Large Occupational Accidents Data Analysis with a Coupled Unsupervised Algorithm: The S.O.M. K-Means Method. An Application to the Wood Industry

safety

MDPI

Article

Large Occupational Accidents Data Analysis with a Coupled Unsupervised Algorithm: The S.O.M. K-Means Method. An Application to the Wood Industry

Lorenzo Comberti [1], Micaela Demichela [1], Gabriele Baldissone [1,*], Gianmario Fois [2] and Roberto Luzzi [2]

[1] SAfeR-Centro Studi su Sicurezza, Affidabilità e Rischi, Dipartimento Scienza Applicata e Tecnologia, Politecnico di Torino, 10129 Torino, Italy; lorenzo.comberti@polito.it (L.C.); micaela.demichela@polito.it (M.D.)

[2] Contarp, INAIL Direzione Regionale del Piemonte, 10100 Torino, Italy; g.fois@inail.it (G.F.); r.luzzi@inail.it (R.L.)

* Correspondence: gabriele.baldissone@polito.it; Tel.: + 39-011-090-4629

Received: 23 July 2018; Accepted: 25 October 2018; Published: 1 November 2018

Abstract: Data on occupational accidents are usually stored in large databases by worker compensation authorities, and by the safety and prevention teams of companies. An analysis of these databases can play an important role in the prevention of accidents and the reduction of risks, but it can be a complex procedure because of the dimensions and complexity of such databases. The SKM (SOM K-Means) method, a two-level clustering system, made up of SOM (Self Organizing Map) and K-Means clustering, has obtained positive results in identifying the dynamics of critical accidents by referring to a database of 1200 occupational accidents that had occurred in the wood industry. The present research has been conducted to validate the recently presented SKM methodology through the analysis of a larger data set of more than 4000 occupational accidents that occurred in Piedmont (Italy), between 2006 and 2013. This work has partitioned the accidents into groups of different accident dynamics families and has quantified the severity and frequency of occurrence of these accidents. The obtained information may be of help to Company Managers and National Authorities to better address preventive measures and policies concerning the clusters that have been identified as being the most critical within a risk-based decision-making framework.

Keywords: clustering; SOM; accident database analysis; accident prevention; safety; risk-based decision making

1. Introduction

Occupational accidents have an important effect on the economies of the whole world, as pointed out by Hamalainen et al. [1].

Reporting and analyzing occupational accidents in order to improve the data available for prevention purposes have been safety management requirements since 1923, when the First International Conference of Labor Statisticians firs defined standards for accident classification. Since 1989, the EU has promoted various policies to reduce the frequency of occupational accidents. The Treaty on the Functioning of the European Union (article 153) in fact states: *'[. . .] the Union shall support and complement the activities of the Member States in the following fields: (a) improvement in particular of the working environment to protect workers' health and safety; [. . .]'*. In January 1990, the European Union launched a European Statistics study on Accidents at Work (ESAW), based on the International

Labor Organization (ILO) standards. As a result of this project, the 'European Statistics on Accidents at Work–Methodology, was published by Eurostat in 2001 and a revised edition was released in 2013 [2].

ESAW describes each occupational accident by means of several parameters, and provides information about the dynamics, time, place, working situation and workers involved.

This large amount of information is analyzed by the EU National Health and Safety Authorities by means of traditional statistical methods, according to Regulation 1338/2008 and Regulation 349/2001 on Community statistics pertaining to public health and health and safety at work.

The results of this approach are published regularly in official reports, by National Health and Safety Authorities, and they highlight such useful and general information on the trends of occupational accidents as: The classes of workers most exposed to accidents, gender effects, the role of the educational level, the age of the injured and various other parameters. In addition, ESAW data have also been analyzed, with reference to a specific field of activity, through a statistics approach to analyze the cause-effect mechanism [3], and information about the trend of accidents and "typical" accidents have been reported in the recent work by Dzwiarek et al. [4] and Kogler et al. [5]. However, these kinds of analyses are only useful to a certain extent to enhance the prevention of accidents in the work environment, as observed by Palamara et al. [6] and Comberti et al. [7], because they do not produce a risk assessment outcome [8].

In addition, the statistical analysis of data characterized by non-numerical variables, such as ESAW data, makes the analysis very difficult, and it requires many a-priori assumptions and tests on the nature of the data distribution (e.g., a CHI-coefficient test). An alternative approach, to overcome the use of statistics, is that of resorting to data mining methods [9,10], which include several different data analysis techniques. Some interesting results, related to ESAW data, have in fact been obtained with Multi Correspondence Analysis (MCA) [8] and Pattern Identification [11], which have allowed the most important accident scenarios to be identified, together with a quantification of the frequency of accidents, but they have not produced a quantification of the associated risk.

A powerful method that has been used in different analysis fields to support risk assessments is the SOM (Self Organizing Map): An unsupervised learning algorithm that is used to generate topologies, while preserving transformations from a high-dimensional data vector space to a low-dimensional map space. In other words, with SOM, it is possible to view a set of multiple-dimension data in a 2-dimensional space. This possibility facilitates data analysis.

SOM algorithms have been applied to different risk-classification problems. Gevrey used SOM to estimate the risk of the establishment of invasive species [12], Liang [13] proposed SOM to classify pipeline sections with the same risk level into different risk patterns, and Asgary [14] used SOM to classify and assess the risk levels of structural fire accidents.

Palamara et al. [6] proposed combining SOM with a clustering algorithm, as previously proposed by Vesanto and Alhoniemi [15], to identify the most critical groups of occupational accidents from ESAW data. This work produced promising results, but suffered from several numerical stability problems—the results were strongly fluctuant when the analysis was repeated.

In 2015, these limits were solved by Comberti et al. [7], who published a sensitivity analysis and set up a revised method named "SKM" (SOM K-Means method). SKM also allows a quantification of the risk, made on the basis of clustering partition, to be associated to the qualitative figures that are represented by SOM maps, and allows the results to be used as a decision making support for prevention purposes, as suggested by Demichela et al. [16], and adopted by Murè et al. [17] and Comberti et al. [18].

This paper describes a research project that has focused on the application of SKM to a large database of occupational accidents that have occurred in the wood industry. The aims of the work have been to test the effectiveness of SKM with a larger data set than in the previous works and to identify occupational accident families, together with a quantification of an awareness of their occurrence and frequency. As discussed in Top et al. [19], the wood industry is mainly characterized by small and medium-sized enterprises—SMEs—whose operators are exposed to multiple hazard factors.

The analysis of the dynamics of accidents that have occurred can help support occupational risk managers identify which hazard have led to the most occupational accidents, and which factors have contributed to the different dynamics—thus guiding prevention actions. Accident-dynamics data are in fact crucial for risk assessments and risk-based decision making, as discussed in Leva et al. [20] and Demichela et al. [16] for high voltage equipment; in Darabnia and Demichela [21,22] for the analysis of human and organizational factors pertaining to maintenance optimization; in Gerbec et al. [23,24] for the design of critical operations, or more in general, for a total safety management, as dealt with in Leva et al. [25,26].

A description of the methodology is given in Section 2. Its application to the wood industry data and the relevant results are shown in Section 3. A discussion and conclusions complete the paper.

2. Materials and Methods

2.1. The SKM Method

SOM is applied in SKM to coded data obtained from an occupational accident database. SOM can represent the occupational data set in a two-dimension map. This process reflects the data similarity within occupational databases: Accidents with similar descriptive parameters are projected into the next units and very different accidents are projected into distant units.

SKM has here been implemented in Matlab® 7.0 (7.0, MathWorks, Natick, MA, USA) coding with an interface designed in Excel® (Excel 2013, Microsoft, Redmond, WA, USA). SKM has been structured in three phases:

1. A pre-processing procedure that pre-treats available data for the subsequent numerical processing;
2. SOM elaboration, which returns a visual map of the occupational accident domain;
3. K-Means calculation, which leads to the final clusters and accident partition.

The SKM structure is shown in Figure 1.

2.1.1. Pre-Processing Phase

The data set used in this study was taken from the INAIL (Italian institution for insurance against accidents at work) database, where accidents are reported according to the ESAW taxonomy.

Each accident is described by more than 20 variables, that is: Geographical location of the accident, time of occurrence, details about the injured party (activity, age ...), dynamics of the accident (deviation from normal procedures, contact and mode of injury) and circumstances of the accident (workstation, working environment).

The combination of the number of elements and the huge number of descriptive variables requires a great calculation effort. Furthermore, most of the variables are categorical elements, whereas the algorithms for SOM and K-Means calculation require numerical ones.

The method requires a pre-processing phase to adapt the data from the occupational accident database to the algorithm characteristic. The pre-processing phase overcomes these two drawbacks by means of a two-step coding procedure.

The first step is focused on the construction of an Accident Matrix (AM). The AM contains the occupational accidents that have to be processed; this matrix has a dimension D, which is obtained from:

$$D = n \times m, \quad (1)$$

where n is the accident number, and m is the number of variables selected from among those available in the ESAW classification to describe each accident.

Each variable can assume different values but, to limit the computational efforts, these values are limited with respect to the hierarchical structure of the ESAW classification. Table 1 shows part of the ESAW taxonomy for the "Activity" variable: According to the coding procedure, the labels from 41 to

49, pertaining to "handling of objects", will be replaced by the upper level label 40, while the labels from 61 to 69, pertaining to "movement", will be replaced by label 60.

The second step involves numerical coding; each accident is coded from a sequence of categorical information to a sequence of numbers.

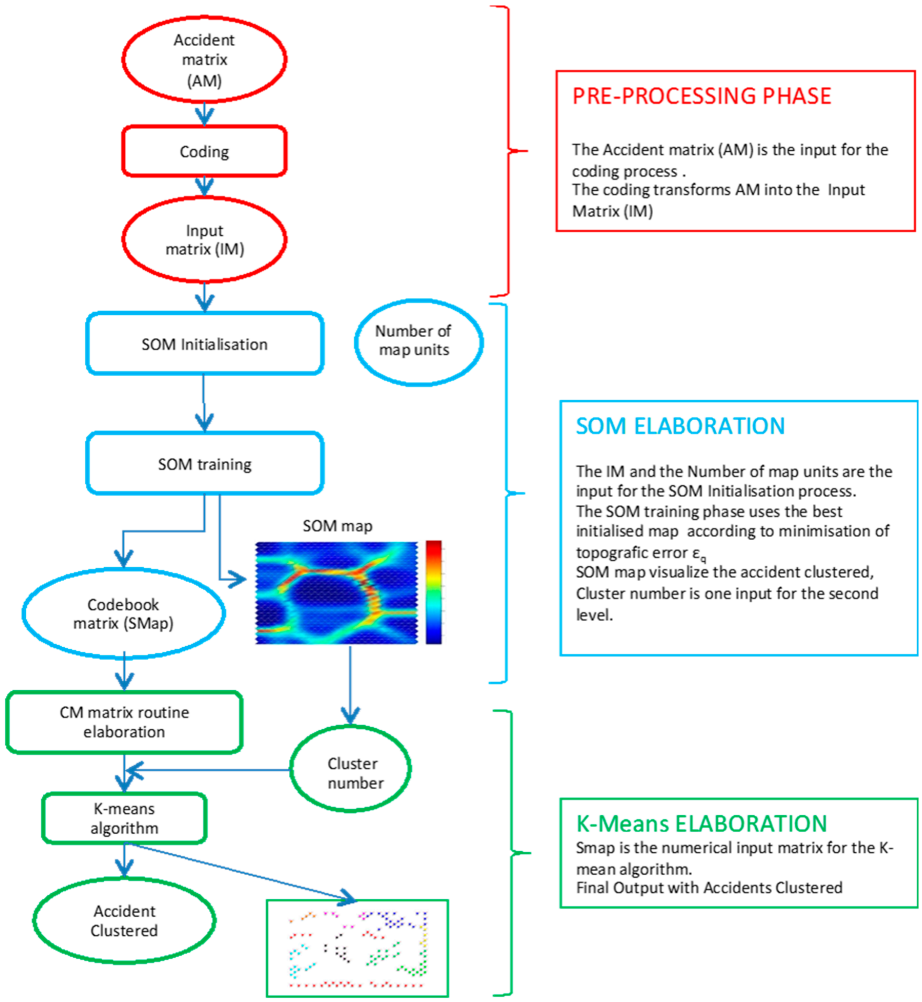

Figure 1. Self Organizing Map K-Means (SKM) scheme.

Table 1. European Statistics study on Accidents at Work (ESAW) hierarchical classification, the upper and lower levels.

40 Handling of Objects	60 Movement
41 Manually taking hold of, grasping, seizing, holding, placing—on a horizontal level	61 Walking, running, going up, going down, etc.
42 Tying, binding, tearing off, undoing, squeezing, unscrewing, screwing, turning	62 Getting in or out
43 Fastening, hanging up, raising, putting up—on a vertical level	63 Jumping, hopping, etc.
44 Throwing, flinging away	64 Crawling, climbing, etc.
45 Opening, closing (box, package, parcel)	65 Getting up, sitting down
46 Pouring, pouring into, filling up, watering, spraying, emptying, baling out	66 Swimming, diving
47 Opening (a drawer), pushing (a warehouse/office/cupboard door)	67 Movements on the spot
49 Other group 40 type Specific Physical Activities not listed above	69 Other group 60 type Specific Physical Activities not listed above

As reported in Palamara et al. [6], each parameter is coded in a numerical vector that contains a sequence of zeros and a single 1. The union of the vectors that describe the variables used for the analysis leads to the complete coding of each accident.

The resulting vector will have as many *1s* as the variables and as many *0s* as the total number of categories for all the variables, less the number of variables.

The "Input matrix" (IM) contains all the accidents coded into numerical vectors; its dimension (D_{input}) is obtained from:

$$D_{input} = n \times p, \quad (2)$$

where n is the number of accidents and p is obtained from the number of variables multiplied by the number of categories used to describe them.

Let us assume that an accident is described by 4 variables and each variable can have 5 possible different categories. The parameter p will thus have a value of 20.

This coding procedure is run automatically through the use of conversion tables that allow an univocal correspondence between categorical values and numerical vectors to be achieved, as shown in Table 2.

Table 2. Coding table for the ESAW "contact" variable.

Contact	Categories	Numeral Coding								
1	Contact with energy	1	0	0	0	0	0	0	0	0
2	Crushing	0	1	0	0	0	0	0	0	0
3	Impact with pitched material	0	0	1	0	0	0	0	0	0
4	Collision with transport system	0	0	0	1	0	0	0	0	0
5	Contact with cutting tool	0	0	0	0	1	0	0	0	0
6	Snugged/sprained	0	0	0	0	0	1	0	0	0
7	Physical effort	0	0	0	0	0	0	1	0	0
8	Violent bump	0	0	0	0	0	0	0	1	0
9	No information	0	0	0	0	0	0	0	0	1

At the end of the pre-processing phase, the AM that originally contained a group of selected occupational accidents is coded into the IM that contains an equivalent number of numerical vectors.

2.1.2. SOM Elaboration

With reference to Figure 1, the first level of SKM contains the Self Organizing Map (SOM) algorithm, which allows multidimensional vectors to be represented in a two-dimensional space, while preserving the topology of the multidimensional space.

SOM is based on a neural network scheme that is formed by two layers: The first layer is made up of the input vectors; the second layer is a map that is characterized by several units that are set by the user.

There are several ways of calculating SOM; SKM is configured with the "batch SOM" approach [27], which guarantees faster and more efficient performances for complex data sets than the traditional approach.

This approach uses an iterative calculation of matrices and it depends on the initial condition, as will be discussed later on.

The input data are fed as a single block, that is, "batch" [27], and the algorithm assigns a random vector of equal size as the input data, called "weight", to each unit during the initialization phase.

In the training phase, the algorithm calculates the Hamming distance [28] between IM elements and all the unit weights.

This is an iterative process in which, at each iteration, the input data set is presented as a batch to the SOM, and the algorithm calculates the distance between each input vector and each unit weight vector. As in a competitive learning algorithm, the units in the map layer compete to represent the input data and, for each input data, the unit whose weight vector is closest to it wins the competition. This unit is called the 'Best Matching Unit' (BMU).

The weight vector values of the winning units are updated, at each iteration, in order to make each output unit representative of a particular kind of input [29], together with those of the surrounding

units. The magnitude of this update depends on the distance between the winning unit in the network and the other units, according to the Gaussian neighborhood function.

The value of the neighborhood function decrees with the distance from the winning unit. In this way, the weight of the units around the winner is modified, while it remains almost unaltered for distant units.

This ensures that the data projected into the next units are similar.

The process ends when each input data is coupled with a BMU.

As mentioned above, this iterative process depends on the initial condition; in order to deal with this dependency, the SKM allows several independent initializations, named seeds, to be made, and these produce several different rough maps.

SKM evaluates, for each map, the topology preservation accuracy that describes how well the data, which are close in the input space, are projected to close units in the SOM.

The topology preservation accuracy is pointed out by the topographic error, which is given by the following equation:

$$\varepsilon_q = \frac{1}{N}\sum_{1}^{N} u(x_i), \tag{3}$$

where N is the data number, x_i is the ith input data and $u(x_i)$ is equal to 1, if the first and the second best matching units are not adjacent units, otherwise it is zero.

The topographic error minimization leads to the identification of the best map among all those generated.

At the end of the training process, the map has organized itself by mapping input data into SOM units and, in particular, by connecting similar input data to neighboring units.

The number of units has to be chosen by the user. There is not an objective criterion to set it up and, as discussed in Comberti et al. [7], a rule of thumb is to set it with a lower value than the number of analysed occupational accidents.

The output of the training process is a bi-dimensional map and a numerical output that is represented by a matrix called SMap.

SMap contains the numerical code of the map and the dimension of this matrix, which is obtained from:

$$\mathrm{D_{SMap}} = \mathrm{U} \times \mathrm{p}, \tag{4}$$

where U is the number of the unit of the map and p is the same as for Equation (2).

Each element is characterized by a sequence of real numbers that represent the weights of each unit, which is also called prototype vector [15]. The weights are basically proportional to the number and type of data that are projected into the corresponding unit, consequently, all the units without projected data are characterized by a similar prototype vector.

SKM defines a new matrix, called Clustering Matrix (CM), from SMap.

CM contains a number of elements that is equal to the number of IM elements, and the prototype vector of the corresponding activated unit defines each element.

The CM matrix and the Cluster number, evaluated from the SOM map interpretation, are the input data for the second level of the method.

2.1.3. K-Means Elaboration

As mentioned in the introduction, the second level of clustering is based on a K-Means algorithm.

K-Means is based on the concept of cluster centers, which are called 'centroids'. A centroid is a point in the data space that represents a cluster. The algorithm finds the positions of the cluster centroids in the input space, and minimizes an objective function E, the 'square-error distortion'.

After each data has been assigned, the centroid of each cluster has clearly changed, on the basis of the positions of the data in the space and on the random initial position of the centroid.

Therefore, a new cluster centroid is calculated in such a way that the sum of the squared distances is minimized.

The process continues with the calculation of the new distances between each input data and each centroid and re-assigning the data to the nearest centroid. This process is repeated until no more changes occur. In other words, the algorithm ends when all the data have been assigned to their nearest centroids.

The K-Means algorithm requires three user-specified parameters: A number of clusters *K*, cluster initialization and a distance metric.

The most critical choice is *K*. Although no perfect mathematical criterion exists, several heuristics criteria [30] are available to choose *K*.

The value of K in SKM is obtained from a SOM map visual evaluation. The CM matrix constitutes the input data for the K-Means algorithm.

The clustering phase provides a data partition that is summarized in a chart, where each occupational accident is attributed to a specific cluster, and a graphical output, dedicated to clustering visualization, is drawn, as shown in Figure 2.

The graph shows the distribution of activated units in the SOM map domain. Each unit is described by different colors, depending on the membership cluster. Each unit is marked by its own number (see the green circle in Figure 2), the number of projected elements (blue circle), and the cluster to which the unit belongs (red circle).

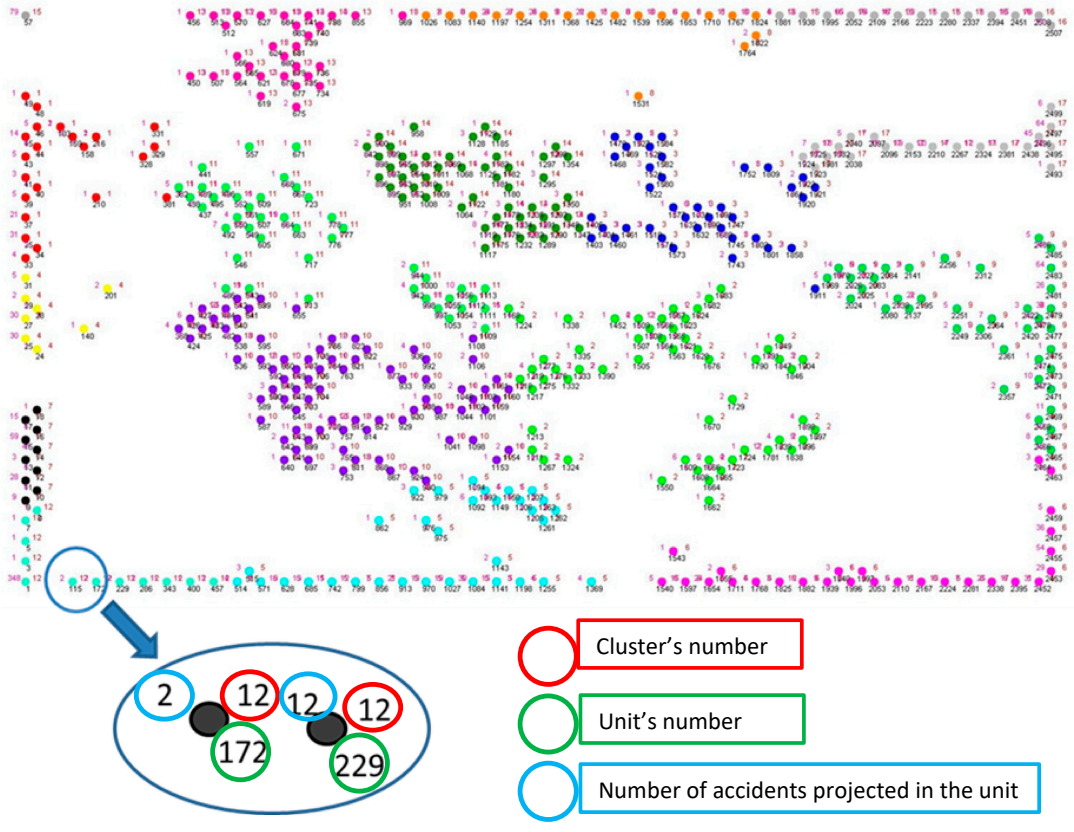

Figure 2. Partition output.

This graphical elaboration makes the comparison between several partitions easier, thus the evaluation of clustering accuracy becomes more immediate and intuitive.

With this visualization, it is also possible to carry out a comparison with the corresponding SOM map.

2.2. Case Study

This work has focused on the analysis of the occupational accident domain of the wood manufactory industry in the north of Italy (the Piedmont Region).

The occupational accident data set was provided by INAIL (Italian National Compensation Authority) and was made up of more than 6000 elements.

Unfortunately, some reports were inaccurate as a great deal of information was missing, and this required a preliminary check of all the available data.

The analysis of the accident database related to the wood manufacturing sector was carried out according to the following criteria:

1. The scope of the study was linked to the accident dynamics analysis in order to define preventive measures and, as a result, the selected descriptive variables were:

    - Activity;
    - Deviation;
    - Material of deviation;
    - Contact;
    - Injured body part;
    - Age of worker involved;

The first five variables were selected because they are closely linked to the accident event; the "Age of worker" was selected to investigate whether there was a possible correlation between the worker's age and the dynamics of the accident.

2. In order to be selected for the AM matrix definition, it was necessary for the first four variables to all be populated at the same time in the accident record.

On the basis of these two criteria, the original data set provided by INAIL led to an AM matrix of 4600 acceptable events.

2.2.1. Coding

The second step involves the transition from AM to IM matrix with the coding phase.

According to the criteria described at Section 2.1.1., 9 possible values were assumed for each variable and they were coded in a numerical sequence, as shown in Table 2; the whole coding table is reported in Appendix A (Ref. Table A1).

The dimension of the IM matrix, according to Equation (2), is:

$$D_{input} = 4600 \times 6 \times 9 = 248{,}400 \text{ cells}$$

2.2.2. SOM Elaboration and Analysis

The SOM was generated, according to the strategy to maximizing the map accuracy, as summarized hereafter:

1. The number of map units was set lower than the number of IM elements;
2. Several initialization seeds were tested, and the map was selected on the basis of a topographic error minimization criterion;
3. A balance between the elaboration time and accuracy was considered, according to the analyst's experience.

The SOM obtained for the case study with 25,000 seeds and 10,000 map units is shown in Figure 3. The visual analysis suggests the presence of at least 18 groups of similar occupational accidents. This value was used to set the *K* value required for the K-Means algorithm.

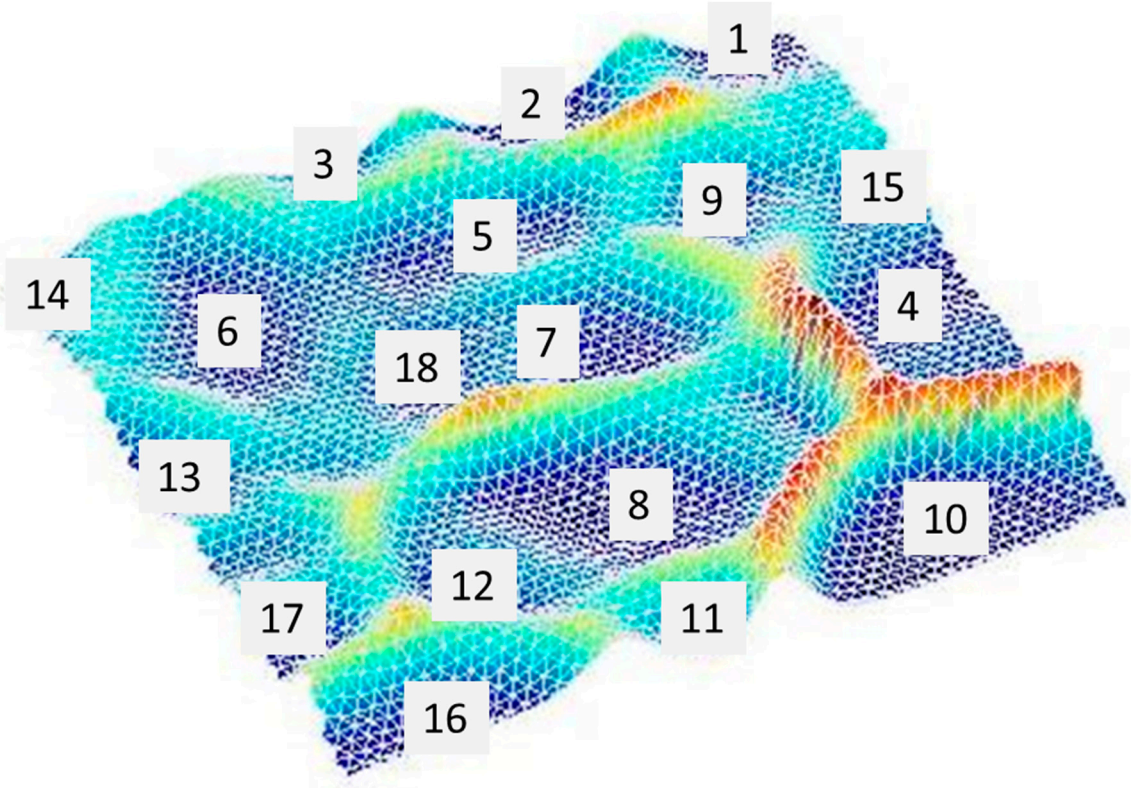

Figure 3. Self Organizing Map (SOM) of the Accident Matrix (AM) matrix based on 10,000 units.

2.2.3. K-Means Clustering and Cluster Identification

As discussed above, numerical clustering is an iterative process, and it was here started from the *K* value that was obtained from the SOM visual analysis.

The final result is a chart of all the accidents clustered into groups on the basis of their numerical similarity; furthermore, a graphic view of the partition is obtained, as shown in Figure 2.

Several independent repetitions of clustering can provide results with a level of variability in the accident cluster attribution that generally involves 8–15% of the data.

In order to manage this numerical variability, two indices can be adopted, as defined in Comberti et al. [7]: "Sequence stability" (Ss) and "sequence membership" (Sm).

The Sm index is calculated for each element. It represents the cluster attribution sequence of that element related to multiple repetitions.

The Ss index represents the number of elements that have the same Sm index.

Table 3 shows an example of the calculation of the Sm and Ss indices for a five element cluster.

Table 3. Sequence membership.

Record	**Clustering Repetition**						
	1°	2°	3°	4°	5°	6°	7°
5	A	A	A	A	A	A	A
2	A	A	A	A	B	A	C
3	A	A	A	A	B	A	A
4	A	A	A	A	B	A	A
1	A	A	A	A	A	A	A

The Sm index for record n. 5 is: AAAAAAA, while the Sm for record n. 2 is AAAABAC.

All the elements that have an Sm without any changes in attribution are represented by an Ss level of 100%. In other words, all the elements that are denoted by a stable sequence of clustering, have an Ss of 100%.

An Ss level equal to 85% corresponds to the number of elements that have an Sm with at least one variation in the cluster attribution.

A total of 85% of the examined data with a stable attribution had an Ss index level of 100%; the amount of stable attribution reached a coverage of 93% of the data for an Ss level equal to 85%.

The use of these indexes allows the clustering stability to be quantified and helps the analyst in the clustering identification. This process leads to a new definition of the clusters as a "group of elements with an assigned sequence stability".

Considering the AM matrix of 4600 occupational accidents in the wood manufacturing sector, and the SOM map obtained that suggested 18 clusters, the K-Means algorithm phase run on three repetitions led to a cluster identification of 21 groups on the basis of an Ss index of 85%, which is represented in Figure 4.

An total of 93% of the data were automatically included in the identified clusters.

Most of the remaining 7% was collocated by the SKM user in the different groups, depending on their level of similarity (272 element), and 78 elements, which were characterized by a very unstable attribution, were all included in a specific cluster called "Other".

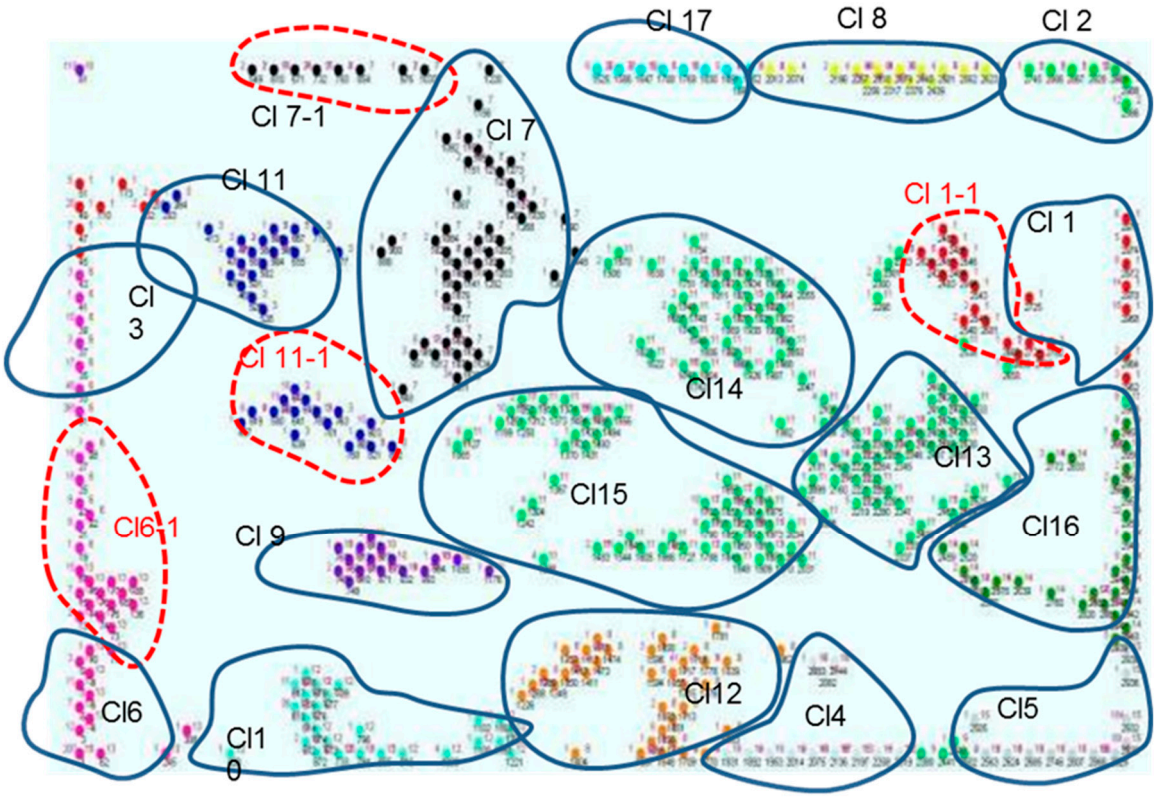

Figure 4. Cluster identification on the basis of the Sm and Ss indices.

3. Results

The application of SKM to the described data set led to the identification of 21 clusters. It was possible to describe all of the clusters according to the level of homogeneity of the data contained within each cluster. For example, Cluster 3 (CL3), which is summarized in Table 4, contains 486 accidents, 94% of which are characterized by "Working with hand tools" as their "Activity".

Table 4. CL3 description.

Variable	Category	%
Activity	Working with hand tools	94
Deviation	Losing control.	91
Deviation material	Hand tools	73
	No information available	15
Contact	Contact with cutting tool	83
Injured body part	Hands	82
Age	Various	-

A total of 91% of the "Deviation" variables is focused on "Losing control" and 73% of the Deviation Material" variables is focused on "Hand Tools". A total of 83% of the "Contact" variables is focused on "Contact with Cutting Tool" and 82% of the "Injured Body Part" is represented by "Hands".

Tables that show the clustering descriptions with a measure of their homogeneity are reported in the annex (Appendix B, Tables A2–A7): The most frequent values of the six descriptive variables selected in the problem definition phase are shown for each cluster.

Some other results could be found by analyzing the number of events of each cluster and the related average days of prognosis.

Figure 5 shows the number of occupational accidents allocated to each cluster. This parameter falls between a minimum value of 40 for cluster 1-1 to a value of 486 for cluster 3. This parameter can be used to estimate the major or minor frequencies of the accident dynamics pertaining to each cluster.

The "Other" label contains a set of heterogeneous accidents that were not assigned to any of the defined clusters.

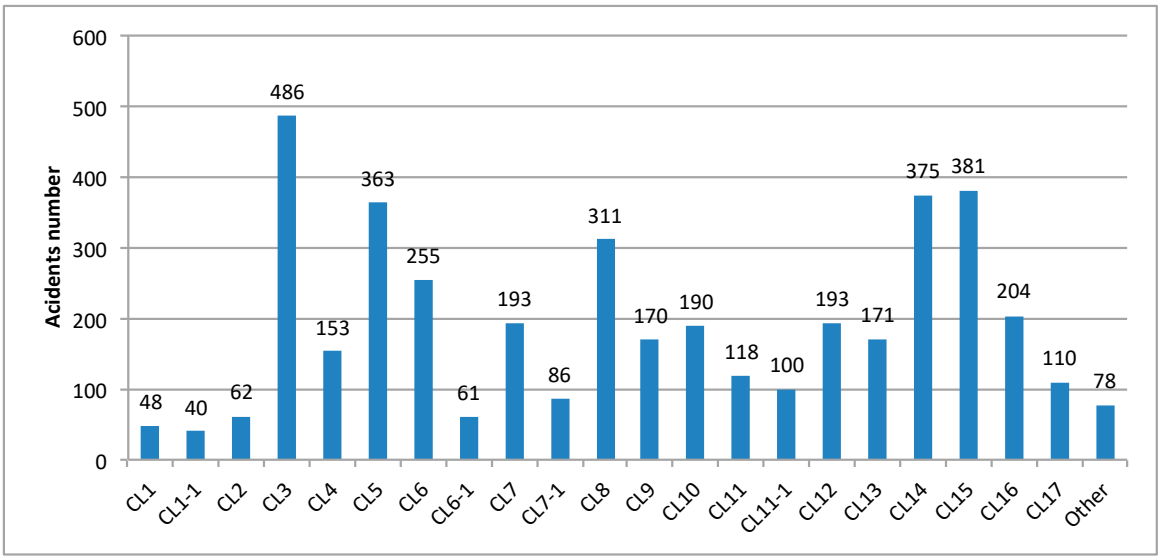

Figure 5. Number of events per cluster.

Figure 6 shows the average number of days of prognosis calculated for each cluster. This parameter showed a variability that ranged from 14.8 days/event for the "CL1-1" cluster to 54.5 days/event for the "CL17" cluster. The average days of prognosis may be used to express the severity of the accidents associated to each cluster, while the frequency of accidents and severity may be used to address preventive measures and policies for those clusters that are characterized by a higher risk.

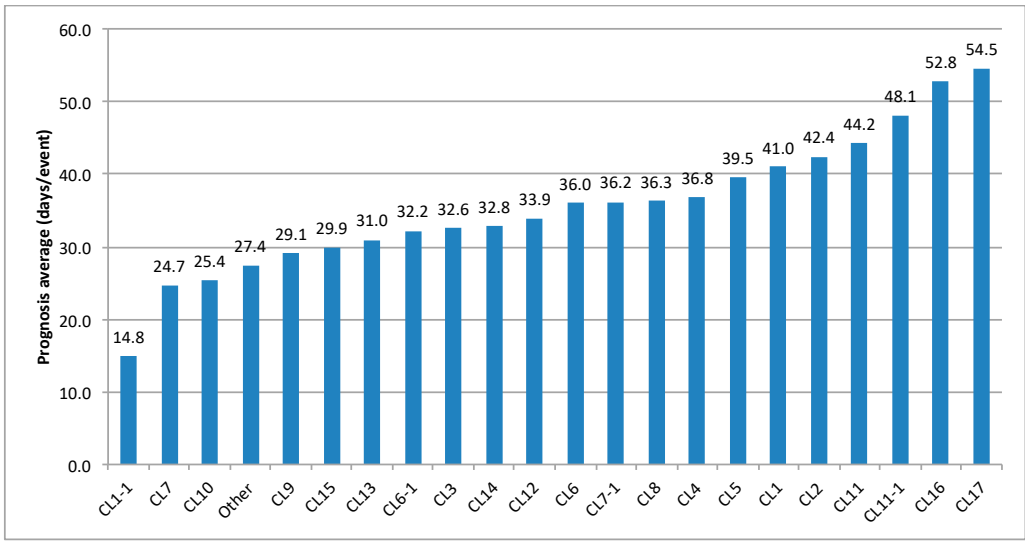

Figure 6. Average days of prognosis.

Figure 7 shows the average age of the workers. It allows the accident types to be associated with the age of the workers. Company managers could thus focus on preventive (as training) or protective (as personal protective devices) measures according to the average age of the workers on the basis of the most relevant accident dynamics characterizing the cluster.

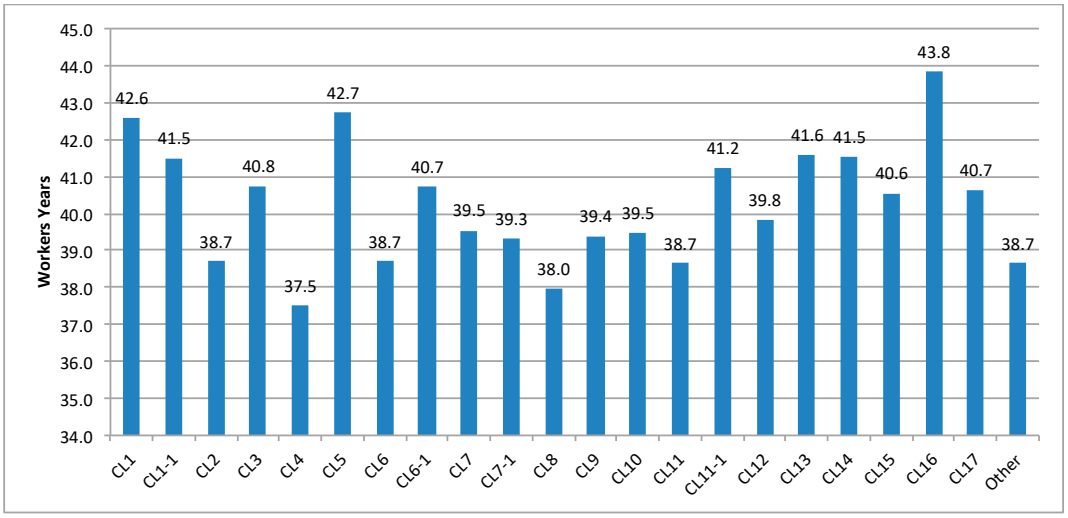

Figure 7. Average age of the workers of each cluster.

4. Discussion

4.1. Opportunities for Prevention: SKM Data Clustering

The results reported in the previous sections highlighted useful information about the ability of SKM to group occupational accidents into clusters.

As far as the cluster descriptions are concerned, Tables A2–A7 show that most of the 21 clusters can easily be characterized by 1 or 2 values of three of the six descriptive parameters, according to their numerousness within the element descriptors.

Activity, Deviation and Contact are generally polarized in one value, and in some cases, they can cover even 90% of the cluster elements, for example, the "CL10" cluster where 99% of the occupational accidents showed the "Handling" label for the "Activity" variable. "CL11" is less

polarized: The "Working with machinery" label covers 69% of the occupational accidents, while the "Manual transport" label covers 23%.

A more distributed division was observed for the "Deviation material", "Age" and "Injured body part" variables.

The results reported in the tables in Annex B suggest that SKM may be used to identify families of occupational accidents that differ according to their accidental dynamics, even though they share the same "Activity". For example, clusters 4, 5 and 16 had the same "activity" value: "Motion".

"CL4" grouped accidents characterized by "Stress movements" as main "Deviation" and "Physical effort" as "Contact". "CL5" grouped accidents characterized by "Fall" as major "Deviation" and "Crushing" for "Contact" and "CL16" identified accident dynamic similar to "CL5", but characterized by a "Contact" value that was polarized to "Contact with cutting tool".

The provided clustering description can easily be compared with additional information calculated for each cluster, with reference to the specific phenomenology of the wood industry.

Figure 5 shows the number of elements for each cluster. "CL3", "CL14" and "CL15" are characterized by the highest number of accidents.

This parameter can be assumed as an estimation of the frequency of accidents and, consequently, can be used to decide on the resources and measures necessary for those clusters identified as the most critical. Another piece of useful information that can be used to support Safety Managers is the average days of prognosis, as summarized in Figure 6.

As far as the above described "CL4", "CL5" and "CL16" clusters, which are taken as an example, are concerned, the average days of prognosis passed from 36.8 days/event ("CL4"—stress movements due to physical efforts) to 39.5 ("CL5"—Falls), and showed a maximum value of 52.8 days/event for "CL16", that is, occupational accidents due to contact with cutting tools. On the other hand, the "CL4" and "CL16" clusters are only moderately populated, while "CL5" is one of the most populated, thus the dynamics therein are among the most frequent in the wood industry. Moreover, with reference to Figure 7, it appears that the accidents resulting from contact with tools can be ascribed to older operators, while those related to movement can be attributed to the younger workers, thus the prevention and protective measures may also be addressed according to age.

According these results, the SKM method is able to distinguish groups of occupational accidents, characterized by different dynamics, and it is able to associate a different quantification of occupational accident frequency and seriousness to each group.

As a consequence, a Risk index was calculated according to the following equation:

$$R = F \times S, \tag{5}$$

where R is the risk, F is the frequency of occurrence, calculated as number of occupational accidents divided by day, and S is the seriousness, calculated as the average days of prognosis.

Equation (5) in Table 5 summarizes the Risk estimation for all the identified clusters.

Table 5. Risk assessment.

Clusters	Frequency (Event/Day)	Seriousness (Day/Event)	Risk
CL1	0.04	41	1.5
CL1-1	0.03	15	0.4
CL2	0.05	42	2.0
CL3	0.37	33	12.0
CL4	0.12	37	4.3
CL5	0.28	40	10.9
CL6	0.19	36	7.0
CL6-1	0.05	32	1.5
CL7	0.15	25	3.6

Table 5. *Cont.*

Clusters	Frequency (Event/Day)	Seriousness (Day/Event)	Risk
CL7-1	0.07	36	2.4
CL8	0.24	36	8.6
CL9	0.13	29	3.8
CL10	0.14	25	3.7
CL11	0.09	44	4.0
CL11-1	0.08	48	3.6
CL12	0.15	34	5.0
CL13	0.13	31	4.0
CL14	0.28	33	9.3
CL15	0.29	30	8.6
CL16	0.15	53	8.2
CL17	0.08	55	4.5
Other	0.06	27	1.6

Risk shows a wide range of variation, that is, from 1.6 for "CL1-1" to 12 for "CL3".

SKM has been able to identify clusters of accidents in the wood industry and to classify them, in terms of minor or greater risk levels. For example, the most critical clusters were "CL3" and "CL5", which are related to manual work with hand-tools ("CL3") and to falls during manual transport or movements ("CL5"). The association of a Risk assessment to each cluster may in fact represent a support to any decision-making process focused on preventive measurement planning.

For example, the high risk of "CL5" suggests there is a need to review the design of the workplace organization in order to optimize the workers' movements inside the working area.

4.2. Opportunities for Prevention: Traditional Data Analysis

Economic and technical resources can be defined to prevent occupational accidents on the basis of the information achievable with the SKM method.

This result cannot be achieved directly with a traditional statistical approach, as mentioned in the introduction. In fact, a statistical analysis performed on an occupational accident database pertaining to the wood industry [31] provided many diagrams and graphical views of the distribution of all the variables used in an ESAW classification. However, this large amount of information did not lead to the identification of occupational accident clusters and did not have the purpose of drawing up a risk quantification, as SKM did. An example of this is shown in Figure 8, where the distribution of three variables that affected the accident dynamics is reported.

Compared to other ESAW data mining techniques, such as MCA [8], the use of SKM offers two main advantages:

1. The here performed SKM analysis was based on six parameters (as described in Section 2.2), but all the other accident details included in the database remained linked to each single accident and could be used to describe the identified clusters. This was done, in the proposed study, with "days of prognosis parameters" and it led to a risk assessment classification, but it could also be done with all the other connected parameters, such as "number of workers employed", "time of accident occurrence", and so on, thus making it possible to conduct several quantified analyses.
2. SKM is a friendly-user method, as it does not require any specific expertise in statistics or data analysis. In fact, once the data set has been coded automatically to the SKM required format, the SKM user simply has to set the number of "SOM units", the number of interaction cycles, and the number of clusters into which dividing the data set should be divided on the basis of the SOM map. This makes the SKM method easier to apply to ESAW data than other more complex data mining techniques.

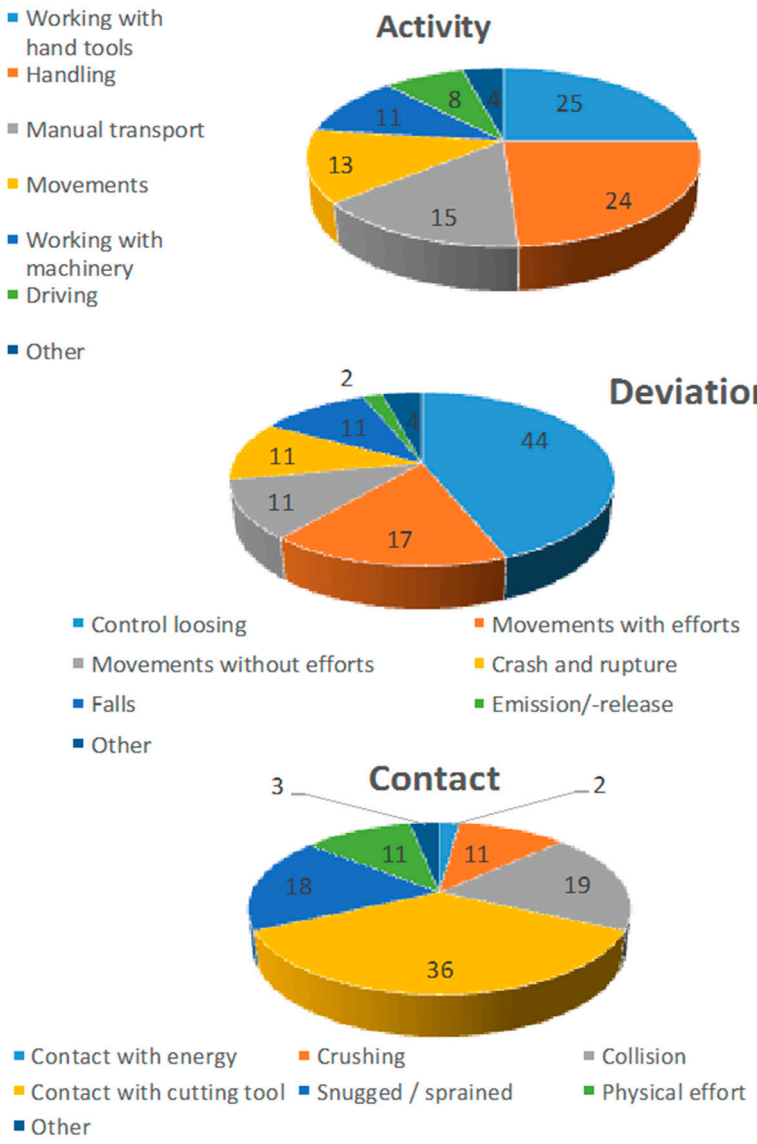

Figure 8. Distribution of the dynamics variables of the wood industry for the Veneto occupational accident database.

5. Conclusions

This paper has focused on the validation of a numerical methodology to deal with an occupational accident database (DB) in order to better address the data analysis, to achieve a reduction in risks and to support the definition of preventive measures.

A data set of more than 4000 occupational accidents that had occurred in the wood industry was selected as a case study, and it was analyzed with the SKM method. SKM was able to successfully identify a set of 21 clusters of accidents based on six variables related to the occurrence dynamics, the injured body part and the age of the involved workers.

The variable distribution of each cluster highlighted that the partition was steered by the four dynamic-related ones, while the variable distributions of the age of the workers and of the injured body part were observed to be more scattered. Some other parameters related to the consequences of each accident (number of days of prognosis) and the number of events (number of accidents) were calculated and associated to each cluster, and this allowed a Risk assessment evaluation to be made.

The two most critical clusters, according to the risk assessment, were related to "manual activity with hand tools" and to "free movements/manual transport" in the working area. This information

suggests, for example, the need to design a different working organization in order to reduce the workers' movements inside the working place.

The results highlight that the proposed methodology represents an advancement in the analyses of occupational accident DBs, since it allows not only the distribution of single parameters, such as statistics, to be identified, but also to rank the dynamics of families of accidents according to such relevant parameters as severity or risk.

More in general, the SKM can help Company Management and the National Authorities to address preventive measures and policies pertaining to those clusters that have been identified as the most critical on the basis of the risk quantification. This additional information represents a useful piece of knowledge that can be used to support risk-based decision-making processes, because it represents a quantification of risk linked to the defined occupational accident groups.

Author Contributions: Conceptualization: M.D. and L.C.; Methodology: L.C. and M.D.; Software: G.B. and L.C.; Validation: G.F. and R.L.; Formal Analysis: all the authors; Data Curation: G.F., R.L., G.B. and L.C.; Writing-Original Draft Preparation and Writing-Review and Editing: all the authors. Supervision: M.D.

Funding: This research was funded by INAIL, Direzione Regionale del Piemonte, grant number Protocollo Accordo INAIL CS&P.

Acknowledgments: The authors gratefully acknowledge the support of INAIL, Direzione Regionale del Piemonte, and in particular its Director, Alessandra Lanza, who believed in the project and allowed this research to be carried out under the CS&P–Centro Studi su Cultura della Sicurezza e Prevenzione, co-financed by INAIL–Direzione Regionale del Piemonte and the Politecnico di Torino.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AM	Accident Matrix
BMU	Best Matching Unit
CM	Clustering Matrix
CL	Cluster
DB	Database
EU	European Union
ESAW	European Statistics on Accidents at Work
INAIL	Istituto Nazionale Assicurazione contro gli Infortuni sul Lavoro
IM	Input Matrix
Sm	Sequence membership
Ss	Sequence stability
SOM	Self Organizing Map
SKM	SOM K-Mean Method

Appendix A

In this Appendix is reported the encoding table used for the data analyses.

Table A1. Coding table for ESAW variable.

Variable	Categories	Numeral Coding								
Activity	Working with machinery	1	0	0	0	0	0	0	0	0
	Working with hand tools	0	1	0	0	0	0	0	0	0
	Driving	0	0	1	0	0	0	0	0	0
	Handing	0	0	0	1	0	0	0	0	0
	Not compiled	0	0	0	0	1	0	0	0	0
	Presence	0	0	0	0	0	1	0	0	0
	Manual transport	0	0	0	0	0	0	1	0	0
	Movement	0	0	0	0	0	0	0	1	0
	Other	0	0	0	0	0	0	0	0	1

Table A1. *Cont.*

Variable	Categories	Numeral Coding								
Deviation	Energy release (fire, explosion, . . .)	1	0	0	0	0	0	0	0	0
	Release	0	1	0	0	0	0	0	0	0
	Material breaking	0	0	1	0	0	0	0	0	0
	Control loosing	0	0	0	1	0	0	0	0	0
	Fall	0	0	0	0	1	0	0	0	0
	Incorrect movement	0	0	0	0	0	1	0	0	0
	Stress movement	0	0	0	0	0	0	1	0	0
	Violence or surprise	0	0	0	0	0	0	0	1	0
	Not information	0	0	0	0	0	0	0	0	1
Deviation material	Surface	1	0	0	0	0	0	0	0	0
	Stored and carved materials	0	1	0	0	0	0	0	0	0
	Absence of deviation material	0	0	1	0	0	0	0	0	0
	Hand tools	0	0	0	1	0	0	0	0	0
	Machinery	0	0	0	0	1	0	0	0	0
	Transport system	0	0	0	0	0	1	0	0	0
	Scraps, dangerous product	0	0	0	0	0	0	1	0	0
	Person or animal	0	0	0	0	0	0	0	1	0
	No information available	0	0	0	0	0	0	0	0	1
Contact	Contact with energy	1	0	0	0	0	0	0	0	0
	Crushing	0	1	0	0	0	0	0	0	0
	Impact with pitched material	0	0	1	0	0	0	0	0	0
	Collision with transport system	0	0	0	1	0	0	0	0	0
	Contact with cutting tool	0	0	0	0	1	0	0	0	0
	Snugged/sprained	0	0	0	0	0	1	0	0	0
	Physical effort	0	0	0	0	0	0	1	0	0
	Violent bump	0	0	0	0	0	0	0	1	0
	No information	0	0	0	0	0	0	0	0	1
Injured body part	Head/neck	1	0	0	0	0	0	0	0	0
	Internal body parts	0	1	0	0	0	0	0	0	0
	Spinal column	0	0	1	0	0	0	0	0	0
	Arms	0	0	0	1	0	0	0	0	0
	Hands	0	0	0	0	1	0	0	0	0
	Legs	0	0	0	0	0	1	0	0	0
	Feet	0	0	0	0	0	0	1	0	0
	Eyes and ears	0	0	0	0	0	0	0	1	0
	Chest	0	0	0	0	0	0	0	0	1
Age	Under 18	1	0	0	0	0	0	0	0	0
	19–27	0	1	0	0	0	0	0	0	0
	28–35	0	0	1	0	0	0	0	0	0
	36–44	0	0	0	1	0	0	0	0	0
	45–55	0	0	0	0	1	0	0	0	0
	55–60	0	0	0	0	0	1	0	0	0
	61–70	0	0	0	0	0	0	1	0	0
	Over 70	0	0	0	0	0	0	0	1	0
	No information	0	0	0	0	0	0	0	0	1

Appendix B

In this Appendix the clusters description is reported.

Table A2. Clusters description (Activity).

Cluster	First Category	%	Second Category	%
CL1	Manually transport	35	Working with hand tools	25
CL1-1	Handling	70		
CL2	Driving	63	Movements	19
CL3	Working with hand tools	94		
CL4	Movements	83	Manual transport	13
CL5	Movements	71	Manual transport	12
CL6	Working with hand tools	31	Working with machinery	25
CL6-1	Working with hand tools	89		
CL7	Working with hand tools	98		
CL7-1	Working with hand tools	95		
CL8	Driving	95		
CL9	Manual transport	98		
CL10	Handling	99		
CL11	Working with machinery	69	Manual transport	23
CL11-1	Working with machinery	94		
CL12	Manual transport	58	Handling	37
CL13	Handling	89		
CL14	Handling	92		
CL15	Handling	78		
CL16	Movements	69	Handling	20
CL17	Driving	95		

Table A3. Clusters description (Deviation).

Cluster	First Category	%	Second Category	%
CL1	Material breaking	98		
CL1-1	Release	73	Material breaking	23
CL2	Violence or surprise	100		
CL3	Control loosing.	91		
CL4	Stress Movement	79	Incorrect movement	21
CL5	Fall	93		
CL6	Incorrect movement	100		
CL6-1	Incorrect movement	100		
CL7	Control loosing.	98		
CL7-1	Control loosing.	98		
CL8	Control loosing.	93		
CL9	Control loosing.	61	Stress Movement	14
CL10	Incorrect movement	87		
CL11	Control loosing.	86		
CL11-1	Control loosing.	79	Incorrect movement	17
CL12	Stress Movement	90		
CL13	Material breaking	85		
CL14	Material breaking	94		
CL15	Material breaking	84		
CL16	Fall	53	Material breaking	38
CL17	Control loosing.	84		

Table A4. Clusters description (Deviation material).

Cluster	First Category	%	Second Category	%
CL1	Scraps, dangerous product	40	Surfaces	29
CL1-1	Scraps, dangerous product	93		
CL2	Absence of deviation material	50	Person or animal	19
CL3	Hand tools	73	No information available	15
CL4	Absence of deviation material	54	Surfaces	26
CL5	Surfaces	58	Stored and carved materials	13
CL6	Absence of deviation material	78		
CL6-1	Hand tools	60	No information available	10
CL7	Stored and carved materials	30	Surfaces	21
CL7-1	Stored and carved materials	73		
CL8	Transport system	95		
CL9	Stored and carved materials	81		
CL10	Absence of deviation material	72	No information available	11
CL11	Machinery	36	Stored and carved materials	31
CL11-1	Machinery	43	No information available	42
CL12	Stored and carved materials	48	No information available	13
CL13	Scraps, dangerous product	52	Hand tools	15
CL14	Hand tools	43	Scraps, dangerous product	25
CL15	Stored and carved materials	99		
CL16	Surfaces	79		
CL17	Transport system	91		

Table A5. Clusters description (Contact).

Cluster	First Category	%	Second Category	%
CL1	Impact with pitched material	88		
CL1-1	Contact with energy	73	Impact with pitched material	23
CL2	Collision with transport system	55	Violent bump	16
CL3	Contact with cutting tool	83		
CL4	Physical effort	87		
CL5	Crushing	99		
CL6	Contact with cutting tool	59		19
CL6-1	Contact with cutting tool	69	Snugged/sprained	11
CL7	Contact with cutting tool	43	Collision with transport system	20
CL7-1	Impact with pitched material	70		
CL8	Collision with transport system	89		
CL9	Impact with pitched material	39	Snugged/sprained	20
CL10	Contact with cutting tool	41	Crushing	16
CL11	Snugged/sprained	47	Contact with cutting tool	27
CL11-1	No information	56	Violent bump	33
CL12	Physical effort	95		
CL13	Contact with cutting tool	89		
CL14	Contact with cutting tool	68	Collision with transport system	11
CL15	Contact with cutting tool	67	Impact with pitched material	13
CL16	Contact with cutting tool	82		
CL17	Crushing	67	Contact with cutting tool	20

Table A6. Clusters description (Injured body part).

Cluster	First Category	%	Second Category	%
CL1	Arms	25	Chest	21
CL1-1	Eyes and ears	60	Head/neck	15
CL2	Hands	27	Scattered	
CL3	Hands	82		
CL4	Legs	56	Hands	18
CL5	Scattered			
CL6	Hands	79		
CL6-1	Hands	77		
CL7	Hands	68		
CL7-1	Scattered		Scattered	
CL8	Spinal column	47	Hands	16
CL9	Hands	54		
CL10	Hands	53	Arms	12
CL11	Hands	81		
CL11-1	Hands	92		
CL12	Spinal column	36	Hands	15
CL13	Scattered			
CL14	Hands	91		
CL15	Hands	50		
CL16	Legs	38	Scattered	
CL17	Legs	25	Hands	24

Table A7. Clusters description (Age).

Cluster	First Category	%	Second Category	%
CL1	36–44	33	Scattered	
CL1-1	36–44	35	44–55	23
CL2	Scattered			
CL3	Scattered			
CL4	36–44	37	Scattered	
CL5	Scattered			
CL6	Scattered			
CL6-1	36–44	31	Scattered	
CL7	45–55	31	Scattered	
CL7-1	36–44	30	Scattered	
CL8	Scattered			
CL9	Scattered			
CL10	36–44	28	44–55	25
CL11	28–35	30		
CL11-1	36–44	31		
CL12	45–55	29	36–44	28
CL13	Scattered			
CL14	Scattered			
CL15	Scattered			
CL16	Scattered			
CL17	Scattered			

References

1. Hamalainen, P.; Takala, J.; Saarela, K.L. Global estimates of occupational accidents. *Saf. Sci.* **2006**, *44*, 137–156. [CrossRef]
2. EUROSTAT. *European Statistics on Accidents at Work (ESAW)—Summary Methodology*; Publications Office of the European Union: Luxembourg, Luxembourg, 2013.
3. Jacinto, C.; Soares, G.C. The added value of the new ESAW/Eurostat variables in accident analysis in the mining and quarrying industry. *J. Saf. Res.* **2008**, *39*, 631–644. [CrossRef] [PubMed]

4. Dźwiarek, M.; Latała, A. Analysis of occupational accidents: prevention through the use of additional technical safety measures for machinery. *Int. J. Occup. Saf. Ergon.* **2016**, *22*, 186–192. [CrossRef] [PubMed]
5. Kogler, R.; Quendler, E.; Boxberger, J. Analysis of occupational accidents with agricultural machinery in the period 2008–2010 in Austria. *Saf. Sci.* **2015**, *72*, 319–328. [CrossRef]
6. Palamara, F.; Piglione, F.; Piccinini, N. Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. *Saf. Sci.* **2011**, *49*, 1215–1230. [CrossRef]
7. Comberti, L.; Demichela, M.; Baldissone, G. Workplace Accidents Analysis with a Coupled Clustering Methods: S.O.M. and K-means Algorithms. *Chem. Eng. Trans.* **2015**, *43*, 1261–1266. [CrossRef]
8. Carrillo-Castrillo, J.A.; Rubio-Romero, J.C.; Guadix, J.; Onieva, L. Identification of areas of intervention for public safety policies using multiple correspondence analysis. *DYNA* **2016**, *83*, 31–37. [CrossRef]
9. Edelstein, H. *Introduction to Data Mining and Knoledge Discovery*, 3rd ed.; Two Crow Corporation: Potomac, MD, USA, 1999; ISBN 1-892095-02-5.
10. Larose, D.T. *Discovering Knowledge in Data-An Introduction to Data Mining*; John Wiley & Sons Inc.: New York, NY, USA, 2005.
11. Silva, J.F.; Jacinto, C. Finding occupational accident patterns in the extractive industry using a systematic data mining approach. *Reliab. Eng. Syst. Saf.* **2012**, *108*, 108–122. [CrossRef]
12. Gevrey, M.; Worner, S.; Kasabov, N.; Pitt, J.; Giraudel, J.L. Estimating risk of events using SOM models: A case study on invasive species establishment. *Ecol. Model.* **2006**, *197*, 361–372. [CrossRef]
13. Liang, W.; Hua, J.Z.; Guo, C.; Lin, W. Assessing and classifying risk of pipeline third-party interference based on fault tree and SOM. *Eng. Appl. Artif. Intell.* **2012**, *25*, 594–608. [CrossRef]
14. Asgary, A.; Sadeghi Naini, A.; Levy, J. Modeling the risk of structural fire incidents using a self-organizing map. *Fire Saf. J.* **2012**, *49*, 1–9. [CrossRef]
15. Vesanto, J.; Alhoniemi, E. Clustering of self-organizing map. *IEEE Trans. Neural Netw.* **2000**, *11*, 586–600. [CrossRef] [PubMed]
16. Demichela, M.; Pirani, R.; Leva, M.C. Human factor analysis embedded in risk assessment of industrial machines: Effects on the safety integrity level. *Int. J. Perform. Eng.* **2014**, *10*, 487–496.
17. Murè, S.; Comberti, L.; Demichela, M. How harsh work environments affect the occupational accident phenomenology? Risk assessment and decision making optimisation. *Saf. Sci.* **2017**, *95*, 159–170. [CrossRef]
18. Comberti, L.; Baldissone, G.; Demichela, M. A combined approach for the analysis of large occupational accident databases to support accident-prevention decision making. *Saf. Sci.* **2018**, *106*, 191–202. [CrossRef]
19. Top, Y.; Adanur, H.; Öz, M. Comparison of practices related to occupational health and safety in microscale wood-product enterprises. *Saf. Sci.* **2016**, *82*, 374–381. [CrossRef]
20. Leva, M.C.; Pirani, R.; Demichela, M.; Clancy, P. Human factors issues and the risk of high voltage equipment: Are standards sufficient to ensure safety by design? *Chem. Eng. Trans.* **2012**, *26*, 273–278. [CrossRef]
21. Darabnia, B.; Demichela, M. Data field for decision making in maintenance optimization: An opportunity for energy saving. *Chem. Eng. Trans.* **2013**, *33*, 367–372. [CrossRef]
22. Darabnia, B.; Demichela, M. Maintenance an opportunity for energy saving. *Chem. Eng. Trans.* **2013**, *32*, 259–264. [CrossRef]
23. Gerbec, M.; Balfe, N.; Leva, M.C.; Prast, S.; Demichela, M. Design of procedures for rare, new or complex processes: Part 1—An iterative risk-based approach and case study. *Saf. Sci.* **2017**, *100*, 195–202. [CrossRef]
24. Gerbec, M.; Baldissone, G.; Demichela, M. Design of procedures for rare, new or complex processes: Part 2—Comparative risk assessment and CEA of the case study. *Saf. Sci.* **2017**, *100*, 203–215. [CrossRef]
25. Leva, M.C.; Balfe, N.; Kontogiannis, T.; Plot, E.; Demichela, M. Total safety management: What are the main areas of concern in the integration of best available methods and tools. *Chem. Eng. Trans.* **2014**, *36*, 559–564. [CrossRef]
26. Leva, M.C.; Kontogiannis, T.; Balfe, N.; Plot, E.; Demichela, M. Human factors at the core of total safety management: The need to establish a common operational picture. In Proceedings of the Contemporary Ergonomics and Human Factors, Daventry, UK, 13–16 April 2015; pp. 163–170.
27. Kangas, J.; Kohonen, T.K.; Laaksonen, J. Variants of self-organizing maps. *IEEE Trans. Neural Netw.* **1990**, *1*, 93–99. [CrossRef] [PubMed]
28. Lourenço, F.; Lobo, V.; Bação, F. Binary-based similarity measures for categorical data and their application in Self-Organizing Maps. In *Internal Report: Instituto Superior de Estatística e Gestão de Informação*; Universidade Nova de Lisboa: Lisbon, Portugal, 2004.

29. Demichela, M.; Palamara, F. Occupational accidents risk analysis using clustering algorithms. In Proceedings of the European Safety and Reliability Conference 2007, ESREL 2007—Risk, Reliability and Societal Safety, Stavanger, Norway, 25–27 June 2007; pp. 1261–1265.
30. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *R. Stat. Soc.* **2001**, *63*, 411–423. [CrossRef]
31. Sarto, F.; Agensi, R.; Veronese, M. *Gli Infortuni sul Lavoro e le Malattie Professionali nel Comparto Industria del LEGNO*; Regione del Veneto. C.O.R.E.O.Centro Stampa: Venezia, Italy, 2009.