2023

AN ENTROPY-BASED RISK INDEX (ERI) OF MINING HEALTH AND SAFETY USING CLUSTERING AND STATISTICAL METHODS

Dharmasai Eshwar Reddy Sirigiri
Michigan Technological University, sirigiri@mtu.edu

AN ENTROPY-BASED RISK INDEX (ERI) OF MINING HEALTH AND SAFETY USING CLUSTERING AND STATISTICAL METHODS

By

Dharmasai Eshwar Reddy Sirigiri

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Mining Engineering

MICHIGAN TECHNOLOGICAL UNIVERSITY

2023

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Mining Engineering.

Department of Geological and Mining Engineering and Sciences

Thesis Advisor: *Dr. Snehamoy Chatterjee*

Committee Member: *Dr. Nathan Manser*

Committee Member: *Dr. Aref Majdara*

Department Chair: *Dr. Aleksey Smirnov*

Table of Contents

List of Figures

List of Tables

Author Contribution Statement

The paper's primary author is the author of this thesis, which is included in Chapter 2. Dr. Snehamoy Chatterjee, the study's second author, serves as the first author's advisor for the Master of Science degree in Mining Engineering. The thesis contains three (3) main chapters, including Chapter 1 discussing the Introduction to the thesis topic, Chapter 2 focuses on the entropy-based risk index of mine health and safety using clustering and statistical methods. The summary and conclusions are presented in Chapter 3, which is based on the results of the entropy-based risk index.

Acknowledgments

I sincerely acknowledge my advisor Dr. Snehamoy Chatterjee for his support and guidance, which was immense throughout my Masters' tenure. Under his direction, the project's study was written and successfully finished. The National Institute for Occupational Health and Safety (NIOSH) deserves special recognition for providing me with partial financial support and aiding me in pursuing my MS degree. I am also grateful to Dr. Manser and Dr. Aref Majdara for serving on my committee and supporting me through the completion of my thesis. I am incredibly grateful to my family and friends for assisting me in relocating to this country to continue my further studies. I would also like to express my gratitude to my parents, who helped me develop an ambitious nature, and to my brother, who supported me through life's most challenging circumstances.

List of Abbreviations

US	: United States
NIOSH	: National Institute for Occupational Safety and Health
MSHA	: Mine Safety and Health Administration
POV	: Pattern of Violation
SPI	: Safe Performance Index
S&S	: Significant & Substantial citations
SPM	: Safety Performance Measurement
ML	: Machine Learning
CFR	: Code of Federal Regulations
NLT	: No Lost Time injuries
LT	: Lost Time injuries
NFDL	: Non-Fatal Days-Lost injuries
NDL	: No Days Lost injuries
IR	: Incidence Rate
AHP	: Analytic Hierarchy Process
BIRCH	: Balanced Iterative Reducing and Clustering using Hierarchies
MANOVA	: Multivariate Analysis of Variance
HSD	: Honestly Significant Difference
ANOVA	: Analysis of Variance

IH : Inspection Hours

Abstract

Over the last decade, the mining industry has seen a significant reduction in the number of fatalities in the United States. However, the annual employee hours have also decreased during the same period. Therefore, it is crucial to evaluate the historical mining data and identify the potential risks of a mine through the mine risk index. The risk indicators also describe the severity of accidents and injuries at mine sites. The variables such as citation, order, significant & substantial citations, lost time, and no lost time injuries and penalties are considered for the determination of the risk index. However, using multiple risk indicators to understand the safety standard of a mine could be a complicated process. The mining industry historically uses arithmetic averaging that considers equal weights for each indicator to calculate the mining risk index. This research proposed a new approach to calculate weight values for different risk indicators for calculating the mining risk index. The weights were calculated through the information entropy approach to understand the degree of dispersion. The influence of these variables on the risk analysis was evaluated to comprehend the risk index. The risk index is validated through hierarchical clustering algorithms such as BIRCH clustering. MANOVA and post-hoc tests were performed to validate the clustering performance. The statistical differences amongst the means of risk index from different clusters were tested through box plots and ANOVA test. The investigation was conducted during the 2011-2020 period utilizing the open-source mine safety and health administration (MSHA) databases. Results from the statistical analysis show that the risk indicators significantly differ from one cluster to the others for all periods. The results also show that the top 25 high-risk mines constitute around 64.8% of coal mines for all periods. The risk index results from the clustering and statistical analysis can help the mining industry to determine the risk index, thereby focusing on ensuring workplace safety.

1. Introduction

1.1. Overview

The mining industry continues to pose a significant risk to worker safety and health. In recent years, industry has significantly improved in reducing major injuries and fatalities. However, multiple research studies and statistical data demonstrate that the number of serious injuries and fatalities remains at a higher level (Dragan et al., 2017). An accident occurs due to an unsafe physical or mechanical work environment. (Rahimi et al., 2022). Therefore, it is crucial to address the violations by evaluating the historic mining data and determining the risk index to ensure workplace safety. The mine safety and health administration (MSHA) of United States inspect mines to identify potentially hazards, thereby eliminating them before an accident which can help in reducing the injuries and fatalities (Milam et al., 2020). MSHA of the United States manages the accidents, violations, employment production, and inspections datasets that provide insights about the mine. MSHA issues citations to a mine operator and owner for an alleged violation of a standard or Section of the Act (Brian, 2021). Research on the determination of risk would help the mining companies to provide significant insights into factors related to mining injuries.

The simplest method which has been applied in numerous research is to assign the criterion equal weights. However, evaluation outcomes largely depend on the criteria weights (Ghorabaee et al., 2021). It is important to consider the determination of the weights of the qualities. The method of determining weight using entropy can eliminate the possibility of subjectivity as well as the impact of human factors (Liu et al., 2021). The entropy measure used to determine criteria weights is that the less the entropy measure of a criterion, the more the weight should be assigned to that criterion among alternatives (Biswas and Sarkar, 2019). To summarize, this research focuses on the entropy-based weights determination of the risk index in the mining industry using clustering and statistical methods.

1.2 Outline

The thesis is organized in the below-mentioned manner:

Chapter 1: An overview of the general introduction to the research issue of this thesis is provided in this chapter, along with a description of the steps and techniques. The specifics include the risk analysis methodology in the mining industry.

Chapter 2: This chapter discusses the procedure for the risk index parameters. The reason for considering the proposed risk index over other indicators is also explained in this section of the thesis. Finally, the results of applying statistical techniques and machine learning algorithms are studied.

Chapter 3: This section of the thesis provides insight into the overall findings drawn from the statistical analysis and the clustering models. In addition, the scope of future work is also discussed.

2. AN ENTROPY-BASED RISK INDEX (ERI) OF MINING HEALTH AND SAFETY USING CLUSTERING AND STATISTICAL METHODS

Dharmasai Eshwar[a*], Snehamoy Chatterjee[a], Rennie Kaunda[b], Hugh Miller[b], Aref Majdara[c]

[a]Michigan Technological University, 1400 Townsend Drive, Houghton, MI, 49931 USA.

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

[c]Washington State University , 14204 NE Salmon Creek Ave, Vancouver, WA, 98686.

(This chapter will be submitted in Safety Science Journal – Elsevier Publications)

Abstract

Over the recent decades, the mining industry has significantly reduced accidents and injuries in the United States. While these statistics are positive, these numbers are confusing due to the declining workforce and employee hours throughout this time. The Mine Safety and Health Administration (MSHA) of the United States has implemented a Pattern of Violation (POV) and Significant & Substantial (S&S) calculator to monitor safety in mines; however, both have their limitations. Different risk indices were proposed to overcome these limitations by utilizing multiple matrices from MSHA databases. However, integrating multiple matrices within a single risk index is the key challenge. This research aims to develop an information entropy-based risk index (ERI) by optimizing the weights of the conflicting matrices. The risk indicators used for the ERI calculation are citation, order, significant & substantial citations, penalty, no lost time, and lost time injury. The proposed ERI was tested using MSHA's underground mines data from 2011 to 2020. The proposed risk index was validated by BIRCH clustering algorithm and statistical analysis. The clustering performance was evaluated by multivariate analysis of variance (MANOVA) test and post-hoc analysis. Box plots and ANOVA test validated the statistical mean difference of the ERI between clusters. MANOVA test and post-hoc results show that BIRCH clustering successfully clustered the seven-dimensional risk indices for all periods. The ANOVA test shows mean risk

index values for at least one cluster is statistically different from other clusters at 95% confidence for all periods. The post-hoc analysis also demonstrated a statistical significance between the means of the risk index of different clusters. Box plot results also support those findings. Finally, the proposed approach was applied to an underground coal mine to show its effectiveness. The study supports that the proposed approach can help the mining company to understand its safety performance and take necessary measures to improve it.

2.1 Introduction

Mining is the process of extracting minerals from veins, seams, or ore bodies that are economically valuable and used in several industries, including steel production, power generation, electronics, construction products, and even agriculture (Groves et al., 2007). Despite these significant contributions, mining is termed the most hazardous environment affecting the safety and health of workers through dust inhalation, roof fall, explosions, rockslides, and humidity (Tawiah et al., 2014; Sanmiquel et al., 2018). Mining is associated with injuries and fatalities with high incidence rates compared to other industries (Onder, 2013). The fatal injury rate in the United States (US) for full-time equivalent workers per 100,000 decreased from 23.5% in 2006 to 11.4% in 2015 (Kia et al., 2017). Mine Safety and Health Administration (MSHA) in the US continuously monitors the fatalities and other injuries, accidents, and violations. The statistics show a decline in fatalities from 242 in 1977 to 28 fatalities in 2017. According to a study of US mine accident fatalities from 1983 to 2018, almost 0.017% of miners died in accidents on average (Rahmi et al., 2022). Figure 1 shows the total number of fatal injuries in the US from 2011 through 2020. However, these numbers are misleading due to the declining workforce and employee hours from 2015 – 2020 (Figure 1). Therefore, to decrease mining accidents and injuries, thus fatalities, a qualitative and quantitative risk analysis investigation helps identify potential risk-related incidents (Stemn, 2019; Grayson et al., 2009).

MSHA of the US and the government have implemented various provisions to improve the protection of miners. MSHA shared an interactive portal based on the mine health

and safety standards containing the mine statistical databases, including accidents, inspections, violations, and degree of seriousness (Grayson et al., 2009). In addition, MSHA initiated several monitoring tools based on citation history, which includes the Pattern of Violation (POV) and Safe Performance Index (SPI) to monitor the potential cause of an accident (Kinilakodi and Grayson, 2011b). However, the complex calculation of POV with ten components and the orders and Significant & Substantial (S&S) citations are significantly challenged in the process. Consequently, the SPI is determined by considering more weightage to the more serious citations (Kinilakodi and Grayson, 2011a). The Safety Performance Measurement (SPM) provides information about the remedial action for risk control rather than focusing on safety in the workplace (Arezes and Miguel, 2003). The SPM tend to be reactive because they only focus on a small part of the day-to-day performance and ignore the situations that potentially caused accidents. The S&S calculator tool, proposed by MSHA, determines a substantial likelihood of incurring a serious injury because of an underlying safety or health hazard. This tool determines the number of S&S citations and orders per 100 inspection hours within a specific time window (MSHA). However, it ignores further citations and only considers two factors: the quantity of S&S citations and the number of inspection hours. Therefore, it is critical to evaluate historical mining injury data and identify the potential risks through the mine risk index.

One of the most critical issues the mining industry deals with is investigating the cause of accidents and quantitative analysis of the risk exposure (Onder, 2013). One such study of a citation-based reliability approach using a pilot sample of underground coal mines assesses the risk of violating safety and health standards. However, the main research focus was limited to underground coal mines with major citations (Kinilakodi and Grayson, 2011a). Another study investigated the modified safe performance index through an equal weightage of accident and citation measures to determine the mine safety performance (Kinilakodi and Grayson, 2011b). Other research examined the risk assessment of underground bituminous mines and risk characterization using citation data in Pennsylvania (Orsulak et al., 2010). However, there is also a need to address the

violations of the regulations and mandatory health and safety standards (Kecojevic, 2011).

The existing tools, including POV and S&S calculator help actively monitor the safety practices in the mines. However, they have their limitations in identifying the potential risks associated with any mine. The MSHA safety indicator metrics that offer understanding of safety are difficult to quantify (Harms-Ringdahl, 2009). Therefore, the research objective is to propose a risk index through leading health and safety indicators to help mining companies assess mine safety performance. An entropy-based risk index (ERI) is proposed to analyze the relative importance of different factors, based on their entropy values, and assign weights accordingly. The ERI is analyzed by utilizing the MSHA databases to assess and optimize the weights of the conflicting matrices. The weight determination of risk indicators was done through the information in an event and the degree of dispersion. The ERI was validated by comparing the results with the unsupervised machine learning (ML) algorithm to determine the clusters with relative risk. As a result, it is possible to calculate the risk index over time, which can reveal trends in the evolution of the amount of risk associated with mining operations. Thus, determining the risk index could bring more insight into the mining industry's risk profile.

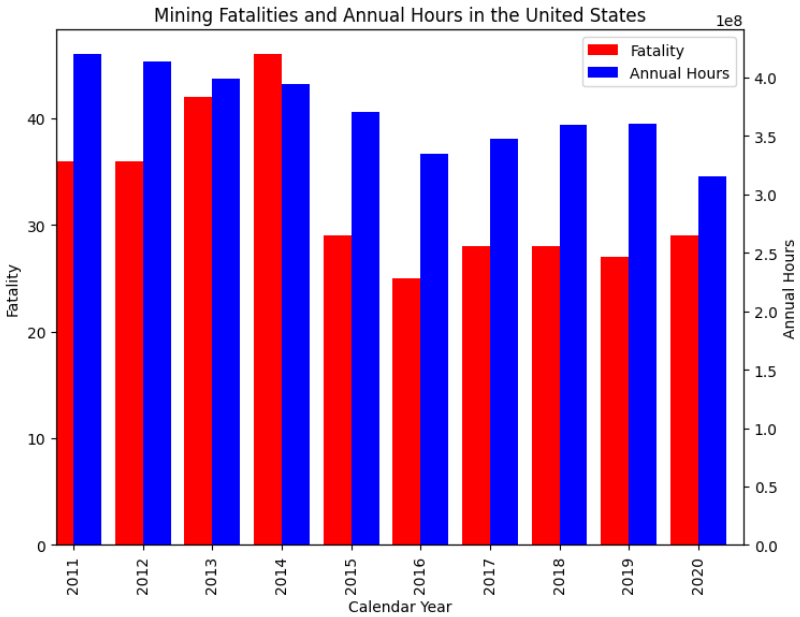


Figure 1 Mining Fatalities and Annual Hours in the United States

2.2 Materials

The MSHA focuses on eliminating fatalities and injuries by implementing mandatory health and safety standards. MSHA databases contain records pertaining to safety and health violations that are updated every week. MSHA databases consist of violations, inspections, employment production and accidents dataset. The violations dataset includes the violations that were reported as a result of MSHA inspections that commenced on 1/1/2000. Each inspection generates a unique event number linked directly to the inspection's dataset. Table 1 includes information on the precise citation, order, or safeguard issued, including the Section of Act, and the occurrences that led up to the violation (MSHA). A citation is issued for violating the mandatory safety and health standards during an inspection. In contrast, an order is issued to the mine operator with penalties when an aggravated activity constitutes greater than ordinary negligence. In addition, a log of failures is maintained for underground subsystems, which can be used in determining the risk hazards and eliminating potential accidents. (Grayson et al., 2009). However, according to MSHA, specific criteria must be proven for a violation to be deemed serious and substantial. Firstly, there needs to be an underlying violation of a regulatory standard. Secondly, a specific safety threat must have contributed to the violation. Lastly, a decent chance is that the risk of hazard will cause an injury. A violation must include these components to be considered significant and substantial (Kinilakodi and Grayson, 2011b). The inspection dataset contains information about the name of the mine that was inspected, the inspection number assigned by MSHA, the inspection hours, and the dates when the inspection started and ended. The employee production dataset contains the annual total of employee hours and coal production reported by mine operators. The accidents dataset contains information on all reported accidents, injuries, and degree of injury by mine operators and contractors (MSHA). The degree of injury from accidents dataset is categorized into ten degrees based on MSHA reports (NIOSH, 2016), including fatal, permanent total or permanent 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. The initial dataset consisted of injuries based on violations, inspections, employee

production, and accident datasets reported to MSHA from 2011 to 2020 thereby calculating the risk index is determined for each year. Based on existing research and analysis, the injured miner's data is examined using risk indicators and their association with the degree of injury (Amoako et al., 2021). This research focuses on MSHA datasets and determining the mine risk index over the years. The complete list of variables considered in this research is provided in Table 1.

Table 1: List of variables used for analysis

Column Name	Description
Violations dataset	
Citations	Type of citation issued
Order	Type of order issued
Safeguard	Type of safeguard issued
Significant and substantial	Gravity of injury
Proposed penalty	Penalty issued for a citation
Inspections dataset	
Total inspection hours	Recorded total inspection hours
Employment Production dataset	
Annual hours	Summation of annual hours
Accidents dataset	Degree of injury or illness to an individual
Degree Injury	

2.3 Methodology

2.3.1 Evaluation of matrix parameters

Incidences are associated with multiple contributing factors, and the metrics provide additional insight into the complex factors associated with these incidents resulting in death or injury. According to Title 30 of the Code of Federal Regulations (CFR), Part 50 mandates that mine operators and contractors submit the MSHA Form 7000-1 with a detailed description of the reportable incident (Sammarco et al., 2016). The accidents in the mines are classified as no lost time injuries (NLT) and lost time injuries (LT) (Robson et al., 2018). Analysis of the lost time data helps prevent fatalities (Grayson et al. 2009).

The LT and NLT injuries were further classified into nonfatal days-lost injuries (NFDL) and fatal and no-days-lost injuries (NDL) (Komljenovic et al., 2008). Therefore, the incidence rate (IR), no-days-lost incidence rates (NDL IR), and nonfatal days-lost incidence rates (NFDL IR) are determined as:

$$\text{IR} = \frac{\text{number of injury occurrences}}{\text{number of employee hours}} \text{ x } 200{,}000 \quad (1)$$

$$\text{NDL IR} = \frac{\text{number of injuries in a category}}{\text{number of employee hours}} \text{ x } 200{,}000 \quad (2)$$

$$\text{NFDL IR} = \frac{\text{number of nonfatal days lost injuries}}{\text{number of employee hours}} \text{ x } 200{,}000 \quad (3)$$

$$\text{SM} = \frac{\text{restricted and lost work days}}{\text{number of employee hours}} \text{ x } 200{,}000 \quad (4)$$

Research shows that the increased penalties for violations can help in improving the health and safety standards in mines (Kecojevic, 2011). The elevated citations for specific violations (such as those categorized as S&S and orders) may indicate failure to manage risks and lead to poor safety performance (Grayson and Kinilakodi, 2011). The frequency of occurrence is the normalized safety measure, and severity is estimated from the relative amount of related loss (Kinilakodi et al., 2012). The leading indicator-type statistics on significant and substantial citations (SS/100 IH), orders (O/100 IH), and citations (C/100 IH) per 100 inspection hours provide a detailed understanding of risk analysis (Grayson et al., 2009). MSHA imposed rules to address safety issues, and penalties for mine safety and health requirements violation that has increased considerably and now constitute a considerable cost to mine operators (Kecojevic, 2011). The penalty for each citation varies considerably, with more severe health and safety violations incurring heavier penalties (Yorio et al., 2014). The penalty per 100 inspection hours (P/100 IH) is an additional indicator to determine the risk index. The SS/100 IH, O/100 IH, P/100 IH and C/100 IH are calculated as:

$$\text{C/100 IH} = \frac{\text{number of citations}}{\text{number of inspection houes}} \text{ x } 100 \quad (5)$$

$$\text{SS/100 IH} = \frac{\text{number of significant and substantial citations}}{\text{number of inspection houes}} \text{ x } 100 \quad (6)$$

$$\text{O/100 IH} = \frac{\text{number of orders}}{\text{number of inspection houes}} \text{ x } 100 \quad (7)$$

$$\text{P/100 IH} = \frac{\text{penalty amount}}{\text{number of inspection houes}} \text{ x } 100 \quad (8)$$

Table 2: Measures used for risk index calculation

Mine ID	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH
1	0.0085	1.0000	0.0000	0.0000	0.1039	0.0000	0.0000
2	0.0797	0.0000	1.0000	1.0000	0.6119	0.5159	0.0847
3	0.1350	0.0000	0.0000	0.0000	0.2189	0.5801	1.0000
4	0.1238	0.0016	0.4025	0.3461	0.3254	0.2861	0.0395
5	0.4356	0.0001	0.0169	0.1088	0.6936	0.7851	0.2206

Table 2 shows the standard injury measures used for calculating the performance index for five example mines. The risk analysis tools have various considerations to determine the risk and ensure safe operations in the mines. However, these techniques are based on a simplistic approach, and the utilization of different information to determine the potential risk is limited. A combination of multiple risk indicators is used to determine and manage the mining-related risks accurately. The risk index is calculated by multiplying the unknown weights (w_1 to w_7) by the individual risk indicators. To ensure simplicity, we can assign equal weight (1/7) to each of the seven component risks. The weights for each risk depend heavily on the indicators and can be quite subjective (Xu et

al., 2019). Equal weights are assigned when the indicators are of equal importance (Becker et al., 2020). A weight is a type of coefficient that indicates the relative importance of one attribute compared to others (Greco et al., 2019). The weights can be calculated using the iterative unsupervised machine learning (ML) algorithms and a weighted ℓ^2 metric determined for the indicator. (Jimenez-Fernandez et al., 2022). There are multiple ways to determine weights, including the analytic hierarchy process, the entropy method, and the expert evaluation method. Of these, the entropy approach is the most objective as it contains no artificial subjective factor, and the weight calculation is based on the variance of data, which can increase the precision of quantitative evaluation results (Xiao et al., 2020). The fundamental concept of the entropy weight methodology is to compute the objective weight of indicators based on the divergence degrees of the unbiased data (Tai et al., 2020). The flow chart for methodology has three key components, as shown in Figure 2: the information entropy approach, machine learning modeling, and statistical analysis. The information entropy approach determines the weights using the degree of dispersion for the risk index. The ERI is validated through unsupervised machine learning models and statistical analysis for an individual year. The following subsections present the detailed methodology proposed in this research.

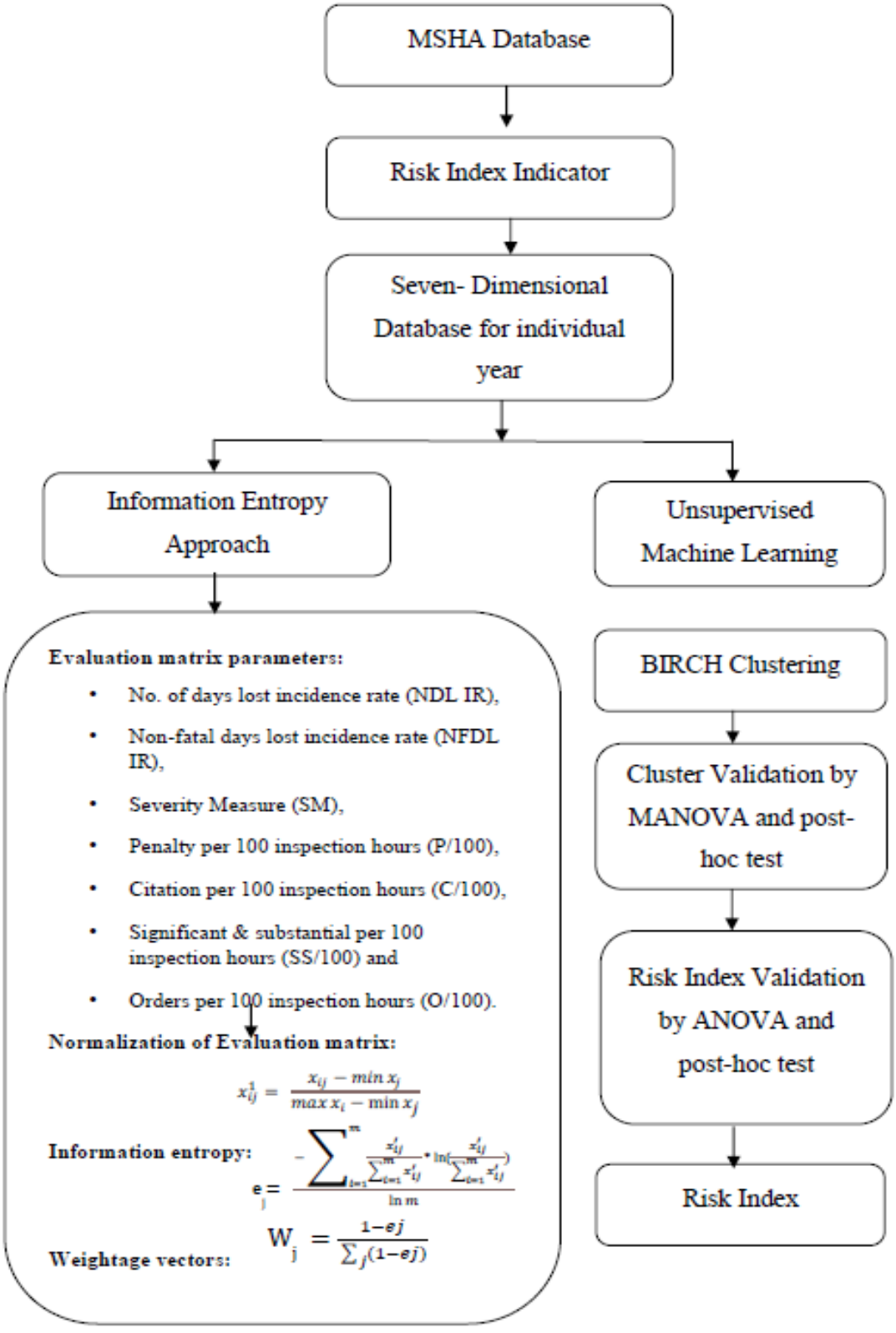


Figure 2 Flow chart to determine the entropy-based risk index

2.3.2 Weight of indexes

The seven-dimensional dataset consisting of NDL IR, NFDL IR, SM, Penalty, C/100, SS/100, and O/100 is considered for each year to determine the weightage values and the risk index. The index weights are typically determined using subjective fixed weight techniques such as the Delphi, expert survey, and analytic hierarchy process (AHP). However, the objective fixed weight methods are based on the inherent information of indexes to generate index weights. In information theory, entropy can assess the degree of disorder and relevance to system information (Li et al., 2011). The entropy-weighting method more precisely avoids the impact of subjective considerations (Zhu et al., 2020). The probability of an event happening, and the amount of information entropy has a negative mathematical relationship. The probability value will be high, and the entropy will be low if an event can be accurately anticipated to occur (Amiri et al., 2014). The attributes with a comparatively higher entropy measure have a higher data distribution between the two extremes of the solution space. The more random the input, the more impact the associated attribute will have on the algorithm's decision-making (Rastogi et al., 2015). The entropy values weaken the impact of some unusual attributes and make the assessment more precise. The essential disagreement between the responses in decision-making can be handled using the entropy technique (Kumar et al., 2021). Therefore, an entropy approach is used to analyze the weight vectors of the indicators on comprehensive evaluation through the degree of dispersion (Ji et al., 2021). The information entropy procedure follows:

Step 1: The seven-dimensional database constructs the evaluation matrix.

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 this study is normalized using the critical value approach. The normalized indicators are determined by:

$$x_{ij}^{1} = \frac{x_{ij} - min\, x_j}{max\, x_i - \min x_j} \quad (9)$$

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 (Zhu et al., 2020). According to the definition of entropy, the entropy of the jth index is determined by:

$$e_j = \frac{-\sum_{i=1}^{m} \frac{x'_{ij}}{\sum_{i=1}^{m} x'_{ij}} * \ln\left(\frac{x'_{ij}}{\sum_{i=1}^{m} x'_{ij}}\right)}{\ln m} \quad (10)$$

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

$$W_j = \frac{1-ej}{\sum_j (1-ej)} \quad (11)$$

The weight calculation helps in determining each indicator's relative importance. To make the ERI unbiased, the sum of indicator weights was kept equal to 1. The risk index was calculated by multiplying the weights with the corresponding values. The proposed steps were followed for each year, and each year's risk index was determined.

2.3.3 Validation

The proposed risk index is validated by clustering seven-dimensional risk matrices for each individual year. If the resultant risk index can capture the seven-dimensional risk matrices, the risk index values from one cluster to others should be statistically different. The MANOVA test and post-hoc techniques were applied to determine the effectiveness of the clustering results. The ANOVA and post-hoc test were performed to ensure the mean value of the risk index from different clusters is statistically different.

2.3.3.1 Unsupervised Machine Learning

With the rapid advancement of ML techniques and algorithms, a greater emphasis has been placed on incorporating machine learning into safety-related studies, which has been proven in several applications (Ji et al., 2021). BIRCH is advantageous as it focuses on intelligent cluster assignment without human participation (Roselin et al., 2021). The

silhouette coefficient was used to determine the number of optimal clusters whereas the BIRCH clustering was applied to determine the cluster labels.

2.3.3.2 Determining the Number of Clusters

The proposed risk index is validated through unsupervised ML algorithms to determine the range of clusters. The ML algorithms learn from the data and, when new data is introduced, recognize the data class using the previously learned features (Mahesh, 2018). Hence, clustering algorithms can be used to determine the similar relationships between the data groups (Lieber, 2013). However, most clustering methods are designed to evaluate the grouping or partition based on a known number of clusters (Kodinariya and Makwana, 2013). To determine the number of clusters, a silhouette method was used (Saputra, 2020). The silhouette coefficient considers the intra-cluster and inter-cluster distances for the cluster number selection (Dinh et al., 2019). The silhouette value represents the similarity of a datapoint to its cluster compared to other cluster centroids. The value ranges from -1 to +1. A higher silhouette score indicates an appropriate match of the datapoint to its cluster centroid. (Zhou and Gao, 2014). The mean for each data point's intra-cluster distance (a) and the nearest-cluster distance (b) are used to determine the silhouette coefficient (Shahapure and Nicholas, 2020):

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

If the mean of the measured silhouette value is significantly high, the number of clusters is at its optimal value (Nanjundan et al., 2019).

2.3.3.3 BIRCH Clustering

A Balanced iterative reducing and clustering (BIRCH) using hierarchies is the fastest clustering algorithm by building the clusters through the cluster feature tree (CF Tree) and height-balanced tree (Lorbeer et al., 2018). The clustering features for hierarchical clustering are stored in a height-balanced tree (CF Tree) with two components: balancing factor B and threshold T. The process of analyzing the data and inserting it into the

appropriate cluster is carried out incrementally as part of building the CF tree. (Venkatkumar and Shardaben, 2016). The balanced tree is created using the BIRCH algorithm by clustering and storing the points in each leaf node. Each leaf node has pointers to the node directly below and above it. After the BIRCH process, these pointers offer quick access to sets provided by the BIRCH algorithm during clustering. (Kovacs and Bednarik, 2011). Here, the algorithm computes the intra-cluster distances using the clustering features until the appropriate number of clusters is reached; the algorithm merges the two closest clusters. Therefore, the BIRCH clustering approach uses multilevel clustering to reduce complexity and increase flexibility. It focuses on determining the best subclusters and multidimensional group metrics to produce clusters of the best quality (Nwadiugwu, 2020).

2.3.3.4 Statistical Methods

The multivariate analysis of variance (MANOVA), which is an extension of the univariate analysis of variance (ANOVA), determines the ability to find the difference between the groups (Smith et al., 2020). MANOVA focuses on three assumptions: independence, multivariate normality, and equality of variance matrices (Appolus and Okoli, 2022). MANOVA determines the simultaneous analysis of correlations that exist between several dependent variables. The null hypothesis (H_0) states that the averages of the investigated parameters are the same. In contrast, the alternative hypothesis states that at least one parameter has a different average value for the populations being compared (Rybak et al., 2023). The MANOVA statistics such as Wilks lambda, Hotelling-Lawley trace, Pillai's trace, and Roy's largest root help in determining the p-value and test for the null hypothesis (H_0) (Anderson and Walsh, 2013). Tukey's Honestly Significant Difference (HSD) was applied as a post hoc test to compare the means of the various cluster labels. If the p-values were less than 0.05, it is determined as a statistically significant difference between the groups. (Korga et al., 2019). After testing the clusters robustness, the risk index’s validity was tested by comparing the means of the risk index value from different clusters. If the proposed weight calculation method successfully captured the seven-dimensional risk indices, the mean value of proposed risk index of

the cluster should be significantly different from the mean values of the risk index of other clusters statistically. The ANOVA test was carried out to verify the mean difference of the risk index from different clusters. Subsequently, the post-hoc test was performed to validate that all means of risk index from different clusters are statistically different. Box plot was also used to visualize the risk index values in other clusters.

2.4 Results

In the United States, surface mines are significantly higher than underground mines. However, the underground mine penalties are substantially higher than surface mine penalties. Figure 3 shows the MSHA statistics from 2011 - 2020, accounting for $410 Million in underground and $222 Million in surface penalties, totaling around 65% of relative penalties in underground mines compared to surface mines. These statistics are based on significant and substantial violations that have a possibility of causing an injury of serious nature. The penalties for underground mines have decreased from about $96.8 Million in 2011 to just over $17.5 Million in 2020. According to the citation statistics, underground mines had about 54,308 more significant and substantial citations than surface mines. According to the Safe Performance Index (SPI) analysis and the MSHA Pattern of Violation (POV) methodology, major hazard-related citations or increased citations have the potential to result in an accident or injury (Kinilakodi and Grayson, 2011b). As a result, underground mines are focused on determining the risk index, thereby ensuring workplace safety.

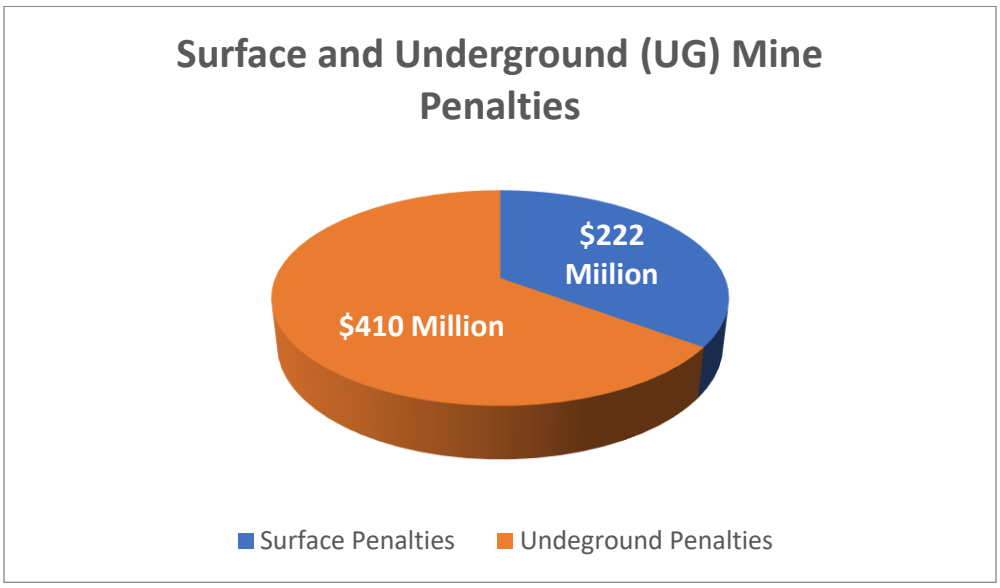


Figure 3 Surface and Underground penalties in mines

2.4.1 Risk Index Determination

Shannon's entropy was applied to determine the relative weights for the ranking procedure. According to their respective importance, hierarchy gives each attribute an overall weight. Table 3 shows the weight values using the information entropy approach for comprehensive evaluation through the degree of dispersion. All the indicator weights vary over the years, with some having more severe amounts than others. However, the significant and substantial violations (SS/100) that lead to an injury are observed more important and assigned greater weight than citations. While citations are considered an effective tool for enforcing safety regulations and promoting compliance, the citations may not completely reflect the presence of risk to the miners. For 2019, the most critical criterion is NDL IR, with around 49% weight. The weights for the severity measures almost remains constant at about 0.2.

Table 3: Weights obtained through the information entropy approach.

CAL YR	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH
2011	0.044	0.291	0.289	0.281	0.008	0.019	0.068
2012	0.092	0.180	0.202	0.290	0.025	0.045	0.166
2013	0.047	0.141	0.342	0.329	0.011	0.026	0.104
2014	0.106	0.199	0.180	0.229	0.023	0.061	0.202
2015	0.088	0.251	0.193	0.210	0.021	0.052	0.185
2016	0.087	0.287	0.148	0.216	0.029	0.059	0.173
2017	0.089	0.184	0.190	0.180	0.025	0.066	0.266
2018	0.110	0.223	0.170	0.200	0.031	0.070	0.195
2019	0.064	0.491	0.119	0.125	0.019	0.045	0.137
2020	0.105	0.197	0.173	0.179	0.041	0.090	0.215

The risk index is calculated by multiplying the determined weights with the corresponding values for individual years. Table 4 shows a sample group of mines with their risk index for the year 2019. Mine 1 has a significantly higher risk than other mines in the example table.

Table 4: Determination of Risk Index

MINE ID	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH	Risk Index
1	0.0085	1.0000	0.0000	0.0000	0.1039	0.0000	0.0000	0.4930
2	0.0797	0.0000	1.0000	1.0000	0.6119	0.5159	0.0847	0.2954
3	0.1350	0.0000	0.0000	0.0000	0.2189	0.5801	1.0000	0.1765
4	0.1238	0.0016	0.4025	0.3461	0.3254	0.2861	0.0395	0.1243
5	0.4356	0.0001	0.0169	0.1088	0.6936	0.7851	0.2206	0.1224

The top 25 mining types that provide the highest risk are shown in Table 5 for each year. The majority of underground coal mines are prone to higher risk than other mining types. The average numbers of coal mines with the higher risk accounted to be 16 out of 25.

Table 5: Top 25 high risk underground mining types from 2011-2020

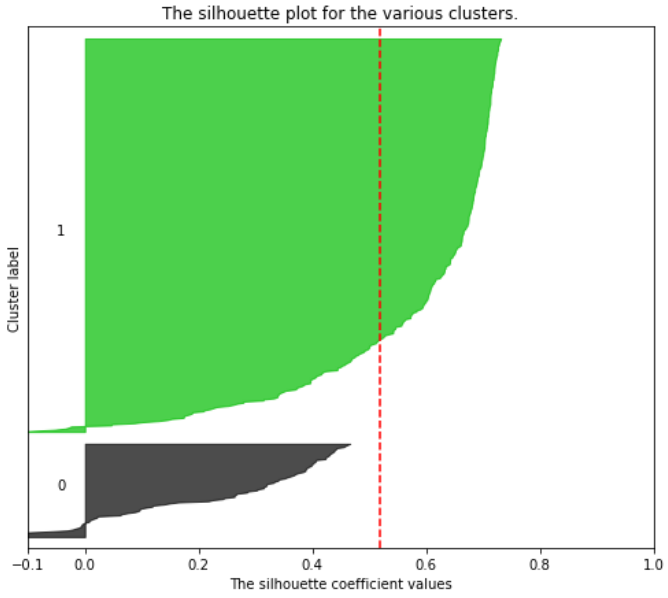
CAL YR	Coal Mines	Metal Mines	Non-metal Mines	Stone
2011	20	2	1	2
2012	17	5	1	2
2013	17	6	1	1
2014	19	5	0	1
2015	11	7	1	6
2016	17	3	0	5
2017	15	6	1	3
2018	17	2	1	5
2019	14	5	0	6
2020	15	4	1	5

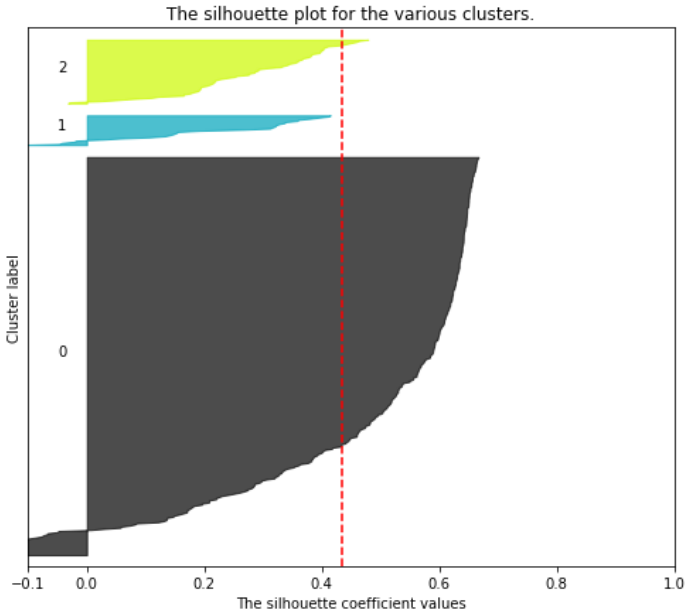
2.4.2 Validation

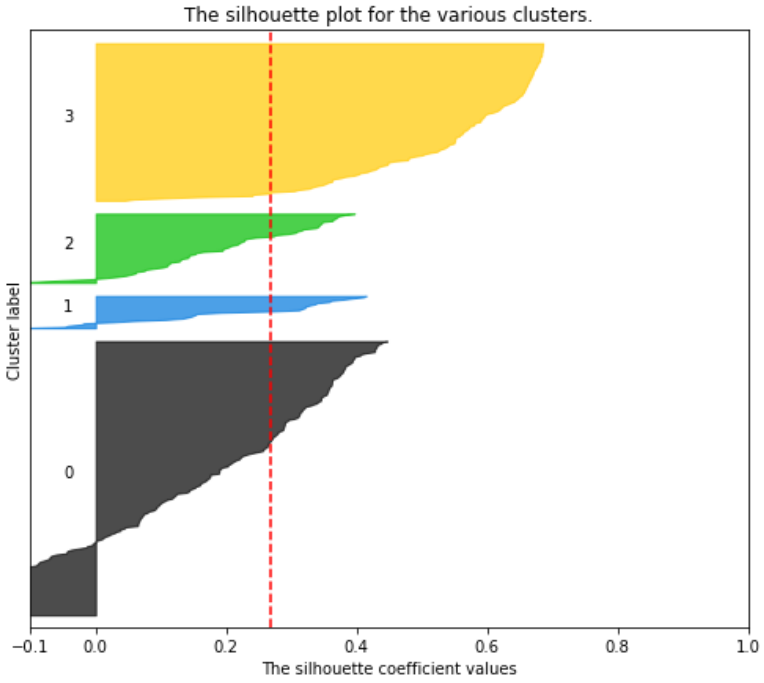
2.4.2.1 Number of clusters

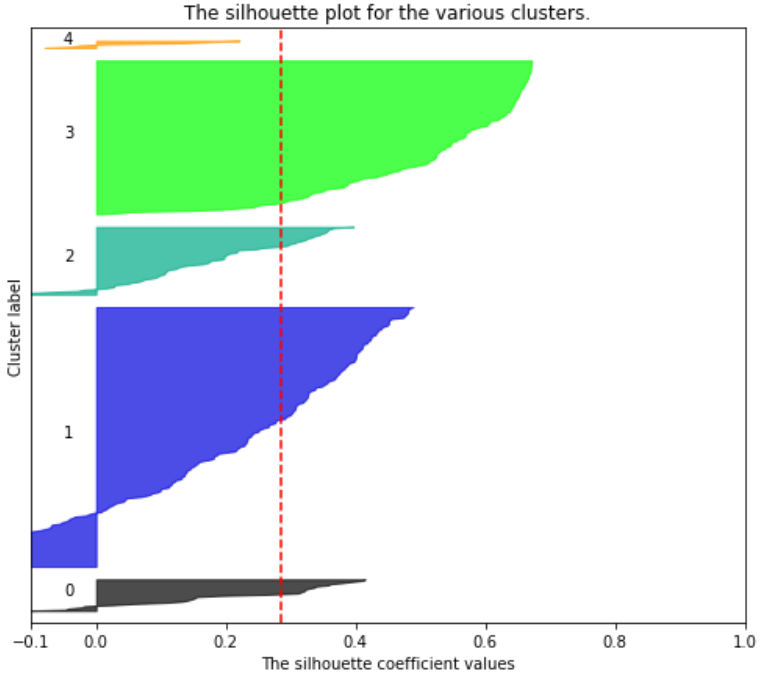
This study uses the clustering technique to group the data, which is then split into (k) clusters. The silhouette plots for the BIRCH clustering indicate the silhouette coefficient values for k clusters. The displayed results for the silhouette analysis are for 2019, as shown in Figure 4. Due to the presence of a negatively dominated cluster and the large fluctuations in the size of the clusters, the clusters values of 4, 5, and 6 are poor choices for this data. The 2 cluster model groups into distinct large clusters based on the

probability score assigned for each data point in the cluster. However, this results in a large group of individual clusters compared to other cluster numbers. The 3 cluster models have fewer overlaps compared to other clusters, and this statement is supported by the average silhouette score of 0.434 which shown as a dashed red line in Figure 4. Therefore, the ideal number of clusters is three, determined by the silhouette analysis. Similarly, the number of clusters is determined for each year, as shown in Table 6.






The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


The silhouette plot for the various clusters.
Cluster label
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values

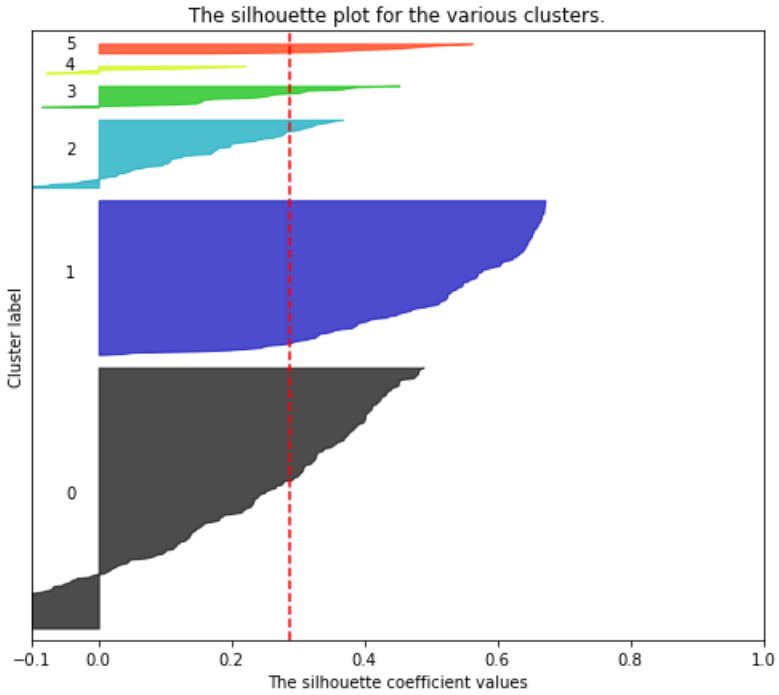


Figure 4 Silhouette analysis for various clusters using BIRCH Clustering

Table 6 Silhouette Analysis from 2011-2020

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

2.4.2.2 Statistical Analysis of Clustering

The multivariate statistical analysis of the BIRCH clustering algorithm is shown in Table 7, which focuses on Wilks lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root and determines the significant difference between the clustered groups. The 2019 dataset consisted of 451 mines, and the optimal number of clusters considered is three based on the silhouette analysis. Seven statistical variables (NDL IR, NFDL IR, SM, C/100, SS/100, O/100, and P/100) were used for clustering. The degrees of freedom for residuals (Den DF) and the number of independent variables (Num DF) are 7.0 and 443.0, respectively. The clustering technique has very low p-values ($0 < 0.05$), along with Pillai's trace indicating a significant difference between the groups and thereby rejecting the null hypothesis. Wilk's lambda has an F-statistic of 55.1727 which produces a p-value that is small enough (0) to be reported. However, the F-statistic is significantly large, indicating a significant difference between the groups. The Wilks lambda shows

around 46.58% of the variance in the dependent variable. Therefore, the statistics show a significant difference between the clusters on the BIRCH cluster labels.

Table 7 Multivariate Statistics for BIRCH clustering

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.3030	7	443	145.6087	0
Pillai's trace	0.6970	7	443	145.6087	0
Hotelling-Lawley trace	2.3008	7	443	145.6087	0
Roy's greatest root	2.3008	7	443	145.6087	0
BIRCH Clustering					
Wilks lambda	0.5342	7	443	55.1727	0
Pillai's trace	0.4658	7	443	55.1727	0
Hotelling-Lawley trace	0.8718	7	443	55.1727	0
Roy's greatest root	0.8718	7	443	55.1727	0

Since a significant difference between the clusters was found, post-hoc test of Tukey's HSD was applied to investigate the difference between the pairs of clusters. As shown in Table 8, there is a significant difference in mean between the cluster groups, and the upper and lower bounds help in determining the confidence interval by focusing on the mean difference. The reject column indicates that the null hypothesis can be rejected for all comparisons and indicates a significant difference between the groups. Similar results were also observed for other years, which are presented in Appendix A – I.

Table 8 MANOVA Post-hoc statistics for the BIRCH algorithm

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.2105	0	0.1849	0.2360	True
0	2	0.1035	0	0.0853	0.1218	True
1	2	-0.1069	0	-0.1368	-0.0771	True

The cluster validation, including MANOVA and post-hoc has shown a significant difference between the clustered groups. However, to compare the difference between

the risk index from one cluster to another, ANOVA and post-hoc tests were performed. The ANOVA test consists of two records: between the clusters and within the cluster's variation, as shown in Table 9. The degrees of freedom and sum of squares between the clusters includes 2.0 and 0.1463. The variation that cannot be explained in the clustering algorithm is termed as residual. The p-value between the clusters is quite low and close to 0 (<0.05) and rejects the null hypothesis. Therefore, there is a significant difference between the cluster groups to the risk index.

Table 9 ANOVA statistics for ERI from different clusters

	df	**sum_sq**	**mean_sq**	**F**	**Pr(>F)**
BIRCH Clustering	2.0	0.1463	0.0732	72.4015	5.6799e-28

The post-hoc statistics were performed for ANOVA to determine the significant mean difference in risk index between the cluster groups (Table 10). The p-adj is quite low, and the null-hypothesis is rejected, indicating a significant difference between the cluster groups. These results revealed that the proposed risk index (ERI) correctly captures the information contained within the seven-dimensional risk indicators. Similar results have also been observed for other years (Appendix A - I).

Table 10 Post- hoc statistics for ERI from different clusters

Group 1	**Group 2**	**Mean Diff**	**P-adj**	**Lower**	**Upper**	**Reject**
0	1	0.0643	0	0.0497	0.0790	True
0	2	0.0321	0	0.0216	0.0425	True
1	2	-0.0323	0	-0.0494	-0.0151	True

For the visual representation of the range of risk index for each cluster group in 2019, a box plot is provided as shown in Figure 5. The box plot can determine the maximum and minimum values from each cluster group. Compared to the other cluster groups shown in the box plot, the range of risk index values for cluster group 0 is the broadest.

Figure 5 Box plot for risk index based on cluster labels

2.5 Case Study

In this section, we considered an underground mine and determined the risk index over time based on the risk indicators. The underground coal mine in the Appalachian coal basin accident in 2010 highlighted the ineffectiveness of the POV. As a result, MSHA highlighted the list of hazardous underground coal mines after the accident and targeted these mines for extra inspections. The determination of SPI included consideration of equal weighting to the accident and citation measures to balance the considerations of accident and citation, thereby focusing on determining the safety performance (Kinilakodi and Grayson, 2011a). To demonstrate the effectiveness of the proposed risk index, a comparative evaluation was performed between SPI and the proposed entropy-based risk index for the study coal mine prior to the accident time. Table 11 shows the weights calculated from 2006 to 2010 using the entropy-based approach. Table 12 shows the entropy-based risk index and SPI values for the study mine during that period.

Table 11 Entropy-based weights from 2006-2011

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

Standard injury measures such as the NDL IR, NFDL IR, and SM, as well as citation-related measures, such as C/100 IH, SS/100 IH, and O/100 IH, were utilized in the calculation of the SPI. SPI shows an increased value over the years, with the highest risk index recorded in 2011. SPI predicts that 2009 and 2010 show reasonably close risk values, ignoring the small values of NDL IR, NFDL IR, and SM. However, the entropy-based risk index (ERI) shows a relatively high-risk index in 2010 and 2011. This can be observed through the methane ignitions at the mine in 2010, which led to terrible outcomes, with the latter resulting in an enormous explosion that affected the whole mine. However, SPI did not determine this, and the values decreased from 2009 to 2010, as shown in Figure 6. The highest number of citations is completely ignored in the SPI as the highest weightage was given to C/100 IH in 2010. The P/100 IH values vary over the years with the highest value recorded in 2008 and the lowest penalties in 2009. ERI is performing better than SPI, indicating that the proposed risk index manages the risks in the case study mine. Therefore, the proposed risk index is anticipated to assist mining industry executives in prioritizing hazards and associated risks and in creating appropriate action plans to eliminate (or lessen) the severity of such risks.

Table 12 Risk index based on entropy weights

CAL YR	P/100 IH	NDL IR	NFDL IR	SM	C/100 IH	SS/100 IH	O/100 IH	SPI	ERI
2006	0.0490	0.0003	0.0003	0.0002	0.1017	0.0784	0.0844	0.1163	0.012
2007	0.1074	0.0100	0.0011	0.0301	0.2368	0.1238	0.0135	0.1670	0.030
2008	0.0865	0.0030	0.0310	0.0143	0.2198	0.1311	0.0341	0.1939	0.034
2009	0.1467	0.0002	0.0003	0.0001	0.3474	0.3375	0.0835	0.2820	0.021
2010	0.1479	0.0228	0.0421	0.0600	0.4243	0.1869	0.0755	0.2673	0.072
2011	1.0000	0.0006	0.0001	0	0.2182	0.2993	0.2740	0.2991	0.070

The plots for the SPI and ERI are shown in Figure 6. The fundamental shortcoming of the equal weights method for evaluating the output is that it cannot evaluate the discriminatory power of input characteristics (Stankovic et al., 2019). Using equal weights, assuming the same relative importance, may introduce bias into estimations (Karagiannis and Karagiannis, 2023). Despite indicating an increasing risk, SPI does not capture the replica of statistics in the mines. Thus, the proposed risk index values can help the mining companies to assess the risk and thereby focus on improving the safety aspects of the mine.

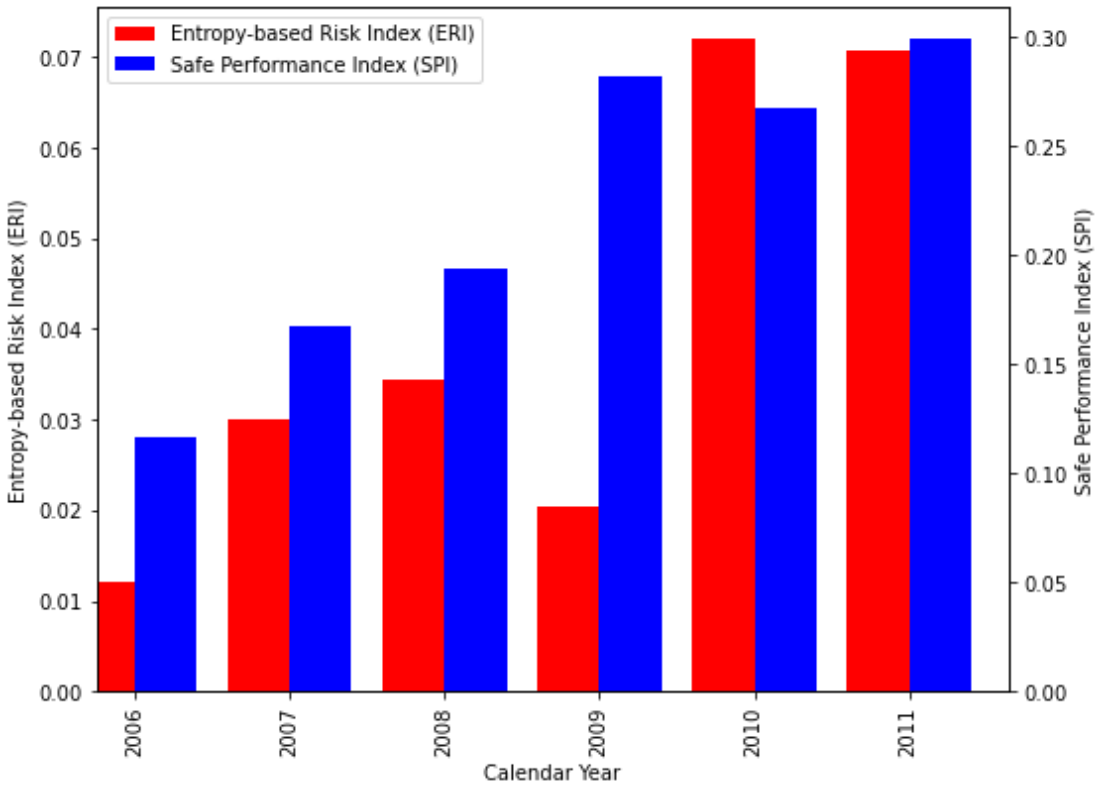


Figure 6 Risk Index based on SPI and ERI

2.6 Conclusion

The research focused on determining and validating the risk index through clustering algorithms and statistical methods. The violations, accidents, employee production, and inspections datasets from the MSHA database were used for this purpose. The entropy-based indicator weights are determined for the individual year, and the risk index is calculated. Parameters like the C/100 IH and SS/100 IH had significantly low weight compared to other indicators over the years. Machine learning algorithms were developed through the seven-dimensional dataset to validate the mine risk index. Statistical techniques such as MANOVA and post-hoc tests are performed on a seven-dimensional dataset to determine the significant difference between the clusters. Results show all the ML models performed with the MANOVA test identified p-value (<0.05), indicating a significant difference between the clusters. Detailed statistical information on mean differences in groups was determined with the help of post-hoc testing. The test statistics

show a p-value that is significantly low (0), and the mean difference between each pair of groups shows a significant difference, thereby rejecting the null hypothesis. However, to determine the significant difference between the risk index from one cluster to another, ANOVA and post hoc tests were performed. The results show a low sum of squares deviation and p-value along with significant mean difference between the clusters. Therefore, the results conclude that the risk index for each year can be determined using a seven-dimensional dataset consisting of (Penalty, no days lost incidence rate, non-fatal days lost incidence rate, severity measure, citation, order, and significant and substantial citations) based on entropy-based weights. The risk index would help the mine operators manage the workplace safety and focus on regulating the safety standards in high-risk mines.

2.7 Future Work

This research will be extended to identify the root causes of accidents by relating the calculated weights with other variables from MSHA databases, including accident narratives. A dashboard will be developed that will provide the user with some quantitative measures for the significance analysis of mine safety performance.

3. Concluding Remarks

This research focuses on the MSHA database to identify the risk indicators causing accidents and injuries and determine the risk index. An understanding of the effects of accident-causing variables assists the determination of risk indicators. Accidents and injuries in the workplace can range from being minor and resulting in no days lost from work to being significant and several days away from work. Entropy-based weights are determined for the risk indicators of individual years and multiplied with the corresponding values to assess the risk index. The mining risk was analyzed using the clustering algorithm and statistical methods. MANOVA test helped in understanding the relationship between the dependent and the independent variables. Post-hoc results helped in analyzing the mean differences between the clusters, thereby rejecting the null hypothesis. The statistical and clustering analysis helped in validating the proposed risk

index based on violations and accidents in the United States. These statistical and clustering analysis interpretations help understand the entropy-based risk index, thereby preventing accidents and injuries, and ensuring a secure environment at the mine site.

Acknowledgement

This study was partially supported by the National Institute for Occupational Safety and Health (NIOSH), under Grant Number 75D30121C12375. In addition, the second author expressed his gratitude to the Witte Family Faculty Fellow in Mining Engineering fund at Michigan Technological University for the assistance.

4. References

Amiri, V., et al. 2014. Goundwater quality assessment using entropy-weighted water quality index (EWQI) in Lenjanat, Iran. Environmental Earth Sciences, 72, 3479 - 3490. https://doi.org/10.1007/s12665-014-3255-0

Amoako, R., et al. 2021. Identifying Risk Factors from MSHA Accidents and Injury Data Using Logistic Regression. Mining, Metallurgy & Exploration,38, 509-527. https://doi.org/10.1007/s42461-020-00347-x

Amponsah-Tawiah, K., et al. 2014. The impact of physical and psychosocial risks on employee well-being and quality of life: The case of the mining industry in Ghana. Safety Science, 65, 28 – 35. https://doi.org/10.1016/j.ssci.2013.12.002

Anderson, M.J., Walsh, D.C.I., 2013. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing. Ecological Monographs, 83, 557 – 574. https://doi.org/10.1890/12-2010.1

Appolus, E.E., Okoli, C.N., 2022. A Robust Comparison Powers of Four Multivariate Analysis of Variance Tests. European Journal of Statistics and Probability, 10(1), 11-20. https://doi.org/10.37745/ejsp.2013/vol10no1pp.11-20

Arezes, M. P., Miguel, A. S., 2003. The role of safety culture in safety performance measurement. Measuring Business Excellence, 7, 20-28. https://doi.org/10.1108/13683040310509287

Biswas, A., Sarkar, B., 2019. Pythagorean fuzzy TOPSIS for multicriteria group decision-making with unknown weight information through entropy measure. International Journal of Intelligent Systems, 34, 1108-1128. https://doi.org/10.1002/int.22088

Becker, W., 2017. Weights and importance in composite indicators: Closing the gap. Ecological Indicators, 80, 12-22. https://doi.org/10.1016/j.ecolind.2017.03.056

Brian, H., 2021. Federal Mine Safety and Health Commission Matters [WWW Document]. Coal Age. URL https://www.coalage.com/legally-speaking/the-federal-mine-safety-and-health-commission-matters/ (accessed 03.08.2023)

Dinh, D.T., et al. 2019. Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient, in: Chen, J., Huynh, V., Nguyen, GN., Tang, X. (eds) Knowledge and Systems Sciences: Communications in Computer and Information Science, 1103, Springer, Singapore. https://doi.org/10.1007/978-981-15-1209-4_1

Dragan, K., et al. 2017. Organization: A new focus on mine safety improvement in a complex operational and business environment. Technology. International Journal of Mining Science and Technology, 27(4), 617-625. https://doi.org/10.1016/j.ijmst.2017.05.006

Ghorabaee, M.K., et al. 2021. Determination of Objective Weights Using a New Method Based on the Removal Effects of Criteria (MEREC). Symmetry, 13(4), 525. https://doi.org/10.3390/sym13040525

Grayson, R.L., et al. 2009. Pilot sample risk analysis for underground coal mine fires and explosions using MSHA citation data. Safety Science, 47, 1371-1378. https://doi.org/10.1016/j.ssci.2009.03.004

Grayson, R.L., Kinilakodi. H., 2011. A comparison of the 2008?2009 post-MINER Act safety performance of union and non-union underground coal mines. International Journal of Mining and Mineral Engineering, 3, 173-193. https://doi.org/10.1504/IJMME.2011.043848

Greco, S., et al. 2019. On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation and Robustness. Soc Indic Res, 141, 61 – 94. https://doi.org/10.1007/s11205-017-1832-9

Groves, W. A., et al. 2007. Analysis of fatalities and injuries involving mining equipment. Journal of Safety Research, 38(4), 461 – 470. https://doi.org/10.1016/j.jsr.2007.03.011.

Harms-Ringdahl, L., 2009. Dimensions in safety indicators. Safety Science, 47, 81-482. doi: 10.1016/j.ssci.2008.07.019

Ji., C., et al. 2021. Development of novel combustion risk index for flammable liquids based on unsupervised clustering algorithms. Journal of Loss Prevention in the Process Industries, 70.
https://doi.org/10.1016/j.jlp.2021.104422

Jimenez-Fernandez, E., et al. 2022. Dealing with weighting scheme in composite indicators: An unsupervised distance-machine learning proposal for quantitative data. Socio-Economic Planning Services, 83.
https://doi.org/10.1016/j.seps.2022.101339

Karagiannis, R., Karagiannis, G., 2023. Nonparametric estimates of price efficiency for the Greek infant milk market: Curing the curse of dimensionality with shannon entropy. Economic Modelling, 121, 106202.
https://doi.org/10.1016/j.econmod.2023.106202

Kecojevic, V., 2011. Analysis of "high-dollar" value safety and health citations and orders for the US coal mines. Safety Science, 49, 658-663.
https://doi.org/10.1016/j.ssci.2010.12.001

Kinilakodi, H., Grayson, R. L., 2011a. A methodology for assessing underground coal mines for high safety-related risk. Safety Science, 49(6), 906-911.
https://doi.org/10.1016/j.ssci.2011.02.007.

Kinilakodi, H., Grayson, R. L., 2011b. Citation-related reliability analysis for a pilot sample of underground coal mines. Accident Analysis and Prevention, 43(3), 1015-1021. doi: 10.1016/j.aap.2010.11.033.

Kinilakodi, H., et al. 2012. Evaluating Equivalence of the Safe Performance Index (SPI) to a Traditional Risk Analysis. Open Journal of Safety Science and Technology, 2(2), 47-54.
doi:10.4236/ojsst.2012.22007

Kovacs, L., Bednarik, L., 2011. Parameter optimization for BIRCH pre-clustering algorithm. IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, 475-480.
doi: 10.1109/CINTI.2011.6108553.

Komljenovic, D., et al. 2008. Injuries in US mining operations – A preliminary risk analysis, Safety Science, 46(5), 792-801.
https://doi.org/10.1016/j.ssci.2007.01.012.

Kodinariya, T.M., Makwana, P., 2013. Review on determining number of clusters in K-means Clustering. International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90-95.

Korga, A., et al. 2019. Inhibition of glycolysis disrupts cellular antioxidant defense and sensitizes HepG2 cells to doxorubicin treatment. FEBS, Open Bio, 9(5), 959-972.
https://doi.org/10.1002/2211-5463.12628

Kumar, R., et al. 2021. Revealing the benefits of entropy weights method for multi-objective optimization in machining operations: A critical review. Journal of materials research and technology, 10, 1471 – 1492.
https://doi.org/10.1016/j.jmrt.2020.12.114

Li, X., et al. 2011. Application of the Entropy Weight and TOPSIS Method in Safety Evaluation of Coal Mines. Procedia Engineering, 26, 2085 - 2091.
doi: 10.1016/j.proeng.2011.11.2410

Lieber, D., et al. 2013. Quality Prediction in Interlinked Manufacturing Processes based on Supervised & Unsupervised Machine Learning. Procedia CIRP,7,193 - 198. doi: 10.1016/j.procir.2013.05.033

Liu, P., et al. 2021. Double hierarchy hesitant fuzzy linguistic entropy based TODIM approach using evidential theory. Information Sciences, 547, 223-243. https://doi.org/10.1016/j.ins.2020.07.062

Lorbeer, B., et al. 2018. Variations on the Clustering Algorithm BIRCH. Big Data Research, 11, 44-53. https://doi.org/10.1016/j.bdr.2017.09.002

Mahesh, B., 2018. Machine Learning Algorithms - A Review. International Journal of Science and Research, 9(1). doi: 10.21275/ART20203995

Milam, O., et al. 2020. Digital Canaries: Identifying Hazardous Patterns in Msha Data Using A Machine Learner. Procedia Computer Science, 177, 227-233. https://doi.org/10.1016/j.procs.2020.10.032

Mine Safety and Health Administration (MSHA). Data sources and calculators [WWW Document]. URL https://www.msha.gov/data-and-reports/data-sources-and-calculators (accessed 03.08.2023)

Nanjundan, S., et al. 2019. Identifying the number of clusters for K-Means: A hypersphere density-based approach. https://doi.org/10.48550/arXiv.1912.00643

NIOSH, 2016. Section 8 Coding Manual [WWW Document].
URL https://www.cdc.gov/niosh/mining/UserFiles/data/codes.pdf (accessed 02.18.2023)

Nowrouzi – Kia, B., et al. 2017. Systematic review: Lost-time injuries in the US mining industry, 67(6), 442-447.
https://doi.org/10.1093/occmed/kqx077

Nwadiugwu, M., 2020. Gene-Based Clustering Algorithms: Comparison Between Denclue, Fuzzy-C, and BIRCH. Bioinformatics and Biology Insights, 14, 1-6.
doi: 10.1177/1177932220909851

Onder, S., 2013. Evaluation of occupational injuries with lost days among opencast coal mine workers through logistic regression models, 59, 86 - 92.
https://doi.org/10.1016/j.ssci.2013.05.002

Orsulak, M., et al. 2010. Risk assessment of safety violations for coal mines. International Journal of Mining Reclamation and Environment, 24(3), 244-254.
doi: 10.1080/17480931003654901

Rahmi, E., et al. 2022. Accident analysis of mining industry in the United States - A retrospective study for 36 years. Journal of sustainable mining, 21(3).
https://doi.org/10.46873/2300-3960.1345

Rastogi, M., et al. 2015. Selection and performance assessment of Phase Change Materials for heating, ventilation and air-conditioning applications. Energy Conversion and Management, 89, 260 - 269.
http://dx.doi.org/10.1016/j.enconman.2014.09.077

Robson, L.S., et al. 2017. Developing leading indicators from OHS management audit data: Determining the measurement properties of audit data from the field. Journal of Safety Research, 61, 93-103.
http://dx.doi.org/10.1016/j.jsr.2017.02.008

Roselin, A.G., et al. 2021. Intelligent Anomaly Detection for Large Network Traffic with Optimized Deep Clustering (ODC) Algorithm. IEEE Access, 9, 47243-47251.
doi: 10.1109/ACCESS.2021.3068172

Rybak, A., et al. 2023. The Impact of Removing Coal from Poland's Energy Mix on Selected Aspects of the Country's Energy Security. Sustainability, 15, 3457.
https://doi.org/10.3390/su15043457

Sammarco, J.J., et al. 2016. An analysis of roof bolter fatalities and injuries in U.S. mining. Trans Soc Min Metall Explor Inc, 340(1), pp. 11-20.
doi: 10.19150/trans.7322

Sanmiquel, L., et al. 2018. Analysis of Occupational Accidents in Underground and Surface Mining in Spain Using Data-Mining Techniques, 15(3), 462.
doi: 10.3390/ijerph15030462

Saputra, D. M., et al. 2020. Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method. Advances in Intelligent Systems Research, 172.
doi: 10.2991/aisr.k.200424.051

Shahapure, K.R., Nicholas, C., 2020. Cluster Quality Analysis Using Silhouette Score. IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 747-748.
doi: 10.1109/DSAA49011.2020.00096

Stankovic, J.J., et al. 2021. The Digital Competitiveness of European Countries: A Multiple-Criteria Approach. Journal of Competitiveness, 13(2), pp. 117-134. https://doi.org/10.7441/joc.2021.02.07

Stemn, E., 2019. Analysis of Injuries in the Ghanaian Mining Industry and Priority Areas for Research. Safety and Health at Work, 10(2), 151-165. https://doi.org/10.1016/j.shaw.2018.09.001.

Tai, X., et al. 2020. A quantitative assessment of vulnerability using social-economic-natural compound ecosystem framework in coal mining cities. Journal of Cleaner Production, 258. https://doi.org/10.1016/j.jclepro.2020.120969

Xiao, W., et al. 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. Ecological Indicators, 109, 105843. https://doi.org/10.1016/j.ecolind.2019.105843

Xu, M., et al. 2019. Supply chain sustainability risk and assessment. Journal of Cleaner Production, 225, 857 – 867. https://doi.org/10.1016/j.jclepro.2019.03.307

Venkatkumar, I.A., Shardaben, S.J.K., 2016. Comparative study of data mining clustering algorithms. 2016 International Conference on Data Science and Engineering (ICDSE), Cochin, India, 1-7. doi: 10.1109/ICDSE.2016.7823946

Yorio, P.L., et al. 2014. Interpreting MSHA Citations Through the Lens of Occupational Health and Safety Management Systems: Investigating Their Impact on Mine Injuries and Illnesses 2003–2010. Risk Anal, 34(8), 1538-53.

Zhou, H.B., Gao, J.T., 2014. Automatic method for determining cluster number based on silhouette coefficient. Advanced Materials Research, 951, 227-230.
doi: 10.4028/www.scientific.net/AMR.951.227

Zhu, Y., et al. 2020. Effectiveness of Entropy Weight Method in Decision-Making. Mathematical Problems in Engineering, 2020, Article ID 3564835. https://doi.org/10.1155/2020/3564835

Appendix

Appendix A

Table 13 Multivariate statistics for BIRCH clustering in 2011

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.2372	7	724	332.5756	0
Pillai's trace	0.7628	7	724	332.5756	0
Hotelling-Lawley trace	3.2155	7	724	332.5756	0
Roy's greatest root	3.2155	7	724	332.5756	0
BIRCH Clustering					
Wilks lambda	0.4527	7	724	125.0320	0
Pillai's trace	0.5473	7	724	125.0320	0
Hotelling-Lawley trace	1.2089	7	724	125.0320	0
Roy's greatest root	1.2089	7	724	125.0320	0

Table 14 MANOVA Post-hoc statistics for BIRCH algorithm in 2011

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	-0.0778	0	-0.0844	-0.0713	True

Table 15 ANOVA statistics for BIRCH clustering and risk index in 2011

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	1.0	0.03984	0.03984	64.506	3.857e-15

Table 16 Post-hoc statistics for BIRCH clustering and risk index in 2011

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	-0.0224	0	-0.0278	-0.0169	True

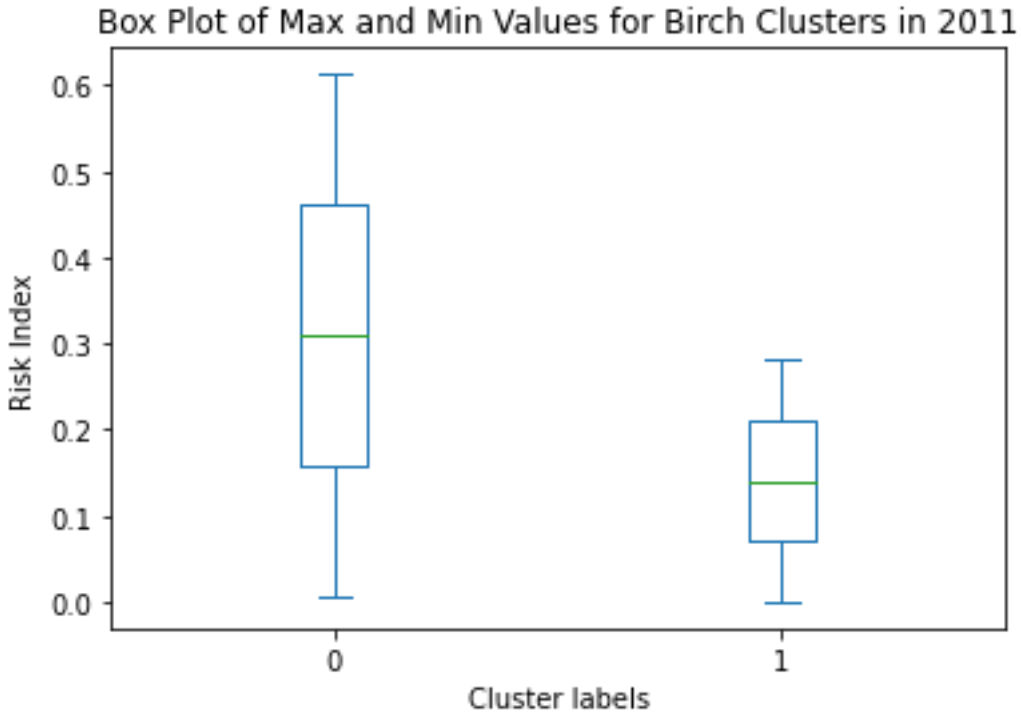


Figure 7 Box plot for 2011

Appendix B

Table 17 Multivariate statistics for BIRCH clustering in 2012

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.3664	7	706	174.3892	0
Pillai's trace	0.6336	7	706	174.3892	0
Hotelling-Lawley trace	1.7291	7	706	174.3892	0
Roy's greatest root	1.7291	7	706	174.3892	0
BIRCH Clustering					
Wilks lambda	0.2796	7	706	259.9215	0
Pillai's trace	0.7204	7	706	259.9215	0
Hotelling-Lawley trace	2.5771	7	706	259.9215	0
Roy's greatest root	2.5771	7	706	259.9215	0

Table 18 MANOVA Post-hoc statistics for BIRCH algorithm in 2012

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.0838	0	0.0751	0.0926	True

Table 19 ANOVA statistics for BIRCH clustering and risk index in 2012

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	1.0	0.1329	0.1329	134.7278	1.1957e-28

Table 20 Post-hoc statistics for BIRCH clustering and risk index in 2012

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.058	0	0.0482	0.0678	True

Figure 8 Box plot for 2012

Appendix C

Table 21 Multivariate statistics for BIRCH clustering in 2013

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.2128	7	625	330.2872	0
Pillai's trace	0.7872	7	625	330.2872	0
Hotelling-Lawley trace	3.6992	7	625	330.2872	0
Roy's greatest root	3.6992	7	625	330.2872	0
BIRCH Clustering					
Wilks lambda	0.4554	7	625	106.7780	0
Pillai's trace	0.5446	7	625	106.7780	0
Hotelling-Lawley trace	1.1959	7	625	106.7780	0
Roy's greatest root	1.1959	7	625	106.7780	0

Table 22 MANOVA Post-hoc statistics for BIRCH algorithm in 2013

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	-0.0631	0	-0.0715	-0.0547	True

Table 23 ANOVA statistics for BIRCH clustering and risk index in 2013

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	1.0	0.0053	0.0053	6.044	0.0142

Table 24 Post-hoc statistics for BIRCH clustering and risk index in 2013

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	-0.008	0.0142	-0.0144	-0.0016	True

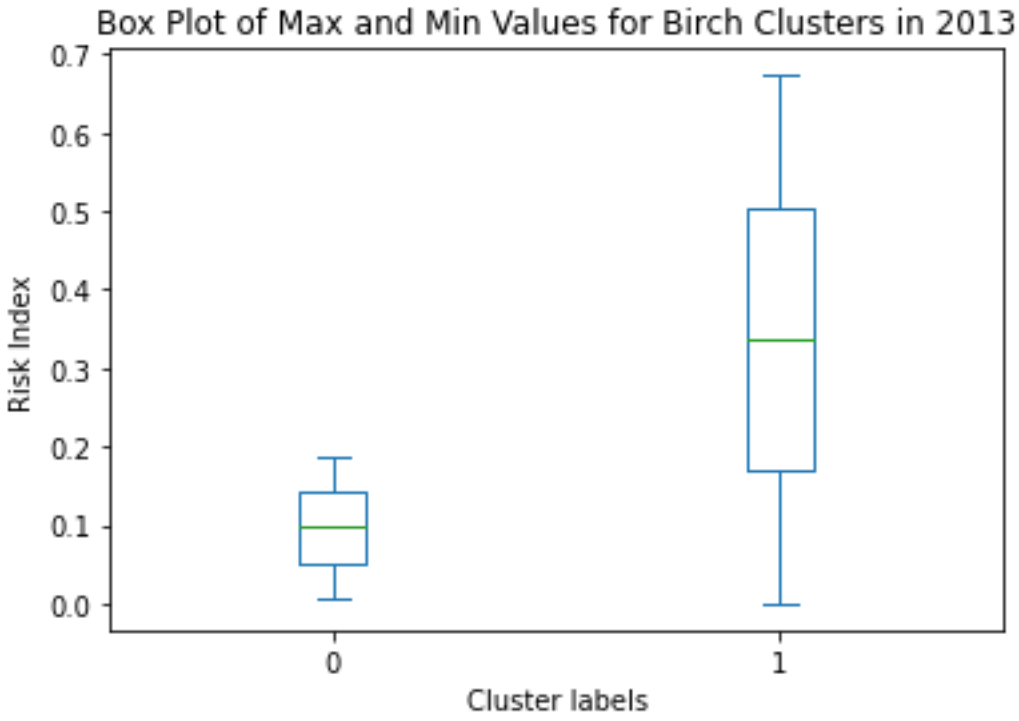


Figure 9 Box plot for 2013

Appendix D

Table 25 Multivariate statistics for BIRCH clustering in 2014

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.2333	7	585	274.6432	0
Pillai's trace	0.7667	7	585	274.6432	0
Hotelling-Lawley trace	3.2863	7	585	274.6432	0
Roy's greatest root	3.2863	7	585	274.6432	0
BIRCH Clustering					
Wilks lambda	0.3612	7	585	147.8192	0
Pillai's trace	0.6388	7	585	147.8192	0
Hotelling-Lawley trace	1.7688	7	585	147.8192	0
Roy's greatest root	1.7688	7	585	147.8192	0

Table 26 MANOVA Post-hoc statistics for BIRCH algorithm in 2014

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.0893	0	0.0801	0.0984	True

Table 27 ANOVA statistics for BIRCH clustering and risk index in 2014

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	1.0	0.1507	0.1507	84.9182	5.393e-19

Table 28 Post-hoc statistics for BIRCH clustering and risk index in 2014

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.0403	0	0.0317	0.0489	True

Figure 10 Box plot for 2014

Appendix E

Table 29 Multivariate statistics for BIRCH clustering in 2015

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.3920	7	535	118.5313	0
Pillai's trace	0.6080	7	535	118.5313	0
Hotelling-Lawley trace	1.5509	7	535	118.5313	0
Roy's greatest root	1.5509	7	535	118.5313	0
BIRCH Clustering					
Wilks lambda	0.3798	7	535	124.8201	0
Pillai's trace	0.6202	7	535	124.8201	0
Hotelling-Lawley trace	1.6332	7	535	124.8201	0
Roy's greatest root	1.6332	7	535	124.8201	0

Table 30 MANOVA Post-hoc statistics for BIRCH algorithm in 2015

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.2661	0	0.2237	0.3084	True
0	2	0.0415	0	0.031	0.052	True
0	3	0.1012	0	0.0858	0.1167	True
1	2	-0.2245	0	-0.2672	-0.1819	True
1	3	-0.1648	0	-0.2089	-0.1207	True
2	3	0.0597	0	0.0434	0.076	True

Table 31 ANOVA statistics for BIRCH clustering and risk index in 2015

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	3.0	0.1516	0.0505	67.4972	4.5517e-37

Table 32 Post-hoc statistics for BIRCH clustering and risk index in 2015

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.1303	0	0.1033	0.1572	True
0	2	0.0097	0.0012	0.003	0.0163	True
0	3	0.0292	0	0.0193	0.039	True
1	2	-0.1206	0	-0.1478	-0.0934	True
1	3	-0.1011	0	-0.1292	-0.073	True
2	3	0.0195	0	0.0091	0.0299	True

Figure 11 Box plot for 2015

Appendix F

Table 33 Multivariate statistics for BIRCH clustering in 2016

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.2762	7	462	172.9663	0
Pillai's trace	0.7238	7	462	172.9663	0
Hotelling-Lawley trace	2.6207	7	462	172.9663	0
Roy's greatest root	2.6207	7	462	172.9663	0
BIRCH Clustering					
Wilks lambda	0.4121	7	462	94.1612	0
Pillai's trace	0.5879	7	462	94.1612	0
Hotelling-Lawley trace	1.4267	7	462	94.1612	0
Roy's greatest root	1.4267	7	462	94.1612	0

Table 34 MANOVA Post-hoc statistics for BIRCH algorithm in 2016

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.1158	0	0.1023	0.1293	True

Table 35 ANOVA statistics for BIRCH clustering and risk index in 2016

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	1.0	0.0827	0.0827	50.6763	4.1231e-12

Table 36 Post-hoc statistics for BIRCH clustering and risk index in 2016

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.04	0	0.029	0.0511	True

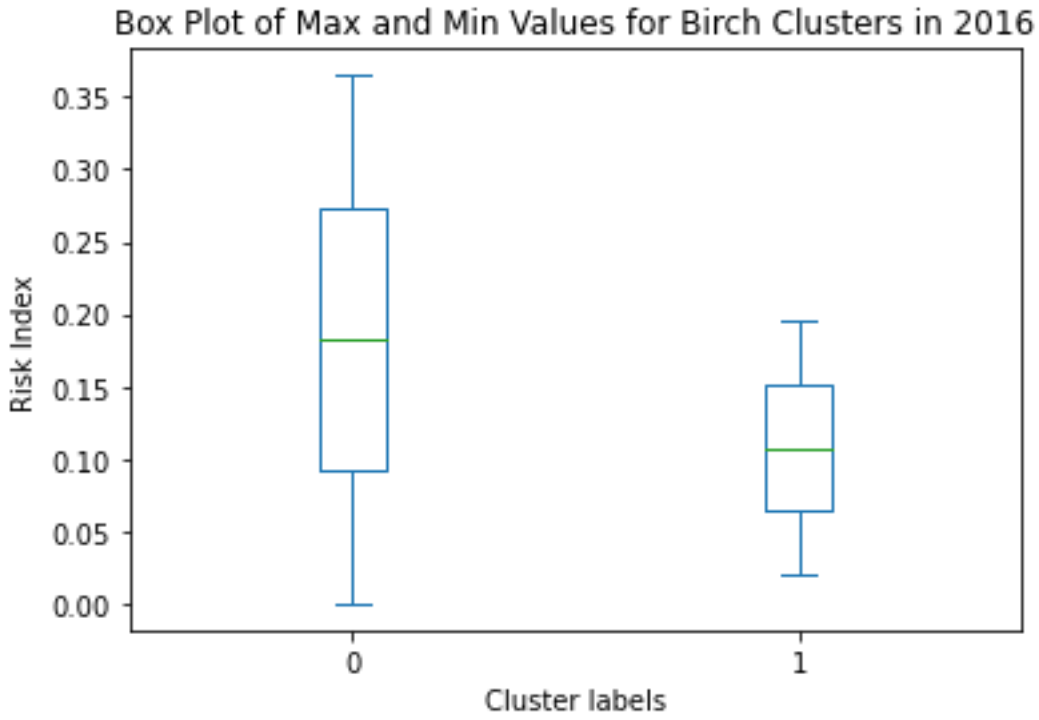


Figure 12 Box plot for 2016

Appendix G

Table 37 Multivariate statistics for BIRCH clustering in 2017

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.2841	7	456	164.1378	0
Pillai's trace	0.7159	7	456	164.1378	0
Hotelling-Lawley trace	2.5197	7	456	164.1378	0
Roy's greatest root	2.5197	7	456	164.1378	0
BIRCH Clustering					
Wilks lambda	0.4710	7	456	73.1587	0
Pillai's trace	0.5290	7	456	73.1587	0
Hotelling-Lawley trace	1.1230	7	456	73.1587	0
Roy's greatest root	1.1230	7	456	73.1587	0

Table 38 MANOVA Post-hoc statistics for BIRCH algorithm in 2017

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	-0.0764	0	-0.0861	-0.0667	True

Table 39 ANOVA statistics for BIRCH clustering and risk index in 2017

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	1.0	0.0580	0.0580	67.2539	2.3826e-15

Table 40 Post-hoc statistics for BIRCH clustering and risk index in 2017

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	-0.0349	0	-0.0432	-0.0265	True

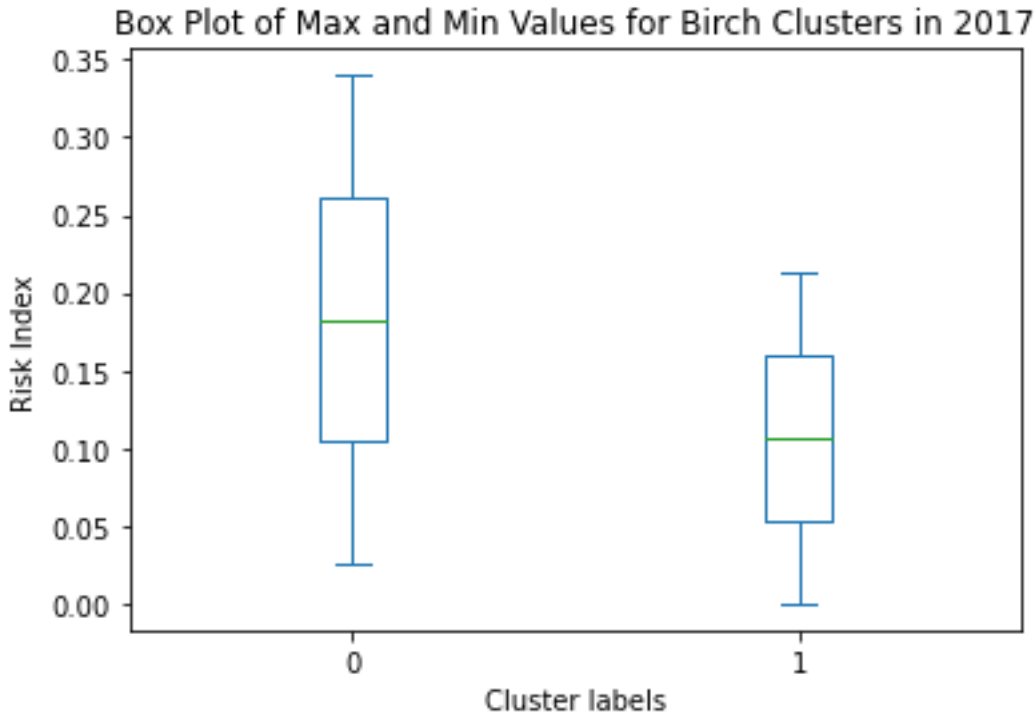


Figure 13 Box plot for 2017

Appendix H

Table 41 Multivariate statistics for BIRCH clustering in 2018

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.3188	7	456	139.1978	0
Pillai's trace	0.6812	7	456	139.1978	0
Hotelling-Lawley trace	2.1368	7	456	139.1978	0
Roy's greatest root	2.1368	7	456	139.1978	0
BIRCH Clustering					
Wilks lambda	0.4568	7	456	77.4770	0
Pillai's trace	0.5432	7	456	77.4770	0
Hotelling-Lawley trace	1.1893	7	456	77.4770	0
Roy's greatest root	1.1893	7	456	77.4770	0

Table 42 MANOVA Post-hoc statistics for BIRCH algorithm in 2018

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.1679	0	0.1478	0.188	True
0	2	0.0798	0	0.0679	0.0917	True
1	2	-0.0881	0	-0.1101	-0.066	True

Table 43 ANOVA statistics for BIRCH clustering and risk index in 2018

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	2.0	0.5649	0.2825	152.936	1.134e-51

Table 44 Post-hoc statistics for BIRCH clustering and risk index in 2018

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.1377	0	0.1178	0.1575	True
0	2	0.0419	0	0.0301	0.0537	True
1	2	-0.0958	0	-0.1175	-0.0740	True

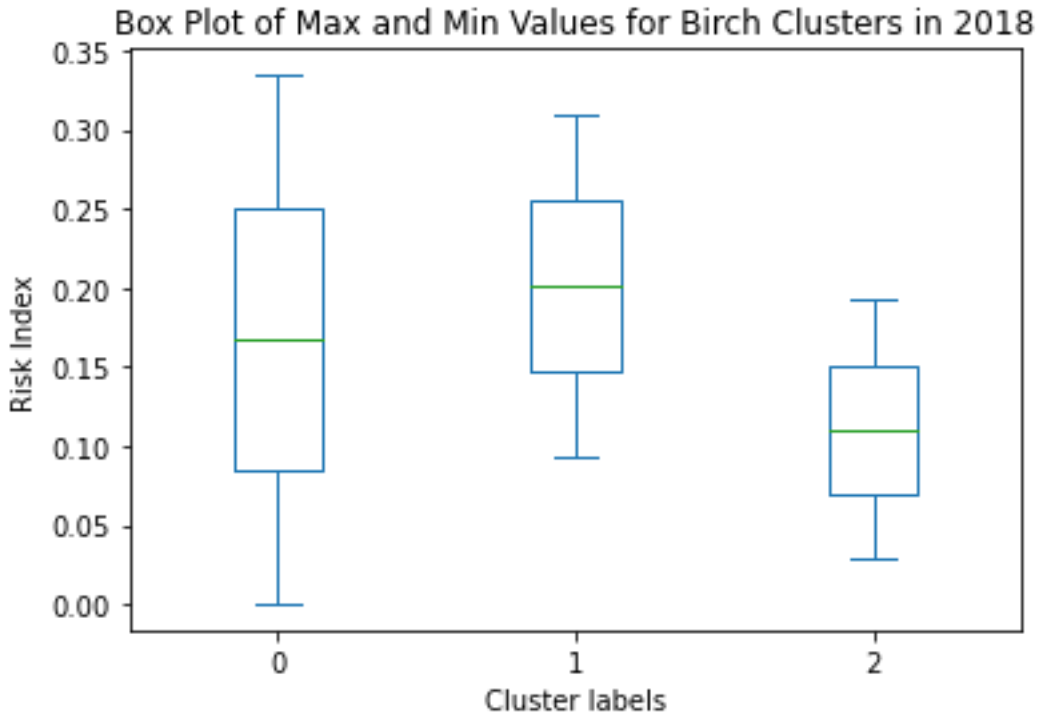


Figure 14 Box plot for 2018

Appendix I

Table 45 Multivariate statistics for BIRCH clustering in 2020

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks lambda	0.1889	7	423	259.5413	0
Pillai's trace	0.8111	7	423	259.5413	0
Hotelling-Lawley trace	4.2950	7	423	259.5413	0
Roy's greatest root	4.2950	7	423	259.5413	0
BIRCH Clustering					
Wilks lambda	0.3879	7	423	95.4670	0
Pillai's trace	0.6124	7	423	95.4670	0
Hotelling-Lawley trace	1.5798	7	423	95.4670	0
Roy's greatest root	1.5798	7	423	95.4670	0

Table 46 MANOVA Post-hoc statistics for BIRCH algorithm in 2020

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	-0.0294	0	-0.0501	-0.0087	True
0	2	-0.1215	0	-0.1341	-0.1088	True
1	2	-0.092	0	-0.1110	-0.0731	True

Table 47 ANOVA statistics for BIRCH clustering and risk index in 2020

	df	sum_sq	mean_sq	F	Pr(>F)
BIRCH Clustering	2.0	0.8421	0.4210	182.351	5.24e-58

Table 48 Post-hoc statistics for BIRCH clustering and risk index in 2020

Group 1	Group 2	Mean Diff	P-adj	Lower	Upper	Reject
0	1	0.0256	0.0115	0.0047	0.0465	True
0	2	-0.0849	0	-0.0976	-0.0722	True
1	2	-0.1105	0	-0.1296	-0.0914	True

Figure 15 Box plot for 2020