Mining, Metallurgy & Exploration (2024) 41:1693–1708
https://doi.org/10.1007/s42461-024-01024-z

An Information Entropy–based Risk (IER) Index of Mining Safety Using Clustering and Statistical Methods

Dharmasai Eshwar[1] · Snehamoy Chatterjee[1] · Rennie Kaunda[2] · Hugh Miller[2] · Aref Majdara[3]

Received: 25 January 2024 / Accepted: 31 May 2024 / Published online: 26 June 2024

Abstract
In recent decades, the mining industry in the United States has made significant strides in reducing accidents and injuries. While these improvements are commendable, interpreting these statistics can be challenging due to concurrent declines in workforce size, employee hours, productivity, and operating systems. The Mine Safety and Health Administration (MSHA) of the United States has instituted tools like the Pattern of Violation (POV) and Significant & Substantial (S&S) calculator to monitor safety in mines. However, both have their respective limitations. Various risk indices have been proposed to address these limitations, leveraging multiple matrices from MSHA databases. Yet, the primary challenge lies in effectively integrating these diverse matrices into a cohesive risk index. This research endeavors to develop an information entropy–based risk (IER) index through the optimization of weights assigned to these sometimes-conflicting matrices. The seven-dimensional risk indicators considered for IER index computation encompass (a) citations, (b) orders, (c) significant & substantial citations, (d) penalties, (e) incidents with no lost time, (f) lost time injuries, and (g) proposed penalty for violation. The efficacy of the proposed IER index was assessed using data from MSHA's underground mines spanning from 2011 to 2020. Validation of the IER index was conducted through application of the BIRCH clustering algorithm in tandem with rigorous statistical analysis. The clustering performance was evaluated using the multivariate analysis of variance (MANOVA) test, followed by post hoc analysis. Box plots and univariate analysis of variance (ANOVA) tests were then employed to substantiate the statistical significance of mean differences in IER index values across clusters. The MANOVA test and subsequent post hoc results underscore the successful clustering of the seven-dimensional risk indices across all time periods using the BIRCH algorithm. The ANOVA test unequivocally demonstrates that the mean risk index values of at least one cluster are statistically distinct from the others at a 95% confidence level for all periods. Post hoc analysis further confirms the statistical significance of differences in mean risk indices between clusters. These findings were further supported by the results obtained from the box plots. Finally, the proposed approach was applied to an underground coal mine to illustrate its practical effectiveness. This study demonstrates that the proposed approach can empower mining companies to comprehensively assess their safety performance and implement necessary measures for improvement.

Keywords Mine safety · BIRCH clustering · Entropy · MANOVA · Post hoc · MSHA

1 Introduction

Mining involves the extraction of economically valuable minerals from veins, seams, or other types of orebodies, playing a pivotal role in various industries such as steel production, power generation, electronics, construction products, and agriculture [14]. Despite its significant contributions, mining stands out as one of the most hazardous industrial environments, posing risks to the safety and health of workers from hazards like dust inhalation, roof collapses, explosions, rockslides, and power haulage accidents [41]. In fact, mining exhibits higher incidence rates

✉ Snehamoy Chatterjee
schatte1@mtu.edu

[1] Michigan Technological University, 625 Dow Environmental Sciences, 1400 Townsend Drive, Houghton, MI 49931, USA

[2] Colorado School of Mines, 1600 Illinois St., Colorado School of Mines, Golden, CO 80401, USA

[3] Washington State University, 14204 NE Salmon Creek Ave, Vancouver, WA 98686, USA

of injuries and fatalities compared to many other industries [33]. The fatal injury rate in mining in the United States (US) for full-time equivalent (FTE) workers per 100,000 saw a notable decrease from 24.18 in 2006 to 11.77 in 2022 (NIOSH, 2024). In 2022, the fatality rate in mining in the US was 11.7, significantly higher than the average work-related fatality rate of 3.7 across all industries in the US (NIOSH, 2024,OSHA, 2024). The US Mine Safety and Health Administration (MSHA) diligently monitors fatalities, injuries, accidents, and violations in the mining industry. Figure 1 illustrates the fatality rate in mining in the US from 2011 through 2020 (NIOSH, 2024). However, these figures can be misleading due to the concurrent decline in workforce size and employee hours from 2015 to 2022 (Fig. 1). Therefore, to mitigate mining accidents, injuries, and fatalities, conducting a comprehensive qualitative and quantitative risk analysis is crucial in identifying potential risk-related incidents [12, 43].

The US Mine Safety and Health Administration (MSHA) has implemented various provisions to enhance miner safety. MSHA has developed an interactive portal focused on mine health and safety standards, which includes comprehensive databases covering accidents, inspections, violations, citations, and production (MSHA, 2024). Additionally, MSHA has introduced monitoring tools such as the Pattern of Violation (POV) and Significant & Substantial (S&S) calculator (MSHA, 2024). The POV utilizes two sets of screening criteria to conduct the review required under 30 CFR §104.2, while the S&S calculator indicates whether the inspector determines the severity to be significant and substantial. These tools can be beneficial for mining companies, but they have limitations. Research by Kinilakodi and Grayson [19] has highlighted a limitation of the POV tool, as it depends on the disposition of the final citation, which involves ten components and can be challenging to calculate and implement. Moreover, the POV provides only a binary answer (YES/NO), making it difficult for mine operators and contractors to use it for risk management and continuous improvement compared to regulatory compliance. In contrast, the S&S calculator is straightforward, provides a quantitative value, and is simple to calculate. However, it only considers two parameters (the number of S&S citations and inspection hours), ignoring other citations and orders.

Incidences involving death or injury are influenced by various contributing factors, and metrics provide valuable insights into the multifaceted nature of these incidents. According to Title 30 of the Code of Federal Regulations (CFR), Part 50, mine operators and contractors are mandated to submit detailed reports of reportable incidents using MSHA Form 7000–1 (Title-30, Chapter-I, Subchapter-I, Part-50, n.d.). Incidents within mines are categorized into two main types: those resulting in no lost-time injuries (NLT) and those leading to lost-time injuries (LT). Analyzing data related to lost-time injuries is crucial in preventing fatalities. Further classification of LT and NLT injuries distinguishes between nonfatal days-lost injuries (NFDL) and fatal and no-days-lost injuries (NDL) (MSHA, 2023). Researchers have applied these tools to understand the potential risks of mining accidents [40, 36, 12, 22]. Research indicates that increased penalties for violations play a crucial role in enhancing health and safety standards within mines [18]. Elevated citations for specific violations, particularly those categorized as S&S and orders, can signify a failure in risk management and lead to a decline in safety

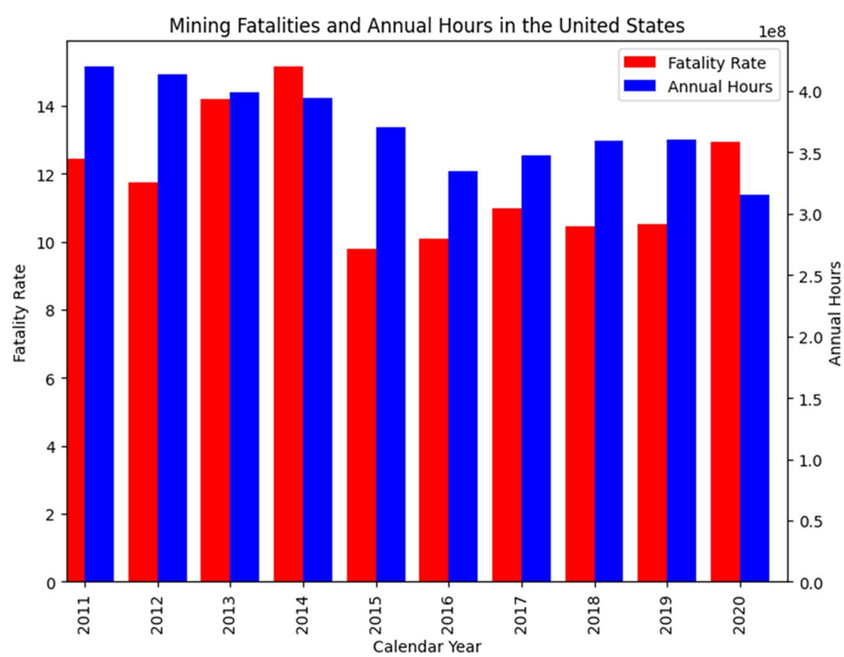

Fig. 1 Mining fatalities and annual hours in the United States

performance [11]. The frequency of occurrence serves as a normalized safety measure, while severity is gauged by the relative extent of associated losses [21]. Historically, these tools have been utilized in the mining industry to measure potential safety issues in mines [9, 46].

To address the limitations of existing tools, Kinilakodi and Grayson [20] proposed the Safe Performance Index (SPI) to assess the overall safety performance of a mine. Unlike previous tools, the SPI places greater emphasis on more severe citations [19]. The SPI, typically expressed as a percentage, allows mining companies to track their safety performance over time and compare it to industry benchmarks. A higher SPI indicates better safety performance, while a lower SPI may highlight areas for improvement in safety practices and procedures. However, the SPI assigns equal weights to accident and citation measures for assessing mine safety performance [20], overlooking the uncertainty associated with each measure. Although the SPI provides a quantitative value for safety performance, Kinilakodi et al. [21] developed a traditional risk analysis tool from the SPI using variable weight values. Nonetheless, the weights were assigned based on subjective judgment rather than evaluating the uncertainty of each index associated with accident and citation measures.

The present study aims to propose a risk index based on safety indicators, enabling mining companies to assess mine safety performance. An information entropy–based risk index (IER) is introduced to analyze the relative significance of various factors based on their entropy values and allocate weights accordingly. The IER is scrutinized using MSHA databases to optimize the weights of conflicting safety indices. Information entropy is crucial for weight calculation when combining multiple indices to develop a new index [26, 34]. For mining risk index development, information entropy helps quantify the uncertainty or unpredictability associated with each safety index. By assigning weights based on the information entropy of each safety index, we can effectively prioritize indices that provide more informative and less redundant information. Validation of the IER was conducted by identifying clusters with relative risk. Consequently, it becomes feasible to compute the IER over time, unveiling trends in the evolution of risk associated with specific mining operations. Thus, discerning the new risk index (i.e., IER) could offer deeper insights into the risk profile of the mining industry.

2 Materials

MSHA's open data portal contains a wealth of information on production (e.g., coal), employment, and injury and fatality statistics for every permitted, active mine in the United States (MSHA, 2024). The data also includes details on inspections, citations, and violations, dating back to inspections initiated on 1/1/2000, each with a unique event number linked directly to the inspection dataset.

For our research, we utilized the violation, inspection, production, and accident datasets from the MSHA open data website (https://arlweb.msha.gov/OpenGovernmentData/OGIMSHA.asp), processing them in a Python environment. All data and code used in this research are available online (link provided later in this paper). The violation data provided citations, orders, S&S information, and proposed penalties, which were integral to our research. Inspection data helped us calculate inspection hours, which were crucial for calculating violation-related risk indices. Production data was used to calculate employee hours, which in turn helped calculate accident-related indices, such as the no-days-lost incidence rate (NDL IR), nonfatal days-lost incidence rate (NFDL IR), and severity measure (SM).

The accident dataset categorizes the degrees of injury into ten levels, including fatal, permanent total or partial disability, nonfatal with days lost only, nonfatal with days lost and days of restricted work activity, nonfatal with restricted work activity only, and nonfatal with no days lost or restricted. This information, along with employee hours' data from the production dataset, was used to calculate the NDL IR, NFDL IR, and SM.

Our research primarily focuses on underground mining. Therefore, we specifically utilized underground data from 2011 to 2020 (MSHA, 2024) for both violation-related risk indices and accident-related risk indices. Subsequently, we calculated the IER index for each year using violation-related risk indices, accident-related risk indices, and an information entropy approach, as detailed in Sect. 3 of our research. Table 1 summarizes the different datasets used and their purposes in our study.

3 Methodology

3.1 Citation and Injury-related Risk Index Calculation

In this research, seven different risk indices were calculated to evaluate the safety performance of each underground operating mine in the US. Four of these indices are related to violation data, while the remaining three are derived from accident injury data. The violation-related indices include Significant and Substantial Citations per 100 Inspection Hours (SS/100 IH), Orders per 100 Inspection Hours (O/100 IH), Citations per 100 Inspection Hours (C/100 IH), and Penalty per 100 Inspection Hours (P/100 IH). The injury-related risk indices are as follows: No-Days-Lost Incidence Rate (NDL IR), Nonfatal Days-Lost Incidence Rate (NFDL IR), and Severity Measure (SM). The indices are calculated

Table 1 List of variables used for analysis and their description

Data source	Description	Purpose
Violation dataset		
Citations	Type of citation issued	Citation index calculation
Order	Type of order issued	Order index calculation
Significant and substantial	Gravity of injury	S&S index calculation
Proposed penalty	Penalty issued for a citation	Proposed penalty index calculation
Inspection dataset		
Total inspection hours	Recorded total inspection hours	Citation, order, S&S, and proposed penalty indices calculations
Production dataset		
Annual employee hours	Employee hours for a calendar year	NDL IR, NFDL IR, SM calculations
Accident dataset		
Degree Injury	Degree of injury or illness to an individual	Accident-related indices (NDL IR, NFDL IR, SM) calculations

using the following equations (MSHA, 2023; OSHA 2024, 30 CFR Part 50, 2024):

$$\mathrm{SS}/100\ \mathrm{IH} = \frac{\text{number of significant and substantial citations}}{\text{number of inspection hours}} \times 100 \tag{1}$$

$$\mathrm{O}/100\ \mathrm{IH} = \frac{\text{number of orders}}{\text{number of inspection hours}} \times 100 \tag{2}$$

$$\mathrm{C}/100\ \mathrm{IH} = \frac{\text{number of citations}}{\text{number of inspection hours}} \times 100 \tag{3}$$

$$\mathrm{P}/100\ \mathrm{IH} = \frac{\text{penalty amount}}{\text{number of inspection hours}} \times 100 \tag{4}$$

$$\mathrm{NDLIR} = \frac{\text{number of NDL injuries}}{\text{number of employee hours}} \times 200{,}000 \tag{5}$$

$$\mathrm{NFDLIR} = \frac{\text{number of NFDL injuries}}{\text{number of employee hours}} \times 200,000 \tag{6}$$

$$\mathrm{SM} = \frac{\text{number of restricted and lost work days}}{\text{number of employee hours}} \times 200,000 \tag{7}$$

Equations (1) to (3) measure S&S risk, order risk, and citation risk, and were utilized in previous research [19]. Equation (4) represents the penalty-related risk index, proposed in this paper after identifying its potential to minimize risk in other industries [15, 24]. To remain consistent with MSHA's other calculators (MSHA, 2024), the per 100 inspection hours metric was used in these four equations. Equations (5) and (6) calculate risk indices for NDL and NFDL injuries. Fatalities are not included in any lost time calculations. The only caveat is if an injured person is hospitalized for a period before dying, this is rare and such cases are usually removed from the dataset once death occurs. Equation (7) represents the severity measure, incorporating lost and restricted workdays to ensure higher visibility of the most severe injuries. The number 200,000 in Eqs. (5) to (7) is used to standardize for 100 full-time employees working 40 h per week, 50 weeks per year.

While current risk analysis tools offer valuable insights for maintaining safe mining operations, they often rely on a simplified approach and have limited capacity to incorporate diverse information in assessing potential risks. Accurate management of mining-related risks requires a combination of multiple risk indicators. The IER index is derived by multiplying the unknown weights (w1 to w7) with the respective risk indicators. For simplicity, assigning equal weights (1/7) to each of the seven component risks is a common practice. The determination of weights heavily relies on specific indicators and can involve subjectivity [49]. Equal weights are assigned when indicators hold equal importance (Becker et al., 2020). A weight serves as a coefficient signifying the relative significance of one attribute compared to others [13]. These weights can be computed through iterative unsupervised algorithms, with a weighted ℓ^2 metric established for the indicator (Jimenez-Fernandez et al., 2022).

Various methods exist for weight determination, including the analytic hierarchy process, the entropy method, and the expert evaluation method. Among these, the entropy approach stands out for its objectivity, as it involves no artificial subjective factors. The weight calculation is grounded in data variance, enhancing the precision of quantitative evaluation results [48]. The core principle of the entropy weight methodology lies in computing objective weights based on the divergence levels within the unbiased data [44].

The methodology flowchart, as depicted in Fig. 2, comprises three key elements: the information entropy approach, clustering algorithm, and statistical analysis. The information entropy approach establishes weights by evaluating the dispersion degree for the risk index. The IER's validity is confirmed through clustering models and subsequent statistical analysis for each individual year. The following subsections present the detailed methodology proposed in this research.

3.2 Weight of Indexes

The seven-dimensional risk indices, comprising NDL IR, NFDL IR, SM, Penalty, C/100, SS/100, and O/100, are taken into account for each year to ascertain the weightage values and determine the IER index. Conventionally, index weights are established through subjective fixed weight techniques such as the Delphi method, expert surveys, and the analytic hierarchy process (AHP). However, objective fixed weight methods rely on the inherent information within indexes to generate index weights.

In information theory, entropy serves as a metric for assessing the degree of disorder and relevance within a system's information [25]. The entropy-weighting method provides a more accurate means to bypass the influence of subjective considerations [51]. There exists an inverse mathematical relationship between the probability of an event occurring and the amount of information entropy. Specifically, when an event can be accurately predicted, the probability value is high, and the entropy value is low [2]. Attributes with relatively higher entropy measures exhibit a more extensive data distribution between the two extremes of the solution space. The more random the input, the greater impact the associated attribute has on the algorithm's decision-making [35]. Entropy values serve to mitigate the influence of unusual attributes, thereby enhancing assessment precision. The fundamental discrepancies in decision-making responses can be addressed using the entropy technique [23].

Consequently, the entropy approach is employed to analyze the weight vectors of the indicators in a comprehensive evaluation based on the degree of dispersion [16]. The information entropy procedure is outlined as follows:

Step 1: The seven-dimensional database constructs the evaluation matrix for each year. The number of rows in this database is the number of operating mines in that year.

Step 2: If n sets of indexes exist in the index system and m mines, then x_{ij} is the value of the j^{th} index in the i^{th} mine. Standardizing indexes is important to remove the impact of index dimension on incommensurability. The assessment matrix in the current study is normalized using the critical value approach. The normalized indicators are determined by:

$$x'_{ij} = \varepsilon + \left(\frac{\left(x_{ij} - min\ x_j\right) * (1 - \varepsilon)}{max\ x_j - \min\ x_j} \right) \tag{8}$$

where x'_{ij} is the normalized value for the element x_{ij}, $min\, x_j$ is the minimum value of column j, $max\ x_j$ is the maximum value of column j, the index i represents the mine, which varies from 1 to m, the j represents the individual risk index, which varies from 1 to 7, m is the number of operating mines in a specific year, and ε represents a very small positive number to avoid the minimum normalized value being zero, whose logarithm is undefined. Please note that m might vary from 1 year to the next due to the closure of an existing mine or the opening of a new mine.

Step 3: The range of entropy value e_{j} is [0, 1]. The greater the value of e_{j}, the higher the degree of differentiation and the more information can be extracted [51]. According to the definition of entropy, the entropy of the jth index is determined by:

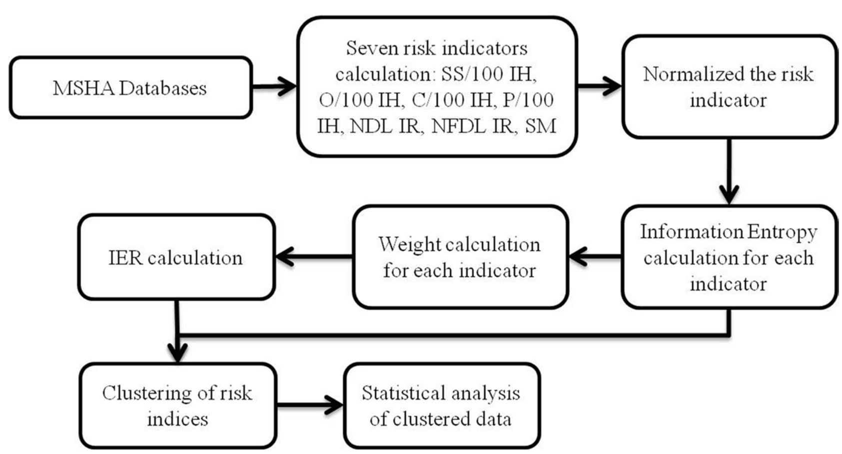

Fig. 2 Flow chart to determine the proposed entropy-based risk index in the current study

$$e_j = \frac{-\sum_{i=1}^{m} \frac{x'_{ij}}{\sum_{i=1}^{m} x'_{ij}} * \ln\left(\frac{x'_{ij}}{\sum_{i=1}^{m} x'_{ij}}\right)}{\ln\ m} \tag{9}$$

Step 4: W_j is defined as the entropy weight of the j^{th} parameter, which is calculated as:

$$w_j = \frac{1 - e_j}{\sum_j (1 - e_j)} \tag{10}$$

Step 5: Finally, the IER index was computed for mine i by multiplying the weights with their corresponding values using the following equation:

$$IER_i = \sum_j w_j * x'_{ij} \tag{11}$$

where $\sum_j w_j = 1$

The weight calculation is instrumental in establishing the relative importance of each indicator. To ensure the IER's unbiasedness, the sum of indicator weights was maintained at 1 [10]. This methodology was consistently applied annually, resulting in the determination of the risk index for each year.

3.3 Validation

The validity of the proposed IER index is confirmed through the clustering of seven-dimensional risk matrices for each respective year. If the resulting risk index effectively encapsulates these matrices, the risk index values between clusters should exhibit statistical differences [6]. This was assessed using the MANOVA test and post hoc techniques. Furthermore, the ANOVA and post hoc tests were employed to verify that the mean values of the risk index across different clusters were statistically different.

3.3.1 Clustering of Risk Indices

Clustering plays a crucial role in various data analysis and machine learning tasks by grouping similar data points together, thus enabling easier data interpretation and pattern recognition [37]. There are many clustering algorithms available for data grouping. The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) stands out for its ability to perform intelligent cluster assignments without the need for human intervention [38]. The number of optimal clusters was determined using the silhouette coefficient, and subsequently, BIRCH clustering was employed to assign cluster labels.

3.3.2 Determining the Number of Clusters

The IER index is validated using statistical analysis of the clustered data. The clustering algorithms learn from the data and, upon encountering new data, categorize it based on previously acquired features [37]. Consequently, clustering algorithms can identify similar relationships among data groups (Lieber, 2013). However, most clustering methods presuppose a known number of clusters for evaluation. To ascertain the appropriate number of clusters, the silhouette method was employed (Saputra, 2020). This method takes into account both intra-cluster and inter-cluster distances when determining the cluster count [8]. The silhouette value gauges a data point's resemblance to its cluster centroid relative to other cluster centroids, with values ranging from -1 to $+1$. A higher Silhouette score signifies a more fitting correspondence of the data point to its cluster centroid [50]. The mean intra-cluster distance (a) and the nearest-cluster distance (b) for each data point are used in the calculation of the silhouette coefficient [50]:

$$\text{Silhouette Coefficient} = \frac{b - a}{\max(a, b)} \tag{12}$$

If the mean silhouette value for a given number of clusters is significantly higher compared to other cluster numbers, it indicates that this number of clusters is optimal [50].

3.3.3 BIRCH Clustering

BIRCH with hierarchies is an efficient clustering algorithm that constructs clusters using the Cluster Feature Tree (CF Tree) and Height-Balanced Tree (Lorbeer et al., 2018). In this hierarchical clustering approach, clustering features are organized in a Height-Balanced Tree (CF Tree), characterized by a balancing factor (B) and a threshold (T) [47]. The process of analyzing and assigning data to the appropriate cluster occurs incrementally as part of building the CF Tree.

The balanced tree is formed using the BIRCH algorithm, which clusters and stores data points in each leaf node. Each leaf node maintains pointers to the node directly below and above it. Following the BIRCH process, these pointers provide swift access to sets as provided by the BIRCH algorithm during clustering (Kovacs and Bednarik, 2011). In this approach, intra-cluster distances are computed using clustering features until the desired number of clusters is achieved. At this point, the algorithm merges the two closest clusters. Therefore, the BIRCH clustering method employs multilevel clustering to reduce complexity and enhance flexibility. It emphasizes identifying the optimal subclusters and multidimensional group metrics to yield clusters of superior quality [31].

3.3.4 Statistical Methods

The multivariate analysis of variance (MANOVA), an extension of the univariate analysis of variance (ANOVA), is employed to assess differences among groups [4]. MANOVA relies on three key assumptions: independence, multivariate normality, and equality of variance matrices [4]. It conducts a comprehensive analysis of the correlations existing among several dependent variables. The null hypothesis (H_0) states that the means of the investigated parameters are equivalent, while the alternative hypothesis suggests that at least one parameter possesses a distinct mean value across the compared populations [39].

Various MANOVA statistics, including Wilks' lambda, Hotelling-Lawley trace, Pillai's trace, and Roy's largest root, aid in determining the *p*-value and evaluating the null hypothesis (H_0) [3]. Additionally, Tukey's honestly significant difference (HSD) test was employed as a post hoc analysis to compare means among different cluster labels. If the *p*-values were below 0.05, it indicated a statistically significant difference between the groups [1].

Following the assessment of cluster robustness, the validity of the developed risk index (IER) was examined by comparing the means of the entropy-based risk index values across different clusters. If the proposed weight calculation method effectively captured the seven-dimensional risk indices, the mean value of the proposed risk index within the cluster should significantly differ from the mean values of the risk index in other clusters, statistically. The ANOVA test was conducted to confirm the mean distinctions of the IER across various clusters. Subsequently, the post hoc test was performed to verify that all means of the risk index from different clusters were statistically different. Box plots were also utilized to visually represent the risk index values across clusters.

4 Results

4.1 Risk Index Determination

Table 2 shows the descriptive statistics of all seven indices from underground mines in this study for 2012. In 2012, a total of 714 active underground mines were identified from MSHA databases. The results from Table 2 indicate that all indices exhibit extremely skewed distributions (skewed to the left, with the median lower than the mean). These findings are further supported by the box plots of these indices (Fig. 3). Table 2 also demonstrates that the different indices have varying ranges, necessitating normalization before calculating the IER index. Table 2 also shows that the minimum value for all seven indices is zero, which could impact the IER index calculation since it involves logarithm operations. The coefficient of variation (CV), which is a standardized measure of dispersion of a frequency distribution, was calculated for all seven indices and is presented in Table 2. The results show that SM has the highest dispersion, followed by NFDL IR and NDL IR, whereas C/100 IH and SS/100 IH show the lowest dispersion.

Table 3 shows the descriptive statistics of the normalized values of the indices. The normalization was performed with an ε value of 1E − 06. The results in Table 3 indicate that all data are normalized within the range of 1E − 06 to 1, while preserving the standardized dispersions (the CV is unchanged). The box plots of the normalized data also demonstrate that the frequency distributions of these indices remain unchanged (Fig. 4). The minimum value (1E − 06) of these indices represents the lowest risk, and the maximum value (1) represents the highest possible risk.

It should be noted that the highest possible risk value (i.e., 1) is very unlikely, as it would require all seven indices to have values equal to 1.

The number of high risk and extremely high-risk mines based on these indices can be identified using the outlier value (OV) threshold (Q_3 + 1.5*IQR) and the extreme outlier value (EOV) threshold (Q_3 + 3*IQR), respectively, as

Table 2 Descriptive statistics of seven risk indices from underground mines in this study for 2012

	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH
Count	714	714	714	714	714	714	714
Mean	7398.50	3.69	3.62	296.69	9.18	2.28	0.29
Std	11,781.79	12.20	14.51	2308.11	8.67	2.09	0.71
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25% (Q_1)	1304.05	0.00	0.00	0.00	5.71	0.96	0.00
50% (Q_2)	3344.07	1.05	0.00	23.27	8.53	1.95	0.04
75% (Q_3)	8645.09	3.57	3.99	201.09	11.44	3.12	0.28
Max	102,356.17	238.66	238.66	57,600.00	200.00	26.37	8.99
CV	1.59	3.31	4.01	7.78	0.95	0.92	2.48

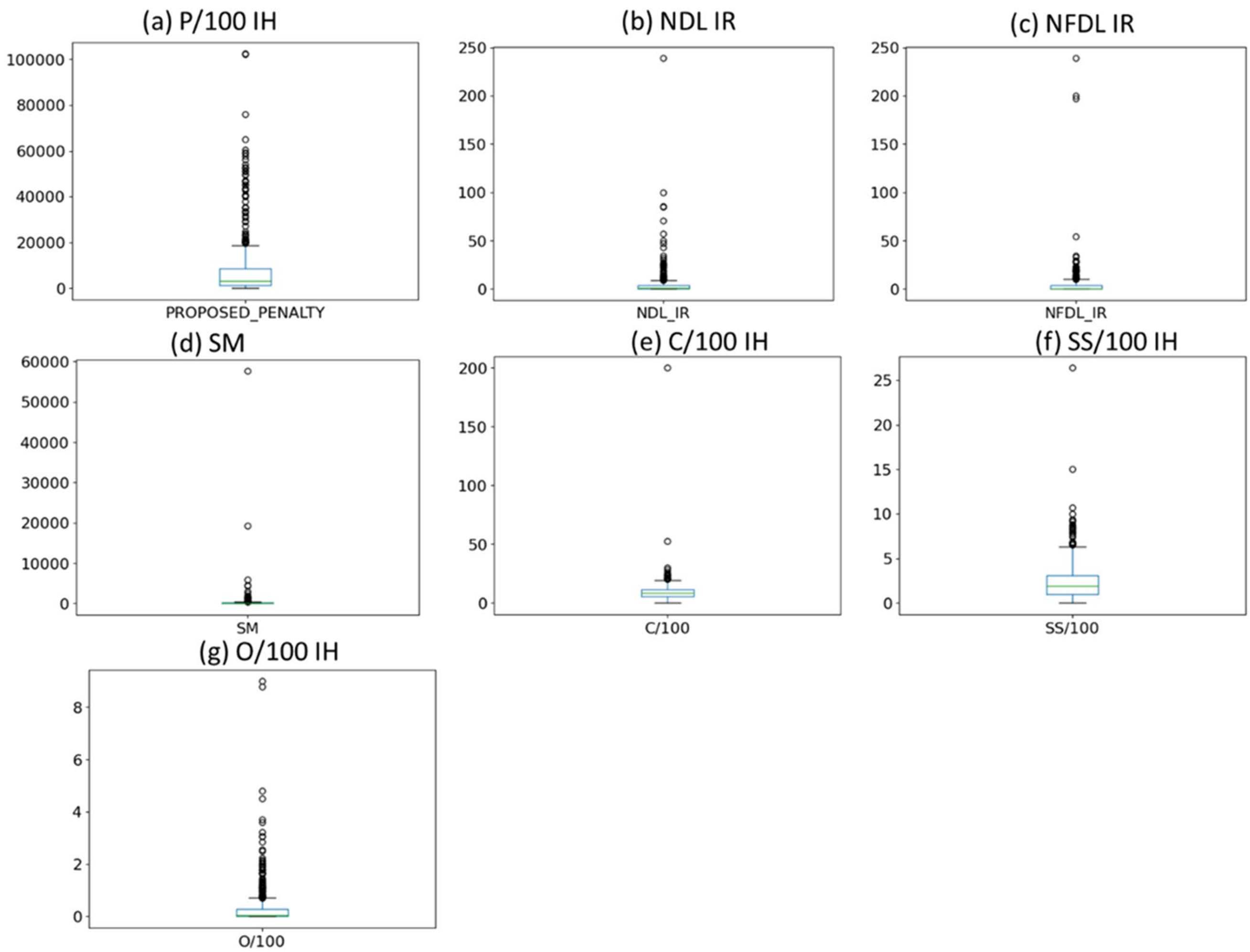

Fig. 3 Box plots of all seven indices for underground mines in 2012

discussed in Kinilakodi et al. [21], where IQR is the interquartile range (Q_3-Q_1). Table 4 shows the extreme outlier value and outlier value thresholds and the number of mines that fall above these threshold values for each index. It was observed that the EOV and the number of mines above EOV thresholds vary significantly from one index to another. The SS/100 IH and P/100 IH have the highest EOV threshold values of 0.36 and 0.29, respectively, whereas NDL IR and NFDL IR have the lowest EOV threshold of 0.06 and 0.067, respectively. The number of mines above the EOV threshold is highest for SM and O/100 IH (45 for both), whereas C/100 IH and SS/100 IH show the lowest number of extreme

Table 3 Descriptive statistics of seven risk indices' normalized data from underground mines in this study for 2012

	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH
Count	714	714	714	714	714	714	714
Mean	0.07	0.02	0.02	0.01	0.05	0.09	0.03
Std	0.12	0.05	0.06	0.04	0.04	0.08	0.08
Min	1E–06	1E–06	1E–06	1E–06	1E–06	1E–06	1E–06
0.25 (Q_1)	0.01	1E–06	1E–06	1E–06	0.03	0.04	1E–06
0.50 (Q_2)	0.03	4.41E–03	1E–06	4.05E–04	0.04	0.07	4.93E–03
0.75 (Q_3)	0.08	0.01	0.02	3.49E–03	0.06	0.12	0.03
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CV	1.59	3.31	4.01	7.78	0.95	0.92	2.48

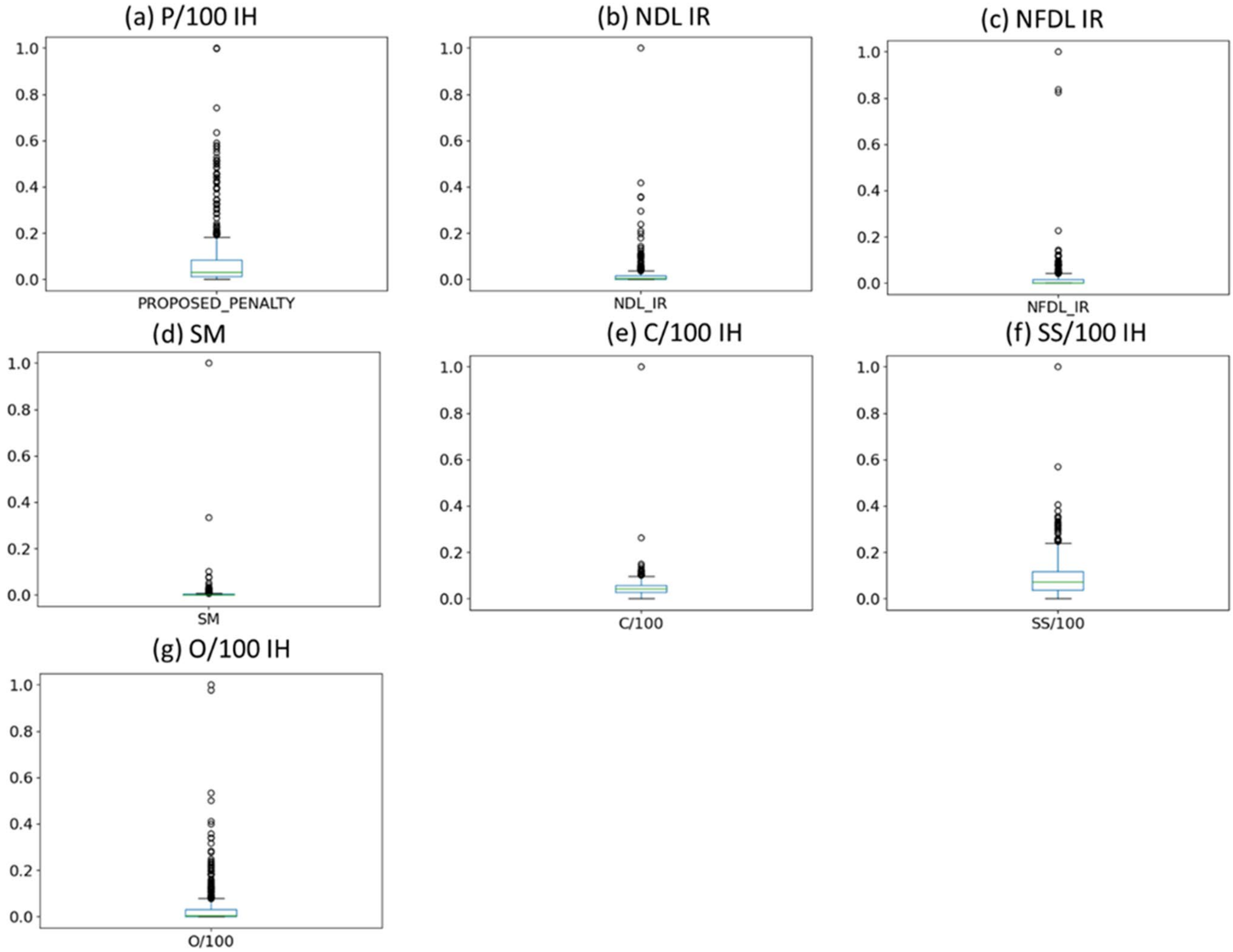

Fig. 4 Box plots of the normalized data all seven indices for underground mines in 2012

high-risk mines, with 3 and 4, respectively. A similar pattern was observed for OV thresholds across all indices. As expected, the number of mines above the OV threshold values is higher than those above the EOV thresholds.

Therefore, the risk quantification of a mine can vary significantly depending on the selected risk index. This result underscores the need for a single risk index that combines all these individual indices. While the risk calculation demonstrated here is based on the EOV, users can choose their own risk ranking based on these normalized values to monitor their mining risk. However, risk categorizations or rankings are beyond the scope of this paper.

The IER index was then calculated using the normalized data, with the weight values associated with each defined risk index for IER calculations presented in Table 5. The results show that SM has the highest weight, followed by NFDL IR and NDL IR. This outcome is consistent with the CV values of the normalized data (Table 3). Additionally,

Table 4 Extreme outlier values and outlier values for all indices and number of mines above the threshold value from underground mines in 2012

	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH	IER
EOV	0.29	0.06	0.067	0.014	0.14	0.36	0.12	0.093
Number of mines above EOV	37	28	25	45	3	4	45	20
OV	0.19	0.037	0.042	0.009	0.1	0.24	0.078	0.061
Number of mines above OV	59	52	44	79	20	26	78	51

Table 5 Weight value for each risk index to calculate IER using different ε for 2012 underground mining

	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH
$\varepsilon = 0.0001$	0.094	0.180	0.202	0.284	0.026	0.046	0.168
$\varepsilon = 1E-06$	0.092	0.180	0.202	0.290	0.025	0.045	0.166
$\varepsilon = 1E-09$	0.092	0.180	0.202	0.290	0.025	0.045	0.166

Table 6 Descriptive statistics of IER and EWR indices from underground mines in this study for 2012

	IER	EWR
Count	714	714
Mean	0.02	0.04
Std	0.03	0.04
Min	1.08E – 04	4.26E – 04
0.25 (Q_1)	0.01	0.02
0.50 (Q_2)	0.02	0.03
0.75 (Q_3)	0.03	0.05
max	0.47	0.33
CV	1.40	0.96

the results indicate that C/100 IH and SS/100 IH have the lowest weight values, which is also reflected in their CV values. The results demonstrate that the weight calculation using the information entropy method can accurately capture the variability of the individual risk indices used for the IER index calculation. Table 5 also highlights that the sum of the weights equals 1, ensuring the unbiasedness of the IER index. The sensitivity of the ε value on the weight calculation was also verified, and Table 5 shows the weight values for different ε values. The table indicates that there is no significant difference in weights when smaller values of ε (1E – 06 or 1E – 09) are used. However, when the ε value is slightly larger, some differences in weights are observed, though they are not very significant. Since ε is used to replace zero, we recommend using an ε value as small as 1E – 06 to ensure it does not impact the solutions.

Table 6 presents the descriptive statistics of IER and equal weights risk (EWR) index values for underground mines in 2012. For the EWR calculation, all seven indices were equally weighted (i.e., 1/7) to obtain the EWR index. The results show that IER values range from as low as 1.08E – 04 to as high as 0.47. Notably, although IER can theoretically reach 1, it only reaches a maximum value of 0.47. The descriptive statistics reveal a very skewed distribution with a long tail at higher index values. The statistics also demonstrate that the IER index captures the distribution shape of all seven indices, which is visualized in the box plot of IER (Fig. 5). EWR index statistics are also presented in Table 6. The results show that the length of the long tail towards higher values is relatively small compared to IER. This is evident from the CV value and the box plot (Fig. 5). Since EWR uses equal weights, it fails to capture the variability of the original data, resulting in higher errors (DelSole, 2013). Although it is challenging to validate this claim, researchers have demonstrated it through experiments (DelSole, 2013). Moreover, by visually comparing the box plot and examining the CV values, it can be concluded that the IER better represents the shape of the frequency distribution and preserves the data statistics better than the EWR index.

Table 7 delineates the weight values determined for IER index using the information entropy approach, evaluating comprehensively through the degree of dispersion. The weights for all indicators exhibit variation over the years, some showing more pronounced values than others. Notably, *significant and substantial* violations (SS/100), which have a direct bearing on injury occurrence, are attributed greater weight than citations. While citations serve as an effective means for enforcing safety regulations and encouraging compliance, they may not wholly encapsulate the presence of risk to miners. The weight values of SM, NFDL IR, and NDL IR are consistently higher compared

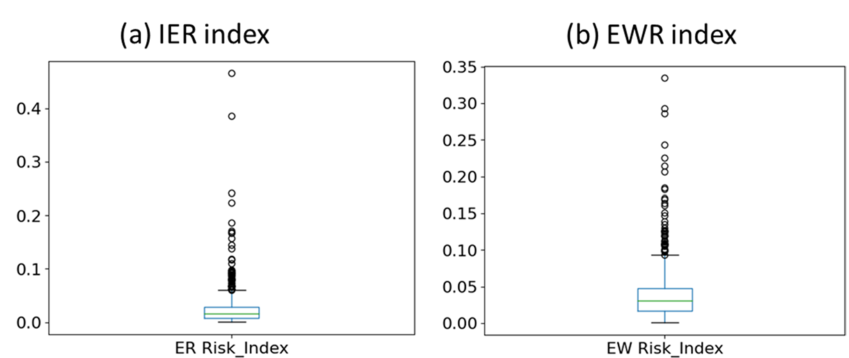

Fig. 5 Box plot of **a** IER and **b** EWR indices for underground mines in 2012

Table 7 Weights for IER index, underground mine number, EOV and OV thresholds for each calendar year from 2011 to 2020 (the number within parentheses in EOV and OV columns represent the number of mines above the threshold value)

CAL YR	Number of samples	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH	EOV threshold	OV threshold
2011	732	0.044	0.291	0.289	0.281	0.008	0.019	0.068	0.023 (28)	0.016 (56)
2012	714	0.092	0.180	0.202	0.290	0.025	0.045	0.166	0.093 (20)	0.061 (51)
2013	633	0.047	0.141	0.342	0.329	0.011	0.026	0.104	0.032 (15)	0.021 (31)
2014	593	0.106	0.199	0.180	0.229	0.023	0.061	0.202	0.173 (13)	0.114 (29)
2015	543	0.088	0.251	0.193	0.210	0.021	0.052	0.185	0.104 (12)	0.069 (24)
2016	470	0.087	0.287	0.148	0.216	0.029	0.059	0.173	0.195 (5)	0.127 (19)
2017	464	0.089	0.184	0.190	0.180	0.025	0.066	0.266	0.111 (8)	0.073 (20)
2018	464	0.110	0.223	0.170	0.200	0.031	0.070	0.195	0.281 (3)	0.182 (20)
2019	451	0.064	0.491	0.119	0.125	0.019	0.045	0.137	0.128 (3)	0.085 (27)
2020	431	0.105	0.197	0.173	0.179	0.041	0.090	0.215	0.322 (2)	0.208 (16)

Table 8 Illustrative summary of mines along with their computed risk indices for the year 2012

MINE ID	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH	IER
1	0.015	0.000	0.838	1.000	0.042	0.105	0.000	0.466
2	0.007	1.000	1.000	0.008	0.038	0.000	0.000	0.386
3	0.512	0.000	0.227	0.336	0.052	0.202	0.248	0.242
4	0.099	0.000	0.000	0.000	0.264	1.000	0.978	0.223
5	0.199	0.000	0.000	0.000	0.039	0.043	1.000	0.187

to other indices. The weight for penalty is consistently over the years. In 2012, the most pivotal criterion is SM, accounting for approximately 29% of the weight. The weights for SM remain relatively consistent, hovering around 0.2. The EOV and OV thresholds for the IER index are also presented in Table 7, showing significant variation from year to year. The number of very high-risk mines (above the EOV threshold) and high-risk mines (above the OV threshold) is also detailed in Table 7. It was observed that the number of very high-risk and high-risk mines has been decreasing over the years. However, it was also noted that the number of operating underground mines has been trending downward, from 732 in 2011 to 431 in 2020.

Table 9 Top 25 high-risk underground mining types from 2011 to 2020

CAL YR	Coal mines	Metal mines	Non-metal mines	Stone
2011	20	2	1	2
2012	17	5	1	2
2013	17	6	1	1
2014	19	5	0	1
2015	11	7	1	6
2016	17	3	0	5
2017	15	6	1	3
2018	17	2	1	5
2019	14	5	0	6
2020	15	4	1	5

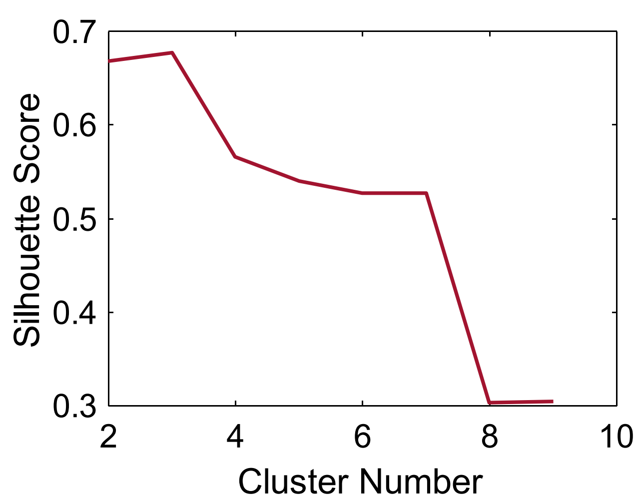

Fig. 6 Silhouette score for different cluster number using BIRCH clustering. The X-axis value represents the Silhouette coefficient. Higher the score, better the cluster separability

Table 10 Silhouette analysis from 2011 to 2020

Calendar year	Number of clusters
2011	2
2012	3
2013	2
2014	2
2015	4
2016	2
2017	2
2018	3
2019	3
2020	3

The IER index is computed by multiplying the assigned weights with their corresponding values for each individual year. Table 8 provides an illustrative selection of mines along with their respective risk indices for the year 2012. It is evident that mine 1 exhibits a considerably higher risk compared to the other mines in the sample table. It should be noted that all five mines are classified as very high-risk, with IER index values exceeding the EOV threshold of 0.093. From this list, it can be observed that these mines are either very risky based on one or two indices (mine 1 has the highest SM score, mine 4 has the highest SS/100 IH score, mine 5 has the highest O/100 IH score, and mine 2 has the highest scores for both NDL IR and NFDL) or contribute moderate risk levels across multiple indices (mine 3 has scores above 0.2 for 6 out of 7 indices).

Table 9 displays the top 25 mining types associated with the highest risks for each year. It is evident that a significant proportion of underground coal mines exhibit higher risks compared to other mining types. The summary indicates that on average, 16 out of the top 25 mines with higher risks are coal mines.

4.2 Validation

This study also employs clustering techniques to categorize the data into *k* clusters. Silhouette scores are used to determine the optimum cluster number, as displayed in Fig. 6 for the year 2012. The Silhouette score is a useful metric for evaluating the quality of clustering and can help choose the optimal number of clusters. It measures how similar a risk index is to its own cluster compared to other clusters. A higher Silhouette score indicates better cluster separability, representing the optimal number of clusters. Figure 6 shows that the Silhouette score increases from cluster number 2 to 3, then gradually decreases as the cluster number increases. Therefore, the optimal cluster number for the 2012 data is 3, with a Silhouette score of 0.677. This process is repeated for each year, and the resulting cluster numbers are presented in Table 10.

The multivariate statistical analysis of the BIRCH clustering algorithm employs metrics including Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root to discern significant distinctions among clustered groups. For the 2012 dataset encompassing 714 mines, an optimal cluster count of three was determined through silhouette analysis. Seven statistical variables (NDL IR, NFDL IR, SM, C/100, SS/100, O/100, and P/100) were utilized for clustering. The degrees of freedom for residuals and the number of independent variables stand at 7.0 and 706.0, respectively. The clustering technique yields notably low p-values (<0.05), confirmed by Pillai's trace which confirms a substantial difference between the groups, thereby rejecting the null hypothesis. Furthermore, Wilks' lambda manifests an F-statistic of 309.31, resulting in a minuscule p-value (0) worthy of note. This sizable F-statistic underscores a significant difference among the groups. Pillai's trace indicates that approximately 75.4% of the variance in the dependent variable is captured. Hence, the statistics affirm a noteworthy distinction between the clusters based on the BIRCH cluster labels.

Given the identified significant difference among the clusters, a post hoc Tukey's HSD test was conducted to further scrutinize the distinctions between pairs of clusters. As depicted in Table 11, noteworthy variances in means among the cluster groups are evident, with the upper and lower bounds instrumental in ascertaining the confidence interval, zeroing in on the mean disparity. The "reject" column signifies that the null hypothesis can be confidently rejected for all comparisons, underscoring a marked distinction between the groups. Analogous outcomes were observed in subsequent years.

Cluster validation, encompassing MANOVA and post hoc analyses, has indeed demonstrated a notable distinction between the clustered groups. To further delineate the differences in IER between clusters, ANOVA and post hoc tests were executed. The ANOVA test encompasses two facets: the variance between clusters and the variance within clusters. Here, the degrees of freedom and sum of squares for inter-cluster variance are denoted as 2.0 and 0.44, respectively. The

Table 11 MANOVA post hoc statistics for the BIRCH algorithm

Group 1	Group 2	Mean diff	P-adj	Lower	Upper	Reject
0	1	0.085	0	0.074	0.095	True
0	2	0.200	0	0.163	0.238	True
1	2	0.115	0	0.077	0.154	True

Table 12 Post hoc statistics for ERI from different clusters

Group 1	Group 2	Mean diff	*P*-adj	Lower	Upper	Reject
0	1	0.059	0	0.051	0.068	True
0	2	0.320	0	0.288	0.352	True
1	2	0.261	0	0.228	0.294	True

residual component encapsulates the variation not accounted for by the clustering algorithm. Notably, the *p*-value associated with inter-cluster differences is exceedingly low, nearly approaching zero (4.6E − 116), firmly rejecting the null hypothesis. Consequently, a significant distinction in IER across cluster groups is evident.

Post hoc statistics were conducted following ANOVA to ascertain the significant mean differences in IER between the cluster groups (Table 12). The adjusted *p*-value (*p*-adj) is notably low, unequivocally rejecting the null hypothesis and affirming a substantial distinction between the cluster groups. These outcomes underscore the efficacy of the proposed IER index in accurately encapsulating the information embedded within the seven-dimensional risk indicators. Comparable findings were also noted in other years.

To visually depict the range of risk index for each cluster group in 2012, a box plot is presented in Fig. 7. This graphical representation allows for the identification of maximum and minimum values within each cluster group. It is evident from the box plot that, in comparison to the other clusters, cluster group 0 exhibits the widest range of risk index values.

5 Case Study

In this section, the IER value of an underground coal mine is calculated. We have selected the Upper Big Branch Mine for this case study application. The Upper Big Branch Mine disaster was a catastrophic mining accident that occurred on April 5, 2010, in West Virginia. The explosion was caused by a buildup of methane gas that ignited, triggering a secondary explosion of coal dust. The mine had been cited numerous times for violations related to its ventilation systems. The accident resulted in 29 fatalities, and several other miners were injured. The 2010 accident at Upper Big Branch Mine underscored the limitations of the POV system. Consequently, MSHA identified and subjected a list of high-risk underground coal mines to intensified inspections in response to the incident.

To showcase the effectiveness of the EIR, a comparative analysis was conducted between the EWR index and the proposed IER index for the Upper Big Branch coal mine leading up to the time of the accident. Table 13 displays the weights computed from 2006 to 2011 utilizing the entropy-based approach. The weights from 2006 to 2011 show very similar trend as we observed before with SM, NFDL IR, and NDL IR are the most contributing factors, and penalty and citations are the least contributing factors.

Table 14 presents the IER index values alongside corresponding EWR index values for the study mine during that period, along with their EOV and OV threshold values. Both the IER and EWR indices show an increasing trend of risk index values for the Upper Big Branch mine from 2006 to 2011, indicating an escalating level of risk for this mine. This trend is clear evidence that there were safety issues in the mining operation as far back as 2006. Although the risk index values for both IER and EWR from 2006 to 2008 are below the EOV and OV limits, the increasing trend over these years indicates emerging safety concerns (Fig. 8).

In 2009, the year before the disaster, the IER index value crossed the OV threshold, classifying the Upper Big Branch mine as a high-risk mine. Unfortunately, the EWR index failed to capture this risk. In 2011, just after the disaster, both the IER and EWR index values showed very high scores, crossing the EOV threshold and categorizing this mine as a very high-risk mine. It was also observed that while the IER index value dropped in 2009 compared to the previous year, the EOV and OV values also dropped significantly from the previous year. This result reveals that solely looking at the index values does not provide a comprehensive insight into the risk unless they are compared with a reference, in this case, the EOV and OV thresholds.

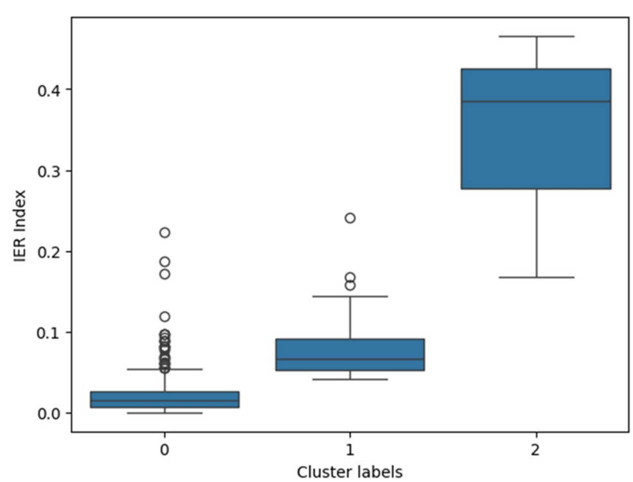

Fig. 7 Box plot for IER index based on cluster labels in 2012 underground mines

Table 13 Entropy-based weights from 2006 to 2011

CAL YR	Number of samples	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH
2006	779	0.054	0.263	0.297	0.279	0.010	0.023	0.076
2007	743	0.101	0.181	0.367	0.157	0.020	0.045	0.139
2008	800	0.104	0.274	0.166	0.223	0.020	0.046	0.167
2009	757	0.038	0.276	0.280	0.308	0.008	0.018	0.073
2010	707	0.095	0.246	0.180	0.231	0.019	0.049	0.180
2011	732	0.044	0.291	0.289	0.281	0.008	0.019	0.068

6 Conclusions and Future Work

The study proposed a new risk index (IER) that combines multiple indices. The proposed index was calculated using a linear weighted method, where the weights are obtained through the information entropy approach. The IER was validated by establishing and validating a risk index using clustering algorithms and rigorous statistical methodologies. To understand the risk levels of mines, two threshold values, EOV and OV, were utilized. The IER index was compared with the EWR index to demonstrate the effectiveness of the proposed risk index. Finally, the risk index for the Upper Big Branch mine was calculated from 2006 to 2011 and compared with the EOV and OV thresholds to assess the mine's risk level. The following conclusions can be drawn from the proposed research:

(a) All individual indices used for the IER index calculations show a very skewed frequency distribution with a long right tail. This is a common phenomenon in risk and reliability data where very few samples fall near the extreme right of the distribution.

(b) The IER preserves the shape and standardized dispersion of frequency distributions of all seven-dimensional indices, making it a suitable tool for evaluating mining risk.

(c) Although, theoretically, the IER should lie between 0 and 1, it is unlikely to see very high values of IER. Mining companies can observe the trend of the IER over the years to understand their risk behavior. While an increasing trend of the IER index can be a concern, companies need to utilize threshold values like EOV and OV to understand where they stand in terms of risk.

(d) Application of IER and EWR for the Upper Big Branch mine shows that the mine had an increasing risk index value from 2006. In 2009, the IER crossed the OV threshold, categorizing it as a high-risk mine. The

Table 14 IER and EWR indices, their EOV and OV threshold values from 2006 to 2011 for the Upper Big Branch mine

	Risk Index		EOV threshold		OV threshold	
	IER	EWR	IER	EWR	IER	EWR
2006	0.012	0.045	0.025	0.144	0.016	0.097
2007	0.030	0.075	0.112	0.206	0.073	0.140
2008	0.034	0.074	0.105	0.254	0.070	0.173
2009	0.020	0.131	0.028	0.248	0.019	0.169
2010	0.072	0.137	0.151	0.366	0.100	0.249
2011	0.071	0.256	0.024	0.159	0.016	0.107

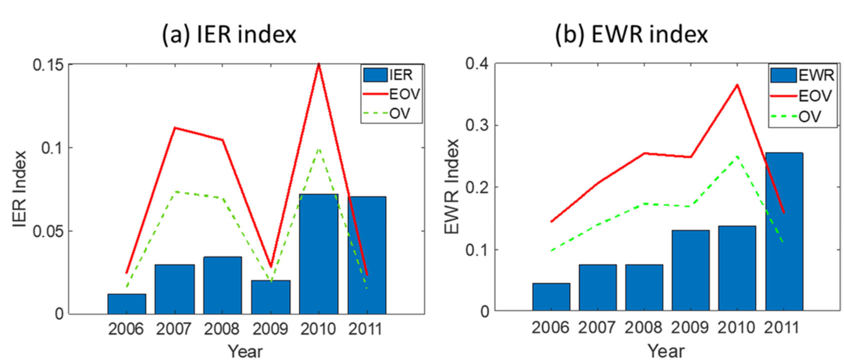

Fig. 8 The risk index: **a** IER and **b** EWR and their associated EOV and OV values from 2006 to 2011 for the Upper Big Branch coal mine

EWR index failed to capture the risk of the mine before the year of the disaster.

(e) The EOV and OV, along with the IER, can be effective tools for identifying high-risk mines, but they cannot categorize mines into different risk levels. A better thresholding approach needs to be developed to categorize mines into different risk levels.

This paper demonstrates how mining companies can utilize the tool to observe their risk levels, but the research has limitations. The indices used for IER calculation do not directly consider fatalities. A suitable index would be required to incorporate fatality information into the risk assessment. Moreover, the IER can only be calculated when all these data are available, so it can only be used for historical risk assessment. An effective risk forecasting tool is needed to evaluate future risks in mining operations. Finally, a better thresholding approach needs to be developed to understand the risk levels of a mine.

Acknowledgements The second author extends his gratitude to the Witte Family Faculty Fellow in Mining Engineering fund at Michigan Technological University for their valuable assistance.

Funding This study received partial support from the National Institute for Occupational Safety and Health (NIOSH), under Grant Number 75D30121C12375.

Code and Data Availability The code and the data that are used in this paper can be downloaded from the Google folder link (https://drive.google.com/drive/u/0/folders/1jslcYxx6_fy10oQJj5Qyu9bNEpDgEFQe). The most up to data MSHA data can be downloaded from this link (https://www.msha.gov/data-and-reports/mine-data-retrieval-system).

References

1. Abdi H, Williams LJ (2010) Tukey's honestly significant difference (HSD) test. Encyclopedia Res Des 3(1):1–5
2. Amiri V, Rezaei M, Sohrabi N (2014) Groundwater quality assessment using entropy weighted water quality index (EWQI) in Lenjanat, Iran. Environ Earth Sci 72:3479–3490
3. Anderson MJ, Walsh DC (2013) PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? Ecol Monogr 83(4):557–574
4. Appolus EE, Okoli CN (2022) A robust comparison powers of four multivariate analysis of variance tests. Euro J Stat Probability 10(1):11–20
5. Becker W, Saisana M, Paruolo P, Vandecasteele I (2017) Weights and importance in composite indicators: closing the gap. Ecol Ind 80:12–22
6. Dalmaijer ES, Nord CL, Astle DE (2022) Statistical power for cluster analysis. BMC Bioinformatics 23(1):205
7. DelSole T, Yang X, Tippett MK (2013) Is unequal weighting significantly better than equal weighting for multi-model forecasting? Q J R Meteorol Soc 139(670):176–183
8. Dinh DT, Fujinami T, Huynh VN (2019) Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20 (pp. 1–17). Springer Singapore
9. Duarte J, Marques AT, Santos Baptista J (2021) Occupational accidents related to heavy machinery: a systematic review. Safety 7(1):21
10. Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, USA
11. Grayson RL, Kinilakodi H (2011) A comparison of the 2008?2009 post-MINER Act safety performance of union and non-union underground coal mines. Int J Min Mineral Eng 3(3):173–193
12. Grayson RL, Kinilakodi H, Kecojevic V (2009) Pilot sample risk analysis for underground coal mine fires and explosions using MSHA citation data. Saf Sci 47(10):1371–1378
13. Greco S, Ishizaka A, Tasiou M, Torrisi G (2019) On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. Soc Indic Res 141:61–94
14. Groves WA, Kecojevic VJ, Komljenovic D (2007) Analysis of fatalities and injuries involving mining equipment. J Safety Res 38(4):461–470
15. Haviland A, Burns R, Gray W, Ruder T, Mendeloff J (2012) A new estimate of the impact of OSHA inspections on manufacturing injury rates, 1998–2005. Am J Ind Med 55(11):964–975
16. Ji C, Jiao Z, Yuan S, El-Halwagi MM, Wang Q (2021) Development of novel combustion risk index for flammable liquids based on unsupervised clustering algorithms. J Loss Prev Process Ind 70:104422
17. Jiménez-Fernández E, Sánchez A, Ortega-Pérez M (2022) Dealing with weighting scheme in composite indicators: an unsupervised distance-machine learning proposal for quantitative data. Socio-econ Plann Sci 83:101339
18. Kecojevic V (2011) Analysis of "high-dollar" value safety and health citations and orders for the US coal mines. Saf Sci 49(5):658–663
19. Kinilakodi H, Grayson RL (2011) A methodology for assessing underground coal mines for high safety-related risk. Saf Sci 49(6):906–911
20. Kinilakodi H, Grayson RL (2011) Citation-related reliability analysis for a pilot sample of underground coal mines. Accid Anal Prev 43(3):1015–1021
21. Kinilakodi H, Grayson RL, Oyewole SA (2012) Evaluating equivalence of the safe performance index (SPI) to a traditional risk analysis. Open J Saf Sci Technol 2(2):47–54
22. Komljenovic D, Groves WA, Kecojevic VJ (2008) Injuries in US mining operations–a preliminary risk analysis. Saf Sci 46(5):792–801
23. Kumar R, Singh S, Bilga PS, Singh J, Singh S, Scutaru ML, Pruncu CI (2021) Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: a critical review. J Market Res 10:1471–1492
24. Levine D, Toffel M, Johnson M (2012) Randomized government safety inspections reduce worker injuries with no detectable job loss. Science 336(6083):907–911
25. Li X, Wang K, Liu L, Xin J, Yang H, Gao C (2011) Application of the entropy weight and TOPSIS method in safety evaluation of coal mines. Procedia Eng 26:2085–2091
26. Libório MP, Karagiannis R, Diniz AMA, Ekel PI, Vieira DAG, Ribeiro LC (2024) The use of information entropy and expert opinion in maximizing the discriminating power of composite indicators. Entropy 26(2):143
27. Lieber D, Stolpe M, Konrad B, Deuse J, Morik K (2013) Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. Procedia Cirp 7:193–198
28. Mine Safety and Health Administration (MSHA) (2023) Mining industry accident, injuries, employment, and production data,

https://arlweb.msha.gov/STATS/STATSCR1.HTM (accessed 12.03.2023)

29. Mine Safety and Health Administration (MSHA) (2024) Data sources and calculators. https://www.msha.gov/data-and-reports/data-sources-and-calculators (accessed 05.03.2024)
30. National Institute of Occupational Safety and Health (NIOSH) (2024) Number and rate of occupational mining fatalities by year, 1983 – 2022. https://wwwn.cdc.gov/NIOSH-Mining/MMWC/Fatality/NumberAndRate
31. Nwadiugwu MC (2020) Gene-based clustering algorithms: comparison between Denclue, Fuzzy-C, and BIRCH. Bioinform Biol Insights 14:1177932220909851
32. Occupational Safety and Health Administration (OSHA) (2024) Commonly used statistics, Federal OSHA coverage. https://www.osha.gov/data/commonstats
33. Onder S (2013) Evaluation of occupational injuries with lost days among opencast coal mine workers through logistic regression models. Saf Sci 59:86–92
34. Qu W, Li J, Song W, Li X, Zhao Y, Dong H, ... Qi Y (2022) Entropy-weight-method-based integrated models for short-term intersection traffic flow prediction. Entropy 24(7):849
35. Rastogi M, Chauhan A, Vaish R, Kishan A (2015) Selection and performance assessment of phase change materials for heating, ventilation and air-conditioning applications. Energy Convers Manage 89:260–269
36. Robson LS, Ibrahim S, Hogg-Johnson S, Steenstra IA, Van Eerd D, Amick BC III (2017) Developing leading indicators from OHS management audit data: determining the measurement properties of audit data from the field. J Safety Res 61:93–103
37. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LDF, Rodrigues FA (2019) Clustering algorithms: a comparative approach. PLoS One 14(1):e0210236
38. Roselin AG, Nanda P, Nepal S, He X (2021) Intelligent anomaly detection for large network traffic with optimized deep clustering (ODC) algorithm. IEEE Access 9:47243–47251
39. Rybak A, Rybak A, Joostberens J (2023) The impact of removing coal from Poland's energy mix on selected aspects of the country's energy security. Sustainability 15(4):3457
40. Sammarco JJ, Podlesny A, Rubinstein EN, Demich B (2016) An analysis of roof bolter fatalities and injuries in US mining. Trans Soc Min Metall Exp Inc 340(1):11
41. Sanmiquel L et al. (2018) Analysis of occupational accidents in underground and surface mining in Spain using data-mining techniques. 15(3):462. https://doi.org/10.3390/ijerph15030462
42. Saputra DM, Saputra D, Oswari LD (2020) Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. In Sriwijaya international conference on information technology and its applications (SICONIAN 2019) (pp. 341–346). Atlantis Press
43. Stemn E (2019) Analysis of injuries in the Ghanaian mining industry and priority areas for research. Saf Health Work 10(2):151–165. https://doi.org/10.1016/j.shaw.2018.09.001
44. Tai X, Xiao W, Tang Y (2020) A quantitative assessment of vulnerability using social-economic-natural compound ecosystem framework in coal mining cities. J Clean Prod 258:120969
45. Title-30, Chapter-I, Subchapter-I, Part-50 (30 CFR Part 50) (2024) Electronic Code of Federal Regulations. https://www.ecfr.gov/current/title-30/chapter-I/subchapter-I/part-50?toc=1 (Accessed on 05.01.2024)
46. Tubis A, Werbińska-Wojciechowska S, Wroblewski A (2020) Risk assessment methods in mining industry—a systematic review. Appl Sci 10(15):5172
47. Venkatkumar IA, Shardaben SJK (2016) Comparative study of data mining clustering algorithms. In 2016 International Conference on Data Science and Engineering (ICDSE) (pp. 1–7). IEEE
48. Xiao W, Lv X, Zhao Y, Sun H, Li J (2020) Ecological resilience assessment of an arid coal mining area using index of entropy and linear weighted analysis: a case study of Shendong Coalfield, China. Ecol Indicators 109:105843
49. Xu M, Cui Y, Hu M, Xu X, Zhang Z, Liang S, Qu S (2019) Supply chain sustainability risk and assessment. J Clean Prod 225:857–867
50. Zhou HB, Gao JT (2014) Automatic method for determining cluster number based on silhouette coefficient. Adv Mater Res 951:227–230
51. Zhu Y, Tian D, Yan F (2020) Effectiveness of entropy weight method in decision-making. Math Probl Eng 2020:1–5