c=1,\dots,C} \{\rho_{q_j|c}\} - \min_{c=1,\dots,C} \{\rho_{q_j|c}\}$.

The algorithm leading to obtaining the AR value can be described as follows:

- the range of conditional probabilities assigned to the q_j value of the variable X_j with regard to the classes of the LCA model is calculated (there are as many probabilities as the latent classes for one q_j),
- the average of all the range values (there are as many ranges as the X_j variable values) obtained for the X_j variable creates the $AR(X_j)$ index.

The discriminating ability of an observed variable, valued with the AR measure, can be interpreted as an assessment of the distinction possibility of latent classes with respect to this variable. The maximum value of any range from formula Eq. 11 is equal to 1. It occurs when the conditional probability of the variable category in one class is 0 and in another class is 1. It means that in the case of perfect distinguishability of at least two classes, the AR measure is equal to 1. The AR value belongs to the interval [0, 1]. The nature of the measure indicates that the closer to unity its value is, the stronger the discriminating ability of the observed variable becomes. In this aspect, the AR measure can also be used as an importance weight of the observed variable in defining profiles as regards the phenomenon under investigation. Extreme cases, when $AR = 0$ or $AR = 1$, for real data almost never occur. In the work, it is assumed that an observed variable has a good discriminating ability if its AR exceeds 0.5.

A series of experiments with a different number of latent classes, from 1 to 16, was performed to determine an acceptable model. Considering that numerical methods are used in the calculations, for each model with a given number of classes, 20 estimates with different initial values for the iterative process were carried out. In the set of C-class models obtained in this way, a representative was selected – the model for which the function Eq. 2 had the lowest value. The final model was selected from among 16 representatives, using the previously described measures. The plots of values of *CAIC*, *BIC*, and *ABIC* information criteria, the *HB* balance ratio, and the $AR(X_j)$, i = 1, ..., 9 measures by each model (with a certain number of classes C = 1, ..., 16) are shown in Fig. 1, Fig. 2, and Fig. 3 respectively. The numbers of latent classes in the LCA model are marked on the horizontal axes (they also indicate the model numbers).

Fig. 1 was elaborated as a scree plot. It shows that the decline in the value of information criteria is relatively gentle, starting with the 7-class or 8-class model. It can be seen that increasing the number of latent classes above eight does not disrupt the linear trend almost parallel to the horizontal axis for all the three measures. The *HB* ratio graph is very diverse (Fig. 2), although the range of values is relatively small – it varies within the interval (0.9, 0.99). The three distinctive points on the polygonal chain suggest a very good balance for the LCA models with the number of latent classes equal to: 3, 7, and 8. According to Fig. 3, five observed variables form the group that plays the greatest role in distinguishing between latent classes: *Activity performed at accident time* (P21), *Injury* type (P08), *Injured body part* (P09), *Material factor as injury source* (P26), and *Casualty injury severity* (P289). Beginning with the 7-class model, all these variables have the *AR* value greater than 0.5. The following features are characterized by a worse discriminating ability: *Casualty age* (P02), *Casualty occupation* (P05) and *Main accident cause* (P27). They have the *AR* value less than 0.3 for the first eight models. The *Enterprise job seniority* (P06) observed variable is located between these two groups, starting from the 6-class model, and the discriminating ability of this indicator can be both high (AR = 0.63) and small (AR = 0.37), depending on the number of classes. Similar chart layouts characterize 6-class to 9-class models. Taking into account all the considered measures and, additionally, the insight into the estimated parameters, finally allowed selecting a 7-class model for further analysis, the qualitative assessment of which was better than that of other models. This model is identified by the *LCA*-7 symbol further in the study.

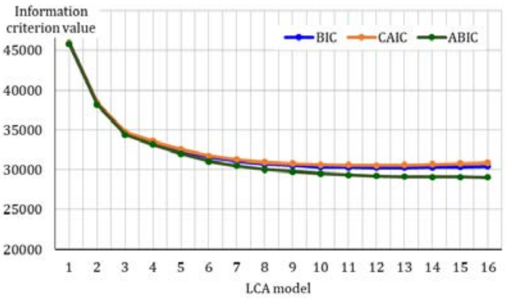

Fig. 1 Information criteria values by the number of classes in the LCA model

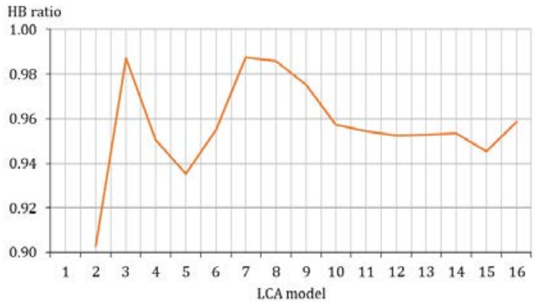

Fig. 2 *HB* balance ratio values by the number of classes in the LCA model

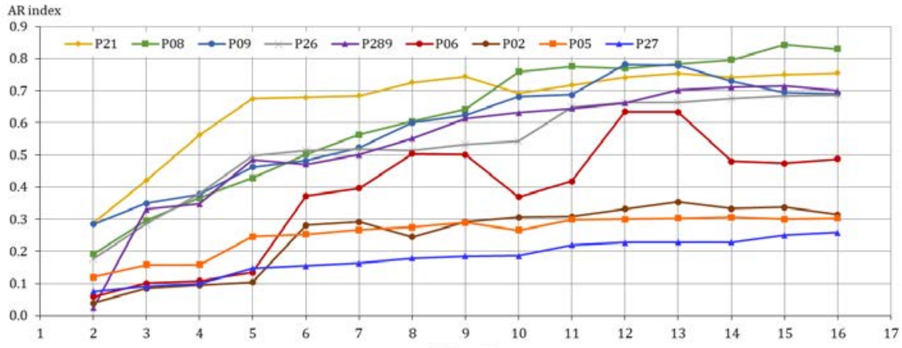

Fig. 3 AR index of observed variables by the number of classes in the LCA model

Table 2 presents the estimated statistics for the latent classes of the *LCA*-7 model. Classes support is quite satisfactory. In each case, the number of observations exceeds 1000, which means that, on average, during a year there are over 100 casualties of accidents at work characterized by the profile of a given cluster.

Table 2 Support for latent classes of the *LCA*-7 model

Latent class identifier	Class K_1	Class K_2	Class K_3	Class K_4	Class K_5	Class K_6	Class K_7
Latent class probability	0.13	0.17	0.12	0.18	0.09	0.14	0.17
Latent class size	1727	2381	1668	2480	1226	1956	2312

5. Results

A synthetic summary on the selected *LCA*-7 model is shown in Fig. 4. The resulting latent classes are illustrated in the form of a heat map. The map layout is defined according to the ordering of the analysed observed variables by their importance (assessed by discriminating ability), starting from the indicator with the highest value of the *AR* measure. The heat map column in front of the vertical line shows the distribution of the corresponding variables for the whole data set and it is the reference for the rest of the map. The cells behind the vertical line reflect the estimated conditional probabilities that the indicator (*P*21, *P*08, etc.) takes a certain value, provided that the observation (accident casualty) it characterizes belongs to a given latent class c ($c = K_1, \ldots, K_7$).

Indicator value	Whole data set	Class K_1	Class K_2	Class K_3	Class K_4	Class K_5	Class K_6	Class K_7
*P*21 - Activity performed at accident time; AR(*P*21) = 0.6847								
Operating machinery	0.47	0.30	0.86	0.00	0.62	0.10	0.28	0.72
Working with tools and objects	0.29	0.31	0.13	0.99	0.20	0.06	0.25	0.20
Transport at workplace	0.15	0.33	0.01	0.01	0.15	0.22	0.30	0.08
Being at accident scene	0.09	0.06	0.01	0.00	0.02	0.62	0.17	0.00
*P*08 - Injury type; AR(*P*08) = 0.5633								
Wounds and superficial injuries	0.56	0.04	0.39	0.86	0.80	0.12	0.56	0.87
Bone fractures	0.16	0.69	0.16	0.05	0.07	0.21	0.01	0.06
Displacements, dislocations ...	0.11	0.18	0.01	0.02	0.07	0.62	0.08	0.02
Traumatic amputations ...	0.07	0.03	0.37	0.01	0.00	0.00	0.00	0.02
Various other injuries	0.09	0.05	0.08	0.06	0.06	0.04	0.35	0.03
*P*09 - Injured body part; AR(*P*09) = 0.5226								
Head, neck	0.07	0.01	0.00	0.09	0.04	0.01	0.27	0.06
Body	0.05	0.06	0.01	0.01	0.01	0.03	0.24	0.00
Upper limbs	0.67	0.58	0.96	0.80	0.81	0.16	0.20	0.86
Lower limbs	0.21	0.35	0.02	0.10	0.14	0.80	0.28	0.07
*P*26 - Material factor as injury source; AR(*P*26) = 0.5173								
Buildings, structures, surfaces	0.07	0.08	0.00	0.03	0.01	0.48	0.07	0.00
Another factor	0.09	0.10	0.03	0.05	0.06	0.24	0.20	0.06
Hand tools	0.09	0.04	0.03	0.56	0.02	0.00	0.03	0.03
Machines and devices	0.40	0.18	0.82	0.00	0.53	0.18	0.19	0.58
Materials, objects, products, machine parts	0.34	0.60	0.11	0.37	0.38	0.09	0.52	0.33
*P*289 - Casualty injury severity; AR(*P*289) = 0.5012								
Minor accident, inability to work for 0-29 days	0.46	0.10	0.09	0.77	0.67	0.36	0.68	0.55
Minor accident, inability to work for 30-89 days	0.38	0.62	0.51	0.19	0.32	0.44	0.20	0.38
Serious accident	0.16	0.28	0.40	0.03	0.01	0.20	0.12	0.07
*P*06 - Enterprise job seniority; AR(*P*06) = 0.3962								
Up to 5 years	0.68	0.63	0.74	0.75	0.97	0.55	0.63	0.38
6 - 10 years	0.16	0.17	0.15	0.14	0.03	0.19	0.19	0.28
Over 10 years	0.16	0.20	0.12	0.10	0.00	0.26	0.18	0.34
*P*02 - Casualty age; AR(*P*02) = 0.29123								
Up to 24 years old	0.17	0.09	0.17	0.25	0.44	0.08	0.11	0.00
25 - 34 years old	0.30	0.28	0.29	0.36	0.45	0.29	0.28	0.16
35 - 44 years old	0.25	0.27	0.25	0.23	0.09	0.30	0.30	0.38
45 - 54 years old	0.19	0.25	0.20	0.12	0.02	0.23	0.21	0.32
Over 54 years	0.08	0.11	0.10	0.04	0.00	0.10	0.10	0.15
*P*05 - Casualty occupation; AR(*P*05) = 0.2667								
Industrial workers, craftsmen ...	0.68	0.68	0.78	0.67	0.69	0.52	0.61	0.71
Operators and assemblers of machines ...	0.32	0.32	0.22	0.33	0.31	0.48	0.39	0.29
*P*27 - Main accident cause; AR(*P*27) = 0.1632								
The defect of material factor	0.16	0.12	0.25	0.12	0.12	0.08	0.17	0.23
Misuse of material factor	0.14	0.19	0.15	0.19	0.15	0.03	0.10	0.14
Inappropriate work organization	0.12	0.15	0.07	0.08	0.10	0.17	0.20	0.08
Safety neglect	0.58	0.54	0.53	0.61	0.62	0.73	0.54	0.56

Fig. 4 The heat map of occupational accident casualty profiles for the *LCA*-7 model

6. Discussion

The individual profiles of the *LCA*-7 model were characterized on the basis of the obtained conditional probabilities and they are presented in Table 3. Descriptions are given according to the importance of the observed variables. If, for important indicators, the differentiation between the classes is small (like for the K_4 and K_7 classes), then the characteristics are completed according to other (less important) variables (in the case of the K_4 and K_7 classes, the difference is noticeable for the variables: *P*06, *Enterprise job seniority*, and *P*02, *Casualty age*).

Table 3 Occupational accident casualty profiles

	Class K_1 – Major limb injuries during manufacturing process or its service
P21	Any activity related to a production process or service of this process (transport) performed during the occurrence of an accident is almost equally possible: P(*P*21 = 1 or *P*21 = 2 or *P*21 = 3 | K_1) = 0.30 + 0.31 + 0.33 = 0.94.
P08	Bone fractures is the most common type of injury: P(*P*08 = 2 | K_1) = 0.69.
P09	Injuries affect limbs; more often the upper ones, less often the lower ones: P(*P*09 = 3 or *P*09 = 4 | K_1) = 0.58 + 0.35 = 0.93.
P26	Most injuries are caused by materials, objects, products, or machine parts: P(*P*26 = 5 | K_1) = 0.60.
P289	Accidents result in long or very long absence from work but also in serious or fatal injuries P(*P*289 = 2 or *P*289 = 3 | K_1) = 0.62 + 0.28 = 0.90.
P06	Mostly employees with the lowest job seniority are involved in accidents – up to 5 years: P(*P*06 = 1 | K_1) = 0.63.
	Class K_2 – Major upper limb injuries when operating machinery
P21	An activity performed by an employee is related to operating machinery: P(*P*21 = 1 | K_2) = 0.86.
P08	Body injuries on the one hand include wounds and superficial injuries (P(*P*08 = 1 | K_2) = 0.39), on the other hand – traumatic amputations (loss of body parts) (P(*P*08 = 4 | K_2) = 0.37).
P09	Upper limbs are injured: P(*P*09 = 3 | K_2) = 0.96.
P26	Machines and devices are the source of injuries: P(*P*26 = 2 | K_2) = 0.82.
P289	An accident leads to a serious injury or death of a casualty: P(*P*289 = 2 or *P*289 = 3 | K_2) = 0.51 + 0.40 = 0.91.
P06	In nearly three quarters of the cases, employees with the lowest job seniority are involved in accidents – up to 5 years: P(*P*06 = 1 | K_2) = 0.74.
	Class K_3 – Slight upper limbs injuries when working with tools or objects
P21	An activity performed by an employee is related to the use of tools and objects: P(*P*21 = 2 | K_3) = 0.99.
P08	Suffered harms refer to wounds and superficial injuries: P(*P*08 = 1 | K_3) = 0.86.
P09	Injuries affect mainly upper limbs: P(*P*09 = 3 | K_3) = 0.80.
P26	In most cases, injuries are caused by hand tools, less often - by materials, objects, products, or machine parts, possibly involving the work of hands: P(*P*26 = 3 or *P*26 = 5 | K_3) = 0.56 + 0.37 = 0.93.
P289	Accidents are minor, resulting in inability to work for no more than a month: P(*P*289 = 1 | K_3) = 0.77. There are almost no serious incidents.
P06	Mostly employees with the lowest job seniority are involved in accidents – up to 5 years: P(*P*06 = 1 | KU-3) = 0.67.
	Class K_4 – Minor upper limbs injuries to people professionally inexperienced during production process
P21	An activity performed by an employee is directly related to a production process; operating machinery or working with tools or objects: P(*P*21 = 1 or *P*21 = 2 | K_4) = 0.62 + 0.20 = 0.82.
P08	Suffered harms refer to wounds and superficial injuries: P(*P*08 = 1 | K_4) = 0.80.
P09	Injuries affect mainly upper limbs: P(*P*09 = 3 | K_4) = 0.81.
P26	In most cases, the injuries are caused by machines and devices, less often – by materials, objects, products, or machine parts, possibly involving the work of hands: P(*P*26 = 4 or *P*26=5 | K_4) = 0.53 + 0.38 = 0.91.
P289	Accidents are minor, resulting in inability to work for no more than three months: P(*P*289 = 1 or *P*289 = 2 | K_4) = 0.67 + 0.32 = 0.99. There are no serious incidents.
P06	Almost all casualties of accidents are employees with the lowest job seniority – up to 5 years: P(*P*06 = 1 | K_4) = 0.97.
P02	These are very young (up to 24 years old) and young (from 24 to 34 years old) people: P(*P*02 = 1 or *P*02 = 2 | K_4) = 0.44 + 0.45 = 0.89.

Table 3 Occupational accident casualty profiles (continuation)

	Class K_5 – Lower limb injuries of varying severity not directly related to production process
P21	An activity performed by a casualty during an accident is not directly related to a production process; most often this is being at the accident scene: P($P21 = 4$ | K_5) = 0.62.
P08	Displacements, dislocations, sprains and strains are the main injuries: P($P08 = 3$ | K_5) = 0.62.
P09	Injuries affect mainly lower limbs: P($P09 = 4$ | K_5) = 0.80.
P26	In nearly half of the cases, the factors behind the injuries relate to buildings, structures or surfaces: P($P26 = 1$ | K_5) = 0.48.
P289	Accidents vary in severity (minor and serious), most often resulting in inability to work from 1 to 3 months: P($P289 = 2$ | K_5) = 0.44.
P06	Slightly more than half of the casualties are employees with the lowest job seniority: P ($P06 = 1$ | K_5) = 0.55, albeit there is one quarter of people with the highest work experience – more than 10 years: P($P06 = 3$ | K_5) = 0.26.
P05	Compared to others, in class K_5, operators or assemblers of machines or devices constitute a significant share of the injured: P($P05 = 2$ | K_5) = 0.48 (nearly half of the observations).
P27	Work safety neglect is the main cause of an accident: P($P27 = 4$ | K_5) = 0.73.
	Class K_6 – Minor injuries of various causes
P21	An activity performed by an employee during the occurrence of the event related to a production process or a service of this process (transportation) is roughly equally possible: P($P21 = 1$ or $P21 = 2$ or $P21 = 3$ | K_6) = 0.28 + 0.25 + 0.30 = 0.83.
P08	Most frequent harms refer to wounds and superficial injuries: P($P08 = 1$ | K_6) = 0.56, secondly – various injuries (heterogeneous category – see Table 1): P($P\ 08 = 5$ | K_6) = 0.35.
P09	Any part of the body can be injured.
P26	In slightly more than half of the cases, injuries are caused by materials, objects, products, or machine parts: P($P26 = 5$ | K_6) = 0.52.
P289	Accidents are minor, resulting in inability to work for no more than a month: P($P289 = 1$ | K_6) = 0.68.
P06	Mostly employees with the lowest job seniority are involved in accidents – up to 5 years: P($P06 = 1$ | K_6) = 0.63.
	Class K_7 – Minor upper limbs injuries to people professionally experienced when operating machinery
P21	An activity performed by an employee is usually related to operating machinery: P($P21 = 1$ | K_7) = 0.72.
P08	Suffered harms refer to wounds and superficial injuries: P($P08 = 1$ | K_7) = 0.87.
P09	Injuries affect mainly upper limbs: P($P09 = 3$ | K_7) = 0.86.
P26	In most cases, injuries are caused by machines and devices, less often – by materials, objects, products, or machine parts, possibly involving the work of hands: P($P26 = 4$ or $P26 = 5$ | K_7) = 0.58 + 0.33 = 0.91.
P289	Accidents are minor, resulting in inability to work for no more than three months: P($P289 = 1$ or $P289 = 2$ | K_7) = 0.55 + 0.38 = 0.93.
P06	Job seniority in an enterprise is very diverse, the majority of which is over 5 years: P($P06 = 2$ or $P06 = 3$ | K_7) = 0.28 + 0.34 = 0.62. In the K_7 latent class, employees with the lowest job seniority occur sporadically.
P02	Very young casualties are absent, and the young ones are rare: P($P02 = 1$ or $P02 = 2$ | K_7) = 0.00 + 0.16 = 0.16.

The most serious accidents relate to operating machinery (class K_2). The weak link in this case may be the failure of machinery or equipment as well as the safety neglect by the employee. This means that the workstation at the machine, both due to the possible technical defects of the equipment and possibly the incorrect way the employee performs their work, may be under insufficient supervision. Therefore, special attention should be paid to provide adequate training regarding the workstation and the machine operated by the employee. Considering serious accidents (class K_1), the work organisation should be under special concern. In particular, it is important to carve appropriate transport routes, to provide a proper storage system, and to adapt a suitable efficient in-house transport equipment. In the case of these two most severe profiles of work accident casualties, it would be reasonable to introduce additional covers or other technical solutions as well as to use more effective ergonomic personal protective equipment to improve the protection of limbs.

As mentioned in the introduction section, there is a small number of publications concerning the analysis of occupational accidents in wood industry, particularly with the use of data mining methods. The four publications cited earlier concern Italy. Palamara *et al.* [13] and Comberti *et al.* [14] considered *the industry of manufacturing of furniture and building elements* that is a different production area from the one studied in this article. In both works, the same data set was used. Contrary to the research presented in this study, the Italian data did not contain information about fatal casualties. Clusters were identified considering sequences of events causing accidents defined by the *Activity*, *Deviation* and *Contact* variables. In the first work, "*The purpose is to discover the most common sequences of events leading to accidents for devising preventive actions*". Fourteen clusters containing sequences were identified. However, the authors did not discuss any profiles. They only stated that the most critical sequences are the loss of control and the incorrect movements during the work with manual tools. They considered working with machinery to be a less critical situation – this is different from the characteristics described in Table 3. The main purpose of [13] seems not to identify patterns but to analyze the effectiveness of the proposed clustering method. Comberti *et al*. [14] presented the same approach; they identified several clusters, but the discussion was concentrated on a coupled clustering methodology (SKM – SOM and *k*-means method). The purpose of the article by Comberti *et al*. [15] was the validation of the SKM method. Data from the wood processing industry and related to a selected region of Italy were used. They were described by six variables: *Activity*, *Deviation*, *Material of deviation*, *Contact*, *Injured body part*, *Age of worker involved*. The research resulted in obtaining 21 clusters of different dynamics (*Activity*-*Deviation*-*Contact*). However, the clusters were not discussed further on (that was beyond the scope of the study). The *Risk index* was defined for the evaluation of the clusters. Only two most critical clusters, according to the risk assessment, related to "manual work with hand-tools" and to "to falls during manual transport or movements", were indicated. In [16], Comberti *et al*. focused on improving the previously used methods. The description of obtained clusters was not given as the research had a methodological aspect (see the introduction section).

7. Final remarks and conclusions

Accidents at work constitute a significant social and economic problem, often causing serious bodily injuries or even death of a casualty. The identification of accident patterns may help in the development of effective tools leading to the improvement of work safety. The research was undertaken to identify patterns and describe profiles of people who suffered accidents at work. The analysis relates to accidents that occurred in connection with production processes in enterprises in Poland in 2008-2017 in the economic activity of branch 16 – *Manufacture of wood and cork products, excluding furniture; manufacture of articles of straw and plaiting materials* – of section C called *Industrial processing* (according to Polish Classification of Activities).

The LCA method, which divides the data set into groups called latent classes, was used enabling the definition of patterns of occupational accident casualties. With the exception of one, in all groups, accident casualties are primarily workers professionally inexperienced in a given job position; their percentage in individual patterns varies from 55 % to 97 %. One pattern relates to *very serious* cases (class K_2), including disability or death. It relates to upper limb injuries and concerns workers operating machinery. The second pattern describes *serious* accidents (class K_1) with long-term consequences. Accidents can arise at various stages of a production process or its service, be generated by various objects (things) of these processes and various activities of employees, and result in limb injuries. The *moderate* accident pattern (class K_5) includes workers not performing a production activity, who due to neglect of work safety (inattention, carelessness), suffered lower limb injuries. The pattern of *minor* accidents (class K_4) characterizes young people with the lowest job seniority, who operate machinery, devices or tools, and suffer from upper limb injuries. There is also a pattern of *minor* accidents (class K_6) difficult to characterize due to its heterogeneity; it includes various activities performed by an injured person, various injuries (not fractures, dislocations and amputations), and injury of any casualty body part. The *slight* accident pattern (class K_3) applies to people using hand tools. Upper limb injuries predominate, leading to a short absence from work. In one pattern (class K_7), professionally inexperienced employees are the least frequent and the youngest ones are absent. Here, wounds and superficial injuries are machine-driven, *minor*, and relate to upper limbs.

During the research, some tools supporting analytical work were proposed. A method was presented to facilitate the decision-making process as regards the number of classes of the LCA model so that the model quality and class distinguishability could be taken into account. A method of assessing the importance of the observed variables in the pattern description was also developed.

It is planned to expand the research scope in various aspects of identifying patterns of occupational accident casualties, considering an enterprise geographical location (a country region) and an industrial plant size. Developing the method of a systematic selection of indicators for the characterization of latent classes is also intended to include it to the whole analysis process.

References

[1] Leber, M., Bastič, M., Moody, L., Schmidt Krajnc, M. (2018). A study of the impact of ergonomically designed workplaces on employee productivity, *Advances in Production Engineering & Management*, Vol. 13, No. 1, 107-117, doi: 10.14743/apem2018.1.277.

[2] Vulanović, S., Delić, M., Kamberović, B., Beker, I., Lalić, B. (2020). Integrated management systems based on risk assessment: Methodology development and case studies, *Advances in Production Engineering & Management*, Vol. 15, No. 1, 93-106, doi: 10.14743/apem2020.1.352.

[3] Müngen, U., Gürcanli, G.E. (2005). Fatal traffic accidents in the Turkish construction industry, *Safety Science*, Vol. 43, No. 5-6, 299-322, doi: 10.1016/j.ssci.2005.06.002.

[4] Rivas, T., Paz, M., Martín, J.E., Matías, J.M., García, J.F., Taboada, J. (2011). Explaining and predicting workplace accidents using data-mining techniques, *Reliability Engineering & System Safety*, Vol. 96, No. 7, 739-747, doi: 10.1016/j.ress.2011.03.006.

[5] Cheng, C.-W., Leu, S.-S., Lin, C.-C., Fan, C. (2010). Characteristic analysis of occupational accidents at small construction enterprises, *Safety Science*, Vol. 48, No. 6, 698-707, doi: 10.1016/j.ssci.2010.02.001.

[6] Alizadeh, S.S., Mortazavi, S.B., Sepehri, M.M. (2015). Assessment of accident severity in the construction industry using the Bayesian theorem, *International Journal of Occupational Safety and Ergonomics*, Vol. 21, No. 4, 551-557, doi: 10.1080/10803548.2015.1095546.

[7] Szóstak, M. (2018). The application of cluster analysis to identify the occupational profile of people injured in accidents in the Polish construction industry, In: *Proceedings of VII International Symposium Actual Problems of Computational Simulation in Civil Engineering,* Vol. 456, Novosibirsk, Russian Federation, Article No. 012027, doi: 10.1088/1757-899X/456/1/012027.

[8] Drozd, W. (2017). Identifying and profiling the patterns of construction accidents using affinity analysis, *Technical Transactions / Czasopismo Techniczne,* Vol. 5, 15-24, doi: 10.4467/2353737XCT.17.065.6422.

[9] Berglund, L., Johansson, M., Nygren, M., Samuelson, B., Stenberg, M., Johansson, J. (2021). Occupational accidents in Swedish construction trades, *International Journal of Occupational Safety and Ergonomics*, Vol. 27, No. 2, 552-561, doi: 10.1080/10803548.2019.1598123.

[10] Ayhan, B.U., Tokdemir, O.B. (2020). Accident analysis for construction safety using latent class clustering and artificial neural networks, *Journal of Construction Engineering and Management*, Vol. 146, No. 3, Article No. 04019114, doi: 10.1061/(ASCE)CO.1943-7862.0001762.

[11] Lee, J.Y., Yoon, Y.G., Oh, T.K., Park, S., Ryu, S.I. (2020). A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry, *Applied Sciences*, Vol. 10, No. 21, Article No. 7949, doi: 10.3390/app10217949.

[12] Zhang, F., Fleyeh, H., Wang, X., Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques, *Automation in Construction,* Vol. 99, 238-248, doi: 10.1016/j.autcon.2018.12.016.

[13] Palamara, F., Piglione, F., Piccinini, N. (2011). Self-organizing map and clustering algorithms for the analysis of occupational accident databases, *Safety Science*, Vol. 49, No. 8-9, 1215-1230, doi: 10.1016/j.ssci.2011.04.003.

[14] Comberti, L., Baldissone, G., Demichela, M. (2015). Workplace accidents analysis with a coupled clustering methods: S.O.M. and K-means algorithms, *Chemical Engineering Transactions*, Vol. 43, 1261-1266, doi: 10.3303/CET1543211.

[15] Comberti, L., Demichela, M., Baldissone, G., Fois, G., Luzzi, R. (2018). Large occupational accidents data analysis with a coupled unsupervised algorithm: The S.O.M. K-means method. An application to the wood industry, *Safety*, Vol. 4, No. 4, Article No. 51, doi: 10.3390/safety4040051.

[16] Comberti, L., Demichela, M., Baldissone, G. (2018). A combined approach for the analysis of large occupational accident databases to support accident-prevention decision making, *Safety Science*, Vol. 106, 191-202, doi: 10.1016/j.ssci.2018.03.014.

[17] Moura, R., Beer, M., Patelli, E., Lewis, J., Knoll, F. (2017). Learning from accidents: Interactions between human factors, technology and organisations as a central element to validate risk studies, *Safety Science,* Vol. 99, Part B, 196-214, doi: 10.1016/j.ssci.2017.05.001.

[18] Moura, R., Beer, M., Patelli, E., Lewis, J. (2017). Learning from major accidents: Graphical representation and analysis of multi-attribute events to enhance risk communication, *Safety Science*, Vol. 99, Part A, 58-70, doi: 10.1016/j.ssci.2017.03.005.

[19] Verma, A., Khan, S.D., Maiti, J., Krishna, O.B. (2014). Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports, *Safety Science*, Vol. 70, 89-98, doi: 10.1016/j.ssci.2014.05.007.

[20] Ghousi, R. (2015). Applying a decision support system for accident analysis by using data mining approach: A case study on one of the Iranian manufactures, *Journal of Industrial and Systems Engineering*, Vol. 8, No. 3, 59-76.
[21] Sanmiquel, L., Bascompta, M., Rossell, J.M., Anticoi, H.F., Guash, E. (2018). Analysis of occupational accidents in underground and surface mining in Spain using data-mining techniques, *International Journal of Environmental Research and Public Health,* Vol. 15, No. 3, Article No. 462, doi: 10.3390/ijerph15030462.
[22] Farina, E., Bianco, S., Bena, A., Pasqualini, O. (2019). Finding causation in occupational fatalities: A latent class analysis, *American Journal of Industrial Medicine*, Vol. 62, No. 2, 123-130, doi: 10.1002/ajim.22936.
[23] Davoudi Kakhki, F., Freeman, S.A., Mosher, G.A. (2019). Segmentation of severe occupational incidents in agribusiness industries using latent class clustering, *Applied Sciences*, Vol. 9, No. 18, Article No. 3641, doi: 10.3390/app9183641.
[24] Nowakowska, M., Pajęcki, M. (2020). Applying latent class analysis in the identification of occupational accident patterns, *Scientific Papers of Silesian University of Technology – Organization & Management*, No. 146, 339-355, doi: 10.29119/1641-3466.2020.146.25.
[25] Collins, L.M., Lanza, S.T. (2010). *Latent class and latent transaction analysis: With applications in the social, behavioral, and health sciences*, John Wiley & Sons, New Jersey, USA, doi: 10.1002/9780470567333.
[26] Kim, S-Y. (2014). Determining the number of latent classes in single- and multiphase growth mixture models, *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 21, No. 2, 263-279, doi: 10.1080/10705511.2014.882690.
[27] Lin, T.H., Dayton, C.M. (1997). Model selection information criteria for non-nested latent class models, *Journal of Educational and Behavioral Statistics*, Vol. 22, No. 3, 249-264, doi: 10.3102/10769986022003249.
[28] Nylund, K.L., Asparouhov, T., Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study, *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 14, No. 4, 535-569, doi: 10.1080/10705510701575396.
[29] Lanza, S.T., Dziak, J.J., Huang, L., Wagner, A.T., Collins, L.M. (2015). Proc LCA & Proc LTA Users' Guide (Version 1.3.2), University Park: The Methodology Center, Penn State, from *https://www.methodology.psu.edu/*, accessed April 10, 2021.
[30] Ordinance of the Minister of Labour and Social Policy of 7 January 2009 on a statistical accident card at work, *Journal of Laws of 2009*, No. 14, item 80, as amended, from *http://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20090140080/O/D20090080.pdf*, accessed June 24, 2021.
[31] Regulation of the Minister of Family, Labour and Social Policy of 4 June 2019 amending the regulation on a statistical accident card at work*, Journal of Laws of 2019*, item 1106, as amended, from *http://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20190001106/O/D20191106.pdf*, accessed June 24, 2021.
[32] Pajęcki, M. (2020) Bezpieczeństwo pracy w sekcji przetwórstwa przemysłowego w Polsce – stan wypadkowości, In: Knosala, R. (ed.), Inżynieria zarządzania. Cyfryzacja produkcji. Aktualności badawcze 3, PWE, Poland, 1223-1232.
[33] Chakrabarti, C.G., Chakrabarty, I. (2005). Shannon entropy: Axiomatic characterization and application, *International Journal of Mathematics and Mathematical Sciences*, Vol. 2005, Article ID 234590, doi: 10.1155/IJMMS.2005.2847.