SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2020
ORGANIZATION AND MANAGEMENT SERIES NO. 146

APPLYING LATENT CLASS ANALYSIS IN THE IDENTIFICATION OF OCCUPATIONAL ACCIDENT PATTERNS

Marzena NOWAKOWSKA[1], Michał PAJĘCKI[2*]

[1] Kielce University of Technology, Faculty of Management and Computer Modelling; spimn@tu.kielce.pl,
ORCID: 0000-0002-6934-523X
[2] Kielce University of Technology, Faculty of Management and Computer Modelling; m.pajecki@tu.kielce.pl,
ORCID: 0000-0003-1593-3615
* Correspondence author

Purpose: The objective of the study is to use selected data mining techniques to discover patterns of certain recurring mechanisms related to the occurrence of occupational accidents in relation to production processes.
Design/methodology/approach: The latent class analysis (LCA) method was employed in the investigation. This statistical modeling technique enables discovering mutually exclusive homogenous classes of objects in a multivariate data set on the basis of observable qualitative variables, defining the class homogeneity in terms of probabilities. Due to a bilateral agreement, Statistics Poland provided individual record-level real data for the research. Then the data were preprocessed to enable the LCA model identification. Pilot studies were conducted in relation to occupational accidents registered in production plants in 2008-2017 in the Wielkopolskie voivodeship.
Findings: Three severe accident patterns and two light accident patterns represented by latent classes were obtained. The classes were subjected to descriptive characteristics and labeling, using interpretable results presented in the form of probabilities classifying categories of observable variables, symptomatic for a given latent class.
Research limitations/implications: The results from the pilot studies indicate the necessity to continue the research based on a larger data set along with the analysis development, particularly as regards selecting indicators for the latent class model characterization.
Practical implications: The identification of occupational accident patterns related to the production process can play a vital role in the elaboration of efficient safety countermeasures that can help to improve the prevention and outcome mitigation of such accidents among workers.
Social implications: Creating a safe work environment comprises the quality of life of workers, their families, thus affirming the enterprises' principles and values in the area of corporate social responsibility.
Originality/value: The investigation showed that latent class analysis is a promising tool supporting the scientific research in discovering the patterns of occupational accidents. The proposed investigation approach indicates the importance for the research both in terms of the availability of non-aggregated occupational accident data as well as the type of value aggregation of the variables taken for the analysis.

http://dx.doi.org/10.29119/1641-3466.2020.146.25

Keywords: occupational accident casualties, variable selection, classification probability.
Category of the paper: research paper.

1. Introduction

Accidents at work play an important role in the functioning of human communities, contributing, among others, to high social security costs, financial costs of organizational units and, most importantly, moral losses of casualties and their families. They also affect productivity, competitiveness and image of enterprises, regardless of their size (Macedo, and Silva, 2005). Therefore, it is necessary to limit all the losses by various activities, in particular by undertaking research leading to a safe working environment, including, in particular, the development of a strategy to reduce accidents and their consequences (Ejdys, 2010). Some researchers even present the concept of Zero Accident Vision (Zwetsloot, Leka, and Kines, 2017), which is a complete elimination of accidents.

In the worldwide scientific literature, the subject has been discussed for many years and it includes a range of occupational accidents (Cheng et al., 2010; Huang, and Hinze, 2003; Kakhki et al., 2019): (1) concerning specific cases (investigations of single accidents), (2) with a specific profile, for example accidents of the same type, (3) in connection with a specific sector of the economy, for example construction or mining, (4) in a global aspect like a region, country or group of countries. Actual data on accidents at work of varying degrees of granularity are considered, and the following methods belong to the most frequently used: statistical measure analysis, frequency analysis, ANOVA, and, from the group of data mining techniques, cluster analysis.

In Polish literature, the issue is undertaken relatively rarely, especially in the context of advanced analyses. In particular, the modeling of selected phenomena related to accidents at work in production plants is much more complex than it may result from the analysis of a single factory (Bogdan, and Boczkowska, 2009; Gajdzik, 2013; Wirkus, and Bajorski, 2017; Zjawin, and Kołodziej, 2018) or from the investigation of aggregated data made available by various data disposers, Statistics Poland (GUS) for example (Roszko-Wójtowicz, 2016; Węgrzyn, 2017).

In order to obtain reliable models concerning a specific aspect of occupational accidents, an advanced data mining analysis should be carried out on a set of detailed data, in which a data unit refers to a single object of observation – an accident or an injured person (like in the work of Szóstak (2018)). One of the methods that can be used for this purpose is the latent class analysis (LCA) presented in this article, owing to which it is possible to identify certain accident patterns.

2. Latent class model – theoretical fundamentals

Latent class analysis is used to model the relationship between categorical variables in multivariate data. The method provides easily interpretable results in the form of probabilities, which means that the above-mentioned relationships are assessed using quantitative measures. A latent variable is an abstract qualitative feature (variable, construct) that cannot be directly observed. It may take one of several values (latent classes) and reveals (manifests) its presence and intensity through other qualitative variables whose values can be observed. These observable variables are symptoms or indicators of an implicit trait. Using the concept of a diagram from path analysis, the model of latent classes can be presented graphically, as in Figure 1, where arrows indicate the direction of the influence of one variable on another variable (Ostasiewicz, 2012). The model assumes no error for the latent variable and the presence of errors for the indicators – in the figure they are denoted by ε_j, $j = 1, ..., J$.

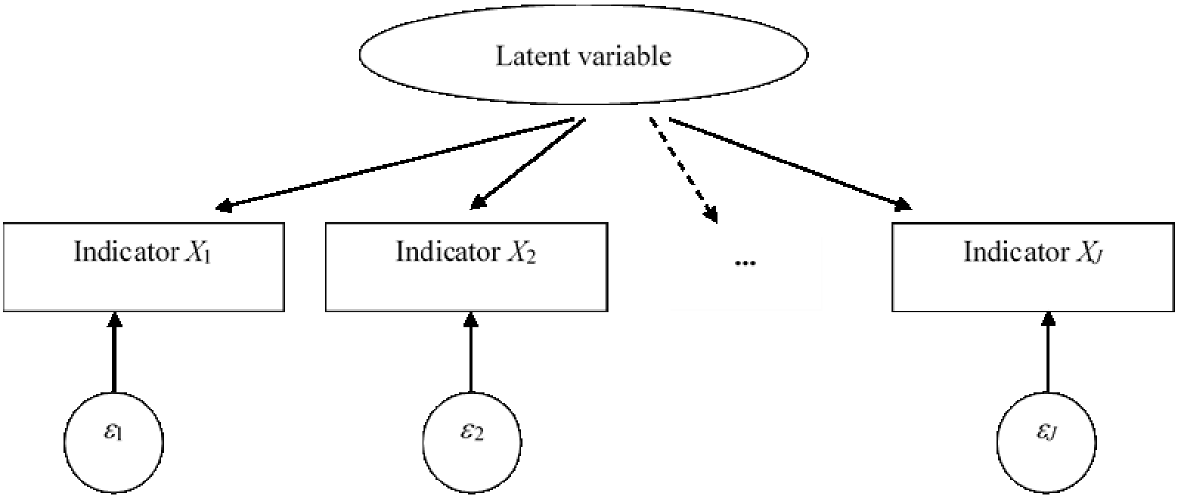

Figure 1. Scheme of dependencies between variables in the latent class model. Source: authors' own elaboration based on (Ostasiewicz, 2012).

In the latent class model, there are no restrictions related to the requirements of linearity, normality of data distribution or homogeneity of variance. It is assumed that each observation belongs to only one latent class c and the indicators within each class are independent (there is a local independence between the observable variables). It means that the conditional membership in a latent class is unambiguous, with observable variables being mutually independent, exhaustive and exclusive. It is also assumed that belonging to different latent classes is the only reason for the possible correlation between the indicators (Lavery, 2011).

Thus, there are the following objectives of applying latent class analysis in the research:

- finding (identifying) an appropriate number of C classes representing some hidden structures, the indicators of which are observable variables,
- defining relationships between observable variables and a latent variable (represented by latent classes) that can be represented by mathematical models.

2.1. Latent class model description

Latent class analysis concerns population units (objects) studied in the context of a specific issue. In this work, the objects are persons employed in production plants who have suffered from accidents at work, and the analysis refers to the characteristics of these individuals. The model is estimated from a sample of the population. The sample comes from a set of data on occupational accidents in Poland provided by the Statistics Poland (due to bilateral agreement). Observable variables are defined by selected qualitative variables characterizing the accident casualties, represented by the *J*-dimensional random vector $X = (X_1, ..., X_J)$. Each of the variables X_j has values in the set of the R_j categories. They constitute the contingency table, which is the basis for defining the model of latent classes (Frątczak, 2016). The number of cells in such a table is equal to $I = \prod_{j=1}^{J} R_j$.

By using latent class analysis it is possible to:

- identify homogeneous groups enabling the identification of profiles among the casualties of accidents at work,
- make group membership forecasts based on information about the values of $X_1, ..., X_J$ variables characterizing the individual.

The basis for building a model of latent classes is the assumption of local independence of indicators. It means that the probability of the product of the corresponding events is the product of the probabilities of the events. Therefore, the probability of the occurrence of the values (realizations) $x = (x_1, ..., x_J)$ of the vector of observable variables $X = (X_1, ..., X_J)$, provided that the object characterized by x belongs to class c is determined by the following relationship:

$$P(X = x\,|c) = \prod_{j=1}^{J} P(X_j = x_j\,|c) \tag{1}$$

where:

x_j – the value of j-th component of the X random vector (the value of the X_j variable).

Taking into account the law of total probability and the relationship (1), it is possible to determine the probability of the x value of the X vector of observable variables:

$$P(X = x) = \sum_{c=1}^{C} P(c) \cdot P(X = x\,|c) = \sum_{c=1}^{C} P(c) \cdot \prod_{j=1}^{J} P(X_j = x_j\,|c) \tag{2}$$

where:

$P(c)$ – the probability of an object belonging to latent class c.

Estimated model parameters are:

- the probabilities of latent classes: $P(c)$, $c = 1, ... C$ – their number is therefore equal to the number C of latent classes,
- the conditional probabilities of observing specific values of the indicator: $P(X_j = x_j\,|c)$. The number of these probabilities depends on the number of cells in the contingency table determined by all indicators $j = 1, ... J$. If the number of indicator categories j is equal to R_j,

then the number of the above-mentioned indicator conditional probabilities for this indicator is equal to: $C \cdot \sum_{j=1}^{J} R_j$.

Considering the issue above as well as the fact that in every latent class *c* each object must have some value for the variable X_j, the number *L* of independent parameters to be determined is equal to:

$$L = C - 1 + C \cdot \sum_{j=1}^{J} (R_j - 1) \tag{3}$$

Using the Bayes theorem, it is possible to re-estimate the probability of an object belonging to class *c* provided that the information about the values *x* of the *X* random vector characterizing this object is given, which means that it is possible to determine the a'posteriori probability:

$$P(c|\, X = x) = \frac{P(c) \cdot P(X = x\, |c)}{P(X = x)} \tag{4}$$

Latent class analysis is therefore applied in order to:

- determine the differences between classes by comparing the conditional probabilities *P(X=x |c)*, which can be used to profile and label groups,
- classify an entity with features defined by the manifested set of values (indicator values) to one of the latent classes based on the a'posteriori probability *P(c| X = x)*.

An ideal classification is obtained when the posterior probability *P(c| X = x)* is equal to 1 for one particular class. However, this is usually not the case. Then, the classification is made according to the highest value among the probabilities calculated from the formula (4) for the successive classes *c*, *c* = 1, ..., *C*, where the probabilities in the numerator and denominator are replaced by the corresponding estimators.

The maximum likelihood method is used to estimate the parameters of the latent class model (Lavery, 2011; Frątczak, 2016).

2.2. Model assessment

The quality of the estimated model of latent classes is assessed on the basis of measures derived from statistics G^2 (Frątczak, 2016):

$$G^2 = 2 \cdot \sum_{r=1}^{I} n_r \cdot \ln\left(\frac{n_r}{\hat{\mathrm{n}}_\mathrm{r}}\right) = 2 \cdot \sum_{r=1}^{I} n_r \cdot \ln\left(\frac{n_r}{N \cdot P(X=x)_r}\right) \tag{5}$$

where:

N – the sample size,

n_r – the empirical frequency in cell *r* of the contingency table, which is in the cell defined by assigned to it vector of values $x = (x_1, , x_J)$ of the $X = (X_1, X_2, \dots, X_J)$ random vector,

$\hat{n}_r$ – the theoretical frequency in cell *r* of the contingency table; $\hat{n}_r = N \cdot P(X = x)_r$, with $P(X = x)_r$ calculated according to formula (2) for table cell *r*.

The statistics is chi-square distributed with *DF* degrees of freedom:

$$DF = \prod_{j=1}^{J} R_j - 1 - \left\{C - 1 + C \cdot \sum_{j=1}^{J}(R_j - 1)\right\} \tag{6}$$

If in each cell of the contingency table the empirical and theoretical counts are equal, then $G^2 = 0$ and the model is a perfect fit. A value greater than zero measures the lack of fit, reporting the strength of the relationship (lack of independence) that remains unexplained.

The G^2 statistic is regarded as the absolute fit measure, which is used to verify the quality of fit of the model with latent *C* classes to the base model (with one latent class) using the likelihood ratio *IW* (Frątczak, 2016):

$$IW(C) = \frac{\left(G^2(1) - G^2(C)\right)}{G^2(1)} \tag{7}$$

The likelihood ratio (7) shows the quality of fitting empirical data to the model after taking into account unobserved relationships (latent classes). This proportional reduction of G^2 can be roughly regarded as R^2.

In order to compare different models of latent classes, information criteria are also used (Frątczak, 2016; Dziak et al., 2020): *AIC* (*Akaike Information Criterion*), *BIC* (*Bayesian Information Criterion*), *CAIC* (*Consistent Akaike Information Criterion*), and *ABIC* (*Adjusted Bayesian Information Criterion*).

The smaller the values of the diagnostics statistics are, the better the assessment of the model is. The most restrictive are the Bayesian criteria, which, with a small sample size, reduce the importance of extended models (with a large number of latent classes and a large contingency table).

3. Characteristics of data for research

In Poland, data on accidents at work are collected by the Statistics Poland. They are taken from statistical accident cards (Ordinance of the Minister of Economy, Labour and Social Policy of 7 January 2009) and concern accident casualties for all sections of PKD – Polish Classification of Activities (economic activities). These resources are used for the needs of official statistics; on their basis, annual collective (aggregated) reports are elaborated, made then available in paper and electronic form (Act of June 29, 1995). The collection of such actual data on accidents at work was used for the pilot studies, which are the subject of the subsequent considerations. It was provided by the administrator (GUS) under a bilateral agreement, in the form of unit (non-aggregated) registers – resources containing records from statistical accident cards from 2008-2017. The collection refers to section C according to the classification of PKD codes, which is manufacturing, and has 14,696 observations.

The statistical accident card contains 29 characteristics of the casualty. In the study, they are identified by the *pxx* symbol, where *xx* stands for the item number from the card (for example, *p01* stands for gender). Most of those features are descriptive, which implies the use of qualitative data analysis methods.

Activity branch 16 of the C section, named *Manufacture of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials* characterized by a high degree of accident risk (Pajęcki, 2020), was selected for the pilot studies. A subset of data covers records on employees injured in occupational accidents in the Wielkopolskie *voivodeship* (*p15 – Voiv. = Wielkopolskie voivodeship*) and were related to the production process (*p18 – Place of the accident = Industrial production sites*; *p19 – Work process = Production*, *processing*, *storage*).

The data were pre-explored by the cleansing process in which observations not conveying relevant information (for example, *p09 – Location of the injury = Unknown or undefined location of the injury*; *p21 – Activity performed by the casualty at the time of the accident = No information available*) and outliers were removed. Then, transformation of variables was proposed; new variables derived from the original indicators were created, by means of the aggregation of selected values of those indicators. Thus, the problem of rare categories was solved. As a result, 15 indicators were obtained – qualitative variables considered in developing the model of latent classes. Table 1 summarizes the indicators obtained after the transformation along with the modification leading to their definition. The codes for the values (consecutive positive integers) and distributions for each indicator were presented. The obtained final data set for the analysis has 2,552 records.

Table 1.
Characteristics of the data set on accidents at work in the Wielkopolskie voivodeship

Indicators and their descriptive values	Value code	[%]
wk – Enterprise size determined on the basis of the number of employees		
Up to 9 employees (without recalculation to full-time employment), including self-employed persons with no employees. *Aggregation of original values: (self-employed) + (up to 9 employees)*	1	6.11
10-49 employees (without recalculation to full-time employment)	2	19.87
50-249 employees (without recalculation to full-time employment)	3	54.27
250-499 employees (without recalculation to full-time employment)	4	12.89
more than 499 employees (without recalculation to full-time employment)	5	6.86
pkd – PKD subcategory		
Sawmilling and planing of wood	1	30.21
Manufacture of products made of wood, cork, straw and plaiting materials	2	69.79
p01 – Gender of the casualty		
Male	1	83.89
Female	2	16.11
p02 – Age of the casualty at the time of the accident		
Up to 24 years old	1	17.32
25 - 34 years old	2	30.53
35 - 44 years old	3	25.24
45 - 54 years old	4	18.65
Over 54 years *Aggregation of original values: (55 - 59 years) + (over 59 years)*	5	8.27

Cont. table 1.

p05 – Occupation of the injured person		
Non-production workers *Aggregation of original values: (representatives of public authorities) + (specialists, technicians and associate professionals) + (office workers) + (farmers, gardeners, foresters and fishermen)*	1	2.86
Industrial workers and craftsmen	2	68.85
Operators and installers of machines and devices	3	25.78
Workers doing simple jobs	4	2.51
p06 – Position job seniority in the enterprise		
Up to 5 years	1	66.03
6-10 years	2	16.54
Over 10 years *Aggregation of original values: (11-15) + (16-20) + (21-30) + (over 30 years)*	3	17.44
p07 – Hours worked between starting work and the time of the accident		
0-3	1	48.2
4-7	2	47.1
8 and more *Aggregation of original values: (8-11) + (12-15) + (16-19) + (20-23)*	3	4.7
p08 – Type of injury		
Wounds and superficial injuries	1	62.3
Bone fractures	2	13.36
Displacements, dislocations, sprains and strains	3	9.95
Traumatic amputations (loss of body parts)	4	5.49
Various other injuries *Aggregation of original values: (Internal injuries) + (Multiple injuries) + (Other injury)*	5	8.89
p09 – Location of the injury		
Head, neck *Aggregation of original values: (Head) + (Neck with cervical spine)*	1	7.56
Body *Aggregation of original values: (Thoracic and lumbar spine) + (Torso and internal organs) + (Whole body and its various parts) + (Other body part)*	2	4.39
Upper limbs	3	67.12
Lower limbs	4	20.92
p16 – Season when the accident happened		
Spring months *Aggregation of original values: (March) + (April) + (May)*	1	26.21
Summer months *Aggregation of original values: (June) + (July) + (August)*	2	23.51
Autumn months *Aggregation of original values: (September) + (October) + (November)*	3	25.82
Winter months *Aggregation of original values: (December) + (January) + (February)*	4	24.45
p17 – Time of accident; after aggregation, the time of the day on which the accident happened		
20:00-3:59 *Aggregation of original values: (20:00-23:59) + (00:00-3:59)*	1	12.03
4:00-11:59 *Aggregation of original values: (4:00-7:59) + (8:00-11:59)*	2	44.44
12:00-19:59 *Aggregation of original values: (12:00-15:59) + (16:00-19:59)*	3	43.53

Cont. table 1.

p21 – Activity performed by the casualty at the time of the accident		
Operating machines	1	46.63
Working with tools and objects *Aggregation of original values: (Working with hand tools) + (Handling objects)*	2	31.23
Workplace transporting *Aggregation of original values: (Driving / driving by means of transport / operation of moving machines and other devices) + (Manual transporting)*	3	14.34
Being at the scene of the accident *Aggregation of original values: (Motion) + (Presence)*	4	7.80
p26 – Material factor that was the source of the injury		
Buildings, surfaces *Aggregation of original values: Buildings, structures and their elements, surfaces (indoor or outdoor, permanent or movable, temporary): (at ground level) + (below ground level) + (above ground level)*	1	7.05
Another factor *Aggregation of original values: (Chemical, radioactive, explosive, biological substances) + (Safety related devices and equipment) + (Office equipment, personal or sports items, weapons) + (People and other living organisms) + (Waste)*	2	8.66
Hand or mechanized tools *Aggregation of original values: (Non-powered hand tools) + (Hand-held or hand guided mechanized tools)*	3	8.93
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	46.59
Materials, objects, products, machine parts	5	28.76
p27 – Main cause of the accident		
The defect of the material factor *Aggregation of original values: (Design defects or inappropriate technical and ergonomic solutions of the material factor) + (Improper use of the material factor) + (Material defects of the material factor)*	1	17.59
Misuse of the material factor *Aggregation of original values: (Inappropriate exploitation of the material factor) + (Non-use or inappropriate handling of the material factor by the employee)*	2	11.29
Inappropriate work organization *Aggregation of original values: (Inadequate overall organization of work) + (Inappropriate organization of the workplace)*	3	11.52
Safety neglect (heterogeneous category) *Aggregation of original values: (Failure to use protective equipment by the employee) + (Psychophysical state of the employee, not ensuring safe work performance) + (Other reason)*	4	3.92
Inappropriate arbitrary behavior of the employee	5	8.70
Inappropriate behavior of the employee (heterogeneous category) *The behavior resulting from the aggregation of the original reasons: (Ignorance or disregarding of the threat) + (Ignorance of the occupational health and safety rules) + (Disregarding the orders of superiors) + (Insufficient concentration) + (Surprise by an unexpected event) + (Incorrect pace of work) + (Inexperience) + (Other reasons)*	6	46.98
p289 - Accident severity (new variable), defined on the basis of variables: *p28* (Effect of the accident) and p29 (Time of incapacity for work)		
Minor accident resulting in inability to work for 0-13 days	1	23.71
Minor accident resulting in inability to work for 14-29 days	2	28.72
Minor accident resulting in inability to work for 30-89 days	3	35.93
Serious accident *Value aggregation: (Severe accident - regardless of the number of days of inability to work) + (Fatal accident) + (Minor accident causing inability to work for more than 90 days)*	4	11.64

Source: authors' own elaboration.

4. Latent class model for occupational accidents

For the selected subset of data (accidents at work in branch 16 of the C section related to the production process, which occurred in enterprises in the Wielkopolskie voivodeship), a series of experiments with a different number of latent classes was performed in order to determine a stable model. The tested number of latent classes varied from 1 to 6. For each variant with a given number of classes, 20 models were estimated (with a different initial value for the iterative process calculating the model). The models estimated for each variant of the number of latent classes were then assessed using diagnostic statistics. The results for the best model within each variant are presented in Table 2. The last column shows the repeatability of the best model in the remaining nineteen models.

The worst repeatability of the best model was obtained for 4-calss and 6-class models – the repeatability rate is equal to 20%, which proves the instability of the results. The most often used when comparing models, *Bayesian Information Criterion* has the lowest value (41148.98) for the model with five latent classes. Similarly, a good assessment of this model can be derived from the values of the other diagnostic statistics. Therefore, as the best one, the 5-class model was selected for the discussion of accident patterns hereafter. It is marked in bold in the table.

Table 2.
Values of statistics for diagnosing latent class models by number of classes

Number of latent classes*	G^2	AIC	BIC	CAIC	ABIC	Entropy statistic	Repeatability rate of the best model
1	42608.06	42696.06	42953.23	42997.23	42813.43	1.00	100%
2	41049.48	41227.48	41747.65	41836.65	41464.87	0.69	95%
3	40276.48	40544.48	41327.66	41461.66	40901.90	0.67	100%
4	39818.87	40176.87	41223.06	41402.06	40654.32	0.74	20%
5	**39391.78**	**39839.78**	**41148.98**	**41372.98**	**40437.27**	**0.76**	**85%**
6	39103.08	39641.08	41213.28	41482.28	40358.60	0.76	20%

* Degrees of freedom (*DF*) in the scientific notation with the order of magnitude (exponent) value of 8 is the same for all models, regardless of the number of latent classes, and is equal to 4.15E + 08.

Source: authors' own elaboration.

The values of the probability estimators of the latent classes of the selected 5-class model are illustrated in Figure 2. Table 3 shows the values of conditional probabilities for the analyzed indicators, provided that they belong to a given latent class – the modal probabilities for rows are marked in italics and bold type and the modal probabilities for the columns have grey background. Cells with a value that is modal both by columns and by rows for a given indicator are marked with bold borders.

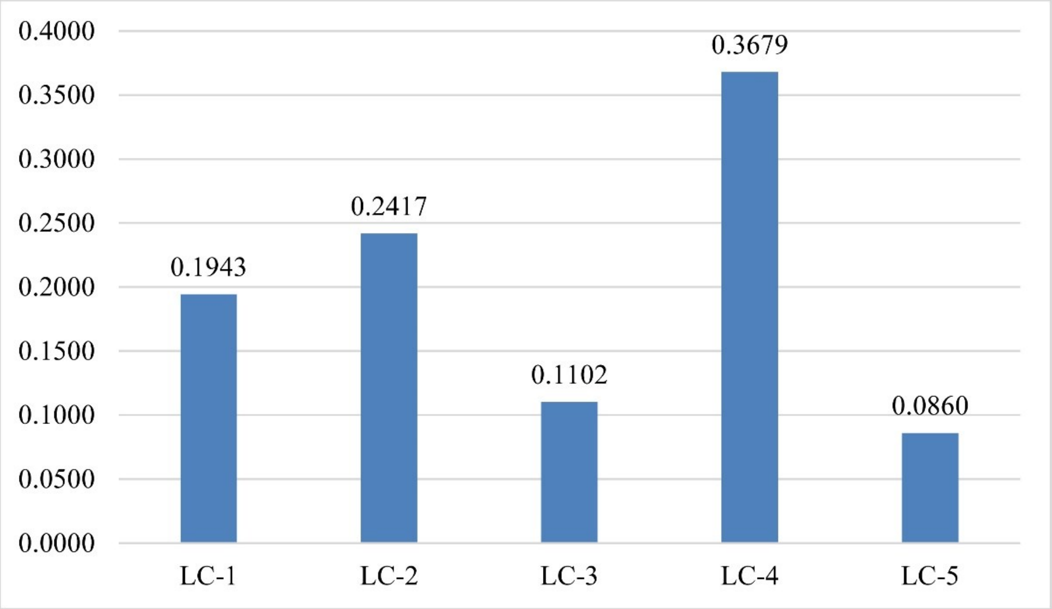

Figure 2. Probabilities of latent classes for the 5-class model. Source: authors' own elaboration.

Table 3.
Conditional probabilities for indicators of the 5-class model

Indicator	Indicator code	LC-1	LC-2	LC-3	LC-4	LC-5
wk	1	0.0720	0.0003	0.0540	***0.1041***	0.0326
	2	***0.2779***	0.1147	0.2120	0.2215	0.1411
	3	0.5364	***0.5864***	0.5665	0.5444	0.3965
	4	0.0836	0.1726	0.0791	0.0994	***0.2986***
	5	0.0301	0.1260	0.0883	0.0306	***0.1312***
pkd	1	***0.4816***	0.3246	0.1893	0.2579	0.1674
	2	0.5184	0.6754	0.8107	0.7421	***0.8326***
p01	1	***0.9017***	0.7928	0.8400	0.8602	0.7345
	2	0.0983	0.2072	0.1600	0.1398	***0.2655***
p02	1	0.1160	0.1631	***0.2208***	0.2145	0.0935
	2	0.2576	0.3563	***0.4071***	0.2761	0.2636
	3	0.2682	0.2593	0.2030	0.2400	***0.3130***
	4	***0.2468***	0.1731	0.1200	0.1694	0.2463
	5	***0.1114***	0.0482	0.0491	0.1000	0.0836
p05	1	0.0149	0.0457	0.0235	0.0070	***0.1105***
	2	0.7126	0.5582	0.7363	***0.7834***	0.5333
	3	0.2384	***0.3583***	0.2048	0.1983	0.3418
	4	0.0341	***0.0378***	0.0354	0.0113	0.0145
p06	1	0.6411	0.6485	0.6778	***0.7122***	0.4923
	2	0.1188	0.2105	***0.2180***	0.1462	0.1581
	3	0.2401	0.1410	0.1042	0.1416	***0.3496***
p07	1	0.4288	0.5049	0.4588	***0.5120***	0.4391
	2	***0.5220***	0.4423	0.4935	0.4466	0.5119
	3	0.0492	***0.0528***	0.0476	0.0414	0.0490
p08	1	0.2588	0.7639	***0.8583***	0.7509	0.2021
	2	***0.3744***	0.0192	0.0821	0.0657	0.2678
	3	0.1523	0.1104	0.0002	0.0095	***0.4619***
	4	0.0331	0.0000	0.0066	***0.1281***	0.0068
	5	***0.1815***	0.1065	0.0528	0.0458	0.0614

Cont. table 3.

p09	1	0.0273	***0.1860***	0.1462	0.0118	0.0574
	2	***0.0994***	0.0738	0.0045	0.0036	0.0570
	3	0.5576	0.4792	0.7827	***0.9456***	0.1518
	4	0.3157	0.2610	0.0665	0.0390	***0.7339***
p16	1	0.2752	0.2169	0.2582	0.2759	***0.3060***
	2	0.2400	0.2315	***0.2761***	0.2280	0.2124
	3	0.2398	***0.3009***	0.2218	0.2508	0.2586
	4	0.2450	***0.2507***	0.2439	0.2454	0.2230
p17	1	0.1088	0.1359	0.1165	0.0957	***0.2123***
	2	0.4525	0.4050	0.3873	***0.5049***	0.3507
	3	0.4386	0.4590	***0.4962***	0.3994	0.4370
p21	1	0.3475	0.3578	0.0080	***0.8172***	0.1267
	2	0.3573	0.2819	***0.9866***	0.1732	0.0273
	3	0.2652	***0.2928***	0.0004	0.0047	0.2250
	4	0.0300	0.0675	0.0050	0.0050	***0.6210***
p26	1	0.0557	0.0462	0.0354	0.0001	***0.5185***
	2	0.1112	0.1060	0.0820	0.0370	***0.1941***
	3	0.0168	0.0081	***0.6908***	0.0217	0.0001
	4	0.2158	0.3437	0.0074	***0.8782***	0.1989
	5	***0.6004***	0.4960	0.1845	0.0630	0.0884
p27	1	0.1940	0.1709	0.0789	***0.2290***	0.0470
	2	0.1227	0.0971	***0.1773***	0.1202	0.0208
	3	0.1374	0.1683	0.0688	0.0514	***0.2483***
	4	0.0359	0.0307	0.0427	0.0293	***0.1079***
	5	0.0746	0.0554	0.0868	***0.1170***	0.0757
	6	0.4354	0.4776	***0.5456***	0.4531	0.5003
p289	1	0.0004	**0.5590**	0.4599	0.1129	0.1123
	2	0.1878	**0.3950**	0.3327	0.2665	0.2394
	3	***0.6225***	0.0459	0.1712	0.4668	0.4271
	4	0.1892	0.0002	0.0362	0.1538	***0.2212***

Source: authors' own elaboration.

5. Occupational accident patterns

The obtained final model consists of latent classes, each of which stands for a certain pattern of occupational accidents. The patterns are discussed in the following subchapters on the basis of the values of selected indicators investigated in the analysis. Each class has been given a label. The labels are short names providing a brief expression of the character of the patterns.

5.1. Latent class model description Latent Class 1 – *Severe accidents, especially limb fractures*

Out of all the observations, 19.43% belong to the first latent class. The class is characterized by accident events resulting in a long or a very long absence from work, including those causing the death of the injured individuals, *P(p289=3|LC-1)* = 0.62*; P(p289=4|LC-1*) = 0.19. The material factors causing the injuries are especially: *Materials, objects, products, machine* *parts*, *P(p26=5|LC-1)* = 0.60. *Inappropriate behavior of the employee* is primarily the cause of the accident, *P(p27=6|LC-1)* = 0.44. Actions performed by the injured person at the time of the accident are: *Operating machines*, *P (p21 =1|LC-1)* = 0.35 or *Working with tools and objects*, *P (p21=2|LC-1)* = 0.36. The injuries are most often located in the *limbs*, mainly *Upper* ones*,* *P(p09=3|LC-1)* = 0.56, slightly less often *Lower* ones, *P(p09=4|LC-1)* = 0.32. The type of injury that distinguishes the class from the entire data set are *Bone fractures*, *P(p08=2|LC-1)* = 0.37, but also *Various other injuries (*covering *Internal* and *Multiple injuries*), *P(p08=5|LC-1)* = 0.19. The accidents are connected with the *Production of sawmill products* PKD subcategory, *P(pkd=1|LC-1)* = 0.48. Men are the casualties, *P(p01=1|LC-1)* = 0.90. Persons aged 55 and over are more likely to be affected in this class than in other classes, *P(p02=5|LC-1)* = 0.11.

5.2. Latent class model description Latent Class 2 – *Minor accidents causing superficial wounds and injuries to the upper body part*

The second latent class includes 24.17% of the observations of the analyzed data set. Almost the entire class relates to minor accidents, with up to 29 days of inability to work, *P((p289 = 1 or p289 = 2)|LC-2)* = 0.95. *Inappropriate behavior of the employee* is the main cause of the accident, *P(p27=6|LC-2)* = 0.48. The most common material factors causing injury are: *Materials, objects, products, machine parts*, *P(p26=5|LC-2)* = 0.50, and *Machines and devices*, *P(p26=4|LC-2)* = 0.34. *Operating machines* and *Workplace transporting* prevail as the activity of the employee at the time of the accident, *P(p21=1|LC-2)* = 0.36 and *P(p21=3|LC-2)* = 0.29 respectively. The injuries mainly affect the upper body of the person – the *Upper limbs* or *Head and neck*, *P((p09=3 or p09=1)|LC-2)* = 0.48 + 0.19 = 0.67. *Wounds and superficial injuries* are the result of the accident, *P(p08=1|LC-2)* = 0.76. The second latent class represents *Industrial workers and craftsmen* as well as *Operators and engineers of machines and devices* as the casualties of occupational accidents, *P(p05=2 or p05=3|LC-2)* = 0.56 + 0.36 = 0.92.

5.3. Latent class model description Latent Class 3 – *Minor accidents caused by the careless use of tools*

The third latent class includes 11.02% of all the observations of the analyzed data set. As in the case of the second latent class, the discussed class is dominated by minor accidents, resulting in inability to work not more than 29 days, *P((p289 = 1 or p289 = 2)|LC-3)* = 0.79. *Inappropriate behavior of the employee* is the main cause of the accident, *P(p27=6|LC-3)* = 0.55, but also *Misuse of the material factor* happens comparatively frequently, *P(p27=2|LC-3)* = 0.18. *Hand or mechanized tools* belong mainly to material factors causing the injury, *P(p26=3|LC-2)* = 0.69, while almost always *Working with tools and objects* is an activity performed by the injured person at the time of the accident occurrence, *P(p21=2|LC-3)* = 0.99. The injuries affect *Lower limbs*, *P(p09=4|LC-3)* = 0.78, and these are mainly wounds and superficial injuries, *P(p08=1|KU-3)* = 0.86. Young people under the age of 35 are involved in accidents that are characteristic for the third latent class, *P((p02=1 or p02=2)|LC-3)* = 0.22 + 0.41 = 0.63.

5.4. Latent class model description Latent Class 4 – *Severe accidents related to careless handling of machinery by short experience employees*

The fourth latent class is the most numerous – it comprises 36.79% of all the observations. The class covers accidents causing a long or a very long absence from work, including those leading to the death of the injured person, *P(p289=3|LC-4)* = 0.47, *P(p289=4|LC-4)* = 0.15. *Inappropriate behavior of the employee* or *Inappropriate arbitrary behavior of the employee* are the main causes of the accident, *P((p27=6 or p27=5)|LC-4)* = 0.45 + 0.12 = 0.57. *Machines and devices* are primary material factors causing the injury, *P(p26=4|LC-4)* = 0.88, while *Operating machines* is an activity performed by the injured person at the time of the accident occurrence, *P(p21=1|LC-4) = 0.82*. The injuries affect *Upper limbs*, *P(p09=3|LC-4)* = 0.95. The predominant type of injury suffered by the worker are *Wounds and superficial injuries*, *P(p08=1|LC-4)* = 0.75, but there are also *Traumatic amputations (loss of body parts)*, *P(p08=4|KU-4)* = 0.13. The accidents concern employees with *Up to 5 years* of experience in the position held in the production plant, *P(p06=1|LC-4)* = 0.71.

5.5. Latent class model description Latent Class 5 – *Severe accidents, caused by inexperience or by routine, not related to employing the device*

The fifth latent class is the least numerous – it comprises 8.60% of all the observations. In the context of incapacity for work, it mostly covers accidents of a significant severity, *P(p289=3|LC-5)* = 0.43 and *P(p289=4|LC-5) = 0.22* (the highest probability compared to other latent classes). The main cause of accidents varies and it may relate to the following behaviors: *Inappropriate behavior of the employee*, *P(p27=6|LC-5)* = 0.50, *Inappropriate work organization*, *P(p27=3| LC-5)* = 0.25, and *Safety neglect P(p27=4| LC -5)* = 0.11. *Buildings, surfaces*, *P(p26=1| LC-5)* = 0.52, and, to a lesser extent, *Another factor*, *P(p26=2| LC-5)* = 0.20, are material factors characteristic of the class as regards the source of the injury. *Being at the scene of the accident* is the most common casualty activity at the time of the accident occurrence, *P(p21=4|LC-5)* = 0.62. *Lower limbs* is the injury location, *P(p09=4|LC-5)* = 0.73, with the basic type of injury being *Displacements, dislocations, sprains and strains* or *Bone fractures*, *P((p08=3 or p08=2)|LC-5)* = 0.46 + 0.27 = 0.73. The most probable value of the job seniority in the position held in the enterprise is the period *Up to 5 years*, *P(p06=1|LC-5)* = 0.49, secondly, *Over 10 years* (experienced workers), *P(p06=3| LC-5)* = 0.35 (the highest percentage compared to other classes). It is worth noting that the class is characterized by a high probability of injured women, *P(p01=2|LC-5)* = 0.27 (a value greater than the frequency observed in the whole population equal to 0.16). This feature of the fifth latent class may be connected with the fact that the occupation performed by the injured person is characterized by a relatively high probability (compared to other classes) of *Non-production workers P(p05=1|LC-5)* = 0.11 (also a value greater than the frequency observed in the entire data set, equal to 0.03).

6. Summary and conclusions

Accidents at work often cause severe or even fatal injuries to the injured person, and also contribute to considerable societal burdens and economic losses. The identification of certain recurring mechanisms related to the occurrence of accidents may help in the development of effective tools leading to the improvement of work safety. Pilot studies were undertaken in the work, which showed that latent class analysis is a promising tool supporting the scientific investigation in that area.

The analysis covers occupational accidents in the economic activity of branch 16 of section C called *manufacturing* (according to Polish Classification of Activities), which occurred in connection with production processes in enterprises in the Wielkopolskie voivodeship in 2008-2017. Three severe accident patterns and two light accident patterns were obtained, which were subjected to descriptive characterizations, using interpretable results presented in the form of classification probabilities of the category of observable variables symptomatic for a particular latent class.

The task, undertaken in the work, of using selected data mining techniques to study the occupational accident phenomenon on the basis of non-aggregated individual data about accident casualties, seems to be justified and promising. Subsequently, in perspective, the research based on a larger data set will be continued along with the analysis extension. There can also be other issues related to the problem which are worth considering in the future, such as improving the process of assessing and selecting a final latent class model out of many possibilities or developing systematic selection of indicators for the characterization of the latent class. However, the presented examples of directions for further research do not exhaust the discussed issue.

References

1. Act of 29 June 1995, on Public Statistics (Journal of Laws of 1995, No. 88, item 439, as amended).
2. Bogdan, M., and Boczkowska, K. (2009). Modelowanie wypadku przy pracy na stanowisku bobiniarki w przedsiębiorstwie produkcyjnym. *Zeszyty Naukowe. Organizacja i Zarządzanie. Politechnika Łódzka, Z. 45, Nr 1064*, pp. 123-140.
3. Cheng, C.-W., Leu, S.-S., Lin, C.-C., and Fan, C. (2010). Characteristic analysis of occupational accidents at small construction enterprises. *Safety Science*, *Vol. 48, Iss. 6,* pp. 698-707, doi: 10.1016/j.ssci.2010.02.001.

4. Dziak, J.J., Coffman, D.L., Lanza, S.T., Li, R., and Jermiin, L.S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics, Vol. 21, Iss. 2,* pp. 553-565, doi: 10.1093/bib/bbz016.
5. Ejdys, J. (Ed.) (2010). *Kształtowanie kultury bezpieczeństwa i higieny pracy w organizacji*. Białystok: Oficyna Wydawnicza Politechniki Białostockiej.
6. Frątczak, E. (Ed.) (2013). *Zaawansowane metody analiz statystycznych*. Warszawa: Oficyna Wydawnicza Szkoła Głowna Handlowa.
7. Gajdzik, B. (2013). Błędy prowokujące wypadki w pracy w przedsiębiorstwie produkcyjnym – analiza case study. *Journal of Ecology and Health, Vol. 17, Nr 2,* pp. 87-90.
8. Huang, X., and Hinze, J. (2003). Analysis of Construction Worker Fall Accidents. *Journal of Construction Engineering and Management*, *Vol. 129, Iss. 3*, pp. 262-271, doi: 10.1061/(ASCE)0733-9364(2003)129:3(262).
9. Kakhki, F.D., Freeman, S.A., Freeman, M., and Mosher, G.A. (2019). Segmentation of Severe Occupational Incidents in Agribusiness Industries Using Latent Class Clustering. *Applied Sciences, Vol. 9, Iss. 18,* 3641, doi: 10.3390/app9183641.
10. Lavery, R. (2011). *An Animated Guide: An Introduction to Latent Class Clustering in SAS®*. Retrieved from https://www.lexjansen.com/phuse/2011/sp/SP07.pdf, 05.02.2020.
11. Macedo, A.C., and Silva, I.L. (2005). Analysis of occupational accidents in Portugal between 1992 and 2001. *Safety Science*, *Vol. 43, Iss. 5-6*, pp. 269-286, doi: 10.1016/j.ssci.2005.06.004.
12. *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).
13. Ostasiewicz, W. (2012). *Myślenie statystyczne* [*Statistical thinking*]. Warszawa: Wolters Kulwer.
14. Pajęcki, M. (2020). Bezpieczeństwo pracy w sekcji przetwórstwa przemysłowego w Polsce – stan wypadkowości. In: R. Knosala (Ed.), *Inżynieria zarządzania. Cyfryzacja produkcji. Aktualności badawcze 2* (pp. 1223-1232). Warszawa: PWE.
15. Roszko-Wójtowicz, E. (2016). The Analysis of the Current State of Accidents at Work in Poland in the Years 2002-2014. *Myśl Ekonomiczna i Polityczna, Nr 3(54),* pp. 32-58.
16. Szóstak, M. (2018). The application of cluster analysis to identify the occupational profile of people injured in accidents in the Polish construction industry. *IOP Conference Series: Materials Science and Engineering*, *Vol. 456, Conference 1*, doi: 10.1088/1757-899X/456/1/012027.
17. Węgrzyn, M. (2017). Analiza danych dotyczących przyczyn wypadków przy pracy oraz liczby osób poszkodowanych. *Zeszyty Naukowe SGSP, Nr 62, Tom 1,* pp. 185-201.
18. Wirkus, M., and Bajorski, J. (2017). Wypadki i sytuacje niebezpieczne w systemie zarządzania bezpieczeństwem pracy. *Zarządzanie przedsiębiorstwem, Nr 1,* pp. 36-42.

19. Zjawin, A., and Kołodziej, S. (2018). Analiza wypadkowości na przykładzie wybranego zakładu produkcyjnego. *Autobusy: technika, eksploatacja, systemy transportowe, Nr 6,* pp. 327-33.
20. Zwetsloot, G., Leka, S., and Kines, P. (2017). Vision zero: from accident prevention to the promotion of health, safety and well-being at work. *Policy and Practice in Health and Safety*, *Vol. 15, Iss. 2,* pp. 88-100, doi: 10.1080/14773996.2017.1308701.