Investigating Occupational and Operational Industrial Safety Data through Business Intelligence and Machine Learning

Nakhal A., A.J.[a], Patriarca, R.[a], Di Gravio, G.[a], Antonioni, G.[b], Paltrinieri, N.[c]*

[a] Department of Mechanical and Aerospace Engineering, Sapienza University, Rome (ITALY)
[b] Department of Civil, Chemical, Environmental, and Materials Engineering, Alma Mater Studiorum – University of Bologna (ITALY)
[c] Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Trondheim (NORWAY)

* Corresponding author: Nakhal Akel, Antonio Javier.
Mail: antonio.nakhal@uniroma1.it Tel: +39 06 44585252.
Sapienza University of Rome, Department of Mechanical and Aerospace Engineering, Via Eudossiana – 18, 00184 Rome (Italy)

Abstract

Learning from previous events represents a crucial element to improve the design and operations of industrial processes, especially considering the many variables characterizing the functioning of a plant. This learning process aims to reduce the frequency of incidents and/or mitigate their severity, which are both continuous and open challenges.

This paper is grounded on a large incident repository, i.e., the Major Hazard Incident Data Service (MHIDAS) database, which was developed in 1986 by the Health and Safety Executive (HSE) to provide a reliable source of data on major hazard incidents involving hazardous materials. The database includes more than 9000 reports collected over five decades(1950s-1990s). This paper aims to provide a novel understanding of the industrial incidents reported in MHIDAS and unveil possible ways of exploring occupational/operational incidents through descriptive and quantitative analyses. Consequently, this paper proposes the implementation of Business Intelligence (BI) tools to facilitate dynamic data visualization and Machine Learning (ML) algorithms for the extraction of knowledge from different data entries. Therefore, after engineering the MHIDAS data model, a set of BI dashboards was designed and complemented with a ML-driven categorization of incidents through representative key variables for occupational/operational incidents.

The manuscript describes the process necessary to create a BI model for safety data management in an industrial context, and its integration with ML solutions that may support an in-depth multi-variate investigation of reported data. The investigation provides evidence on the importance of a precise reporting of safety events, thus unveiling the potential for lessons learned in the process industry.

Keywords

Occupational safety, Operational safety, MHIDAS, industrial processes, Business Intelligence, Machine Learning

1. Introduction

Occupational Safety and Health (OSH) should be a widespread and necessary concern for all companies globally and across all economic sectors (Adaku et al., 2021; Blanc and Escobar Pereira, 2020; Sarkar and Maiti, 2020). Traditionally, accidents are defined either as the consequence of unsafe acts or unsafe physical working conditions, or failures of technological systems (Dekker, 2019; Stefana et al., 2019; Väyrynen et al., 2015).

Occupational safety is a branch of safety science aiming to provide the employers with a clean and safe operational environment, free from known dangers. Law and standards help employers to prevent workplace injuries, illnesses, and deaths (Dekker, 2019; Manuele, 2008). Previous research in this area mainly focused on five key points: (i) Comprehensive and widening OSH both as a specific process and as embedded in the work system and processes; (ii) Total quality by employers' participation; (iii) Excellence in leadership and management; (iv) Regulation-based and voluntary reporting; (v) Effectiveness of policies (Väyrynen et al., 2015). In addition to traditional developments, modern advocacy towards occupational safety promotes increasing interest in establishing consistent safety efforts and taking more effective precautions against hazards (Sarkar and Maiti, 2020). Exemplary recent research conducted in this area describe the paths for a successful implementation of COHSMS (Certified Occupational Health and Safety Management System) through internal contextual elements that need to be approved before starting a certification request (Uhrenholdt Madsen et al., 2020).

Operational safety is a field of safety science that focuses on prevention, mitigation, and response to major accidents. While these terms are sometimes used interchangeably, it remains clear that operational safety is a broader branch that encompasses process safety. In this regard, operational risk assessment refers to the overall process of risk identification, risk analysis, and risk evaluation during operational phases(Monteiro, 2020; Monteiro et al., 2020; Yang et al., 2018). For each identified risk, the risk manager should develop and implement strategies that: (i) seek to prevent or minimize the occurrences causing losses; (ii) investigate, evaluate, defend, and settle claims resulting from such occurrences; and (iii) establish risk financing mechanisms to pay for the losses (Blanc and Escobar Pereira, 2020; Dekker, 2019).

Over recent years, the approaches in operational risk management have been oriented towards considering precautions and engineering work processes against hazards. Some examples refer to the analysis of the impact an organizational structure may receive in case of a major incident (Monteiro, 2020; Monteiro et al., 2020). These analyses can be extended through systemic modeling, as documented in an operational risk management restricted to project management (Abdulkhaleq et al., 2015). A consistent and systematic analysis is therefore required to ensure the highest possible capability for reducing adverse events and preventing major consequences in case of safety events (Micán et al., 2019; Zhang et al., 2019).

The analysis of an incident should involve an in-depth data management that spans over technical, human, and organizational issues (Lees', 2012). In turn, a thorough investigation is expected to generate a large quantity of data stored in dedicated reporting systems. Ideally, these reports should be standardized to allow a higher data quality and the systematic recreation of events, while still

ensuring a certain level of flexibility to deal with different industries and domains (Marle and Vidal, 2016). To ensure organizational or even inter-organizational learning, incidents must be registered in dedicated database. On this basis, safety reporting systems are used to manage the data surrounding accidents and injuries in the workplace and among the clients of an organization. Common elements reported in safety reporting systems relate to the event, including date, time, nature of the event, cause, injuries, name of injured parties, description of the event, location, witnesses, medical care required and mitigation (Lees', 2012; Zarei et al., 2019).

Specifically, this paper focuses on MHIDAS database, whose structure is discussed in the 'Exploring MHIDAS database' section. MHIDAS includes a large repository of industrial incidents reports from all continents. The use of safety reports stored in MHIDAS has been relevant since its early adoption and it has improved over time, as proven by different applications (Carol, Vilchez, and Casal 2002). Some examples include: the analysis of past events to support and improve risk prevention for future industrial processes and to determine critical factors for undesired events (Tauseef et al., 2011; Villafañe et al., 2011); the study of the cascading nature of incidents as for the domino theory (Abdolhamidzadeh et al., 2011; Swuste et al., 2019); and a risk assessment in industrial facilities, focused on main variables for reactive severity evaluation (Paltrinieri et al., 2020).

Despite ad hoc modeling solutions, the design of a database-wide Business Intelligence (BI) solution can be used to fully grasp information hidden in reported incident data, thus empowering flexible analyses. BI usually involves the delivery and integration of relevant and useful business information in an organization (Chaudhuri et al., 2011; Sharda, 2020; Watson and Wixom, 2007). Consequently, BI can improve the decision-making processes at all levels of management by integrating different data sources to a unique and consistent environment for safety analysis (Dorsey et al., 2020; Patriarca et al., 2016). In many cases, such decisions are routinely automated, thus eliminating the need for managerial interventions. BI prospered since computer applications moved away from transaction processing and monitoring activities towards problem analysis and solution applications. Nowadays many of the BI activities are performed through cloud-based technologies, in many cases accessed through mobile devices. The growth in hardware, software, and network capacities facilitated the implementation of the above-mentioned solutions. However, it is also possible to identify further developments that contributed to this expansion, some of the main are (Al-Aqrabi et al., 2015; Ciampi et al., 2021): (i) Gain Insights into data behavior to improve the visibility of the data and to group communication and collaboration teams. (ii) To Turn Data into Actionable Information: a BI system can help to better understand the implications of various processes and enhance the ability to identify suitable opportunities for the data, thus allowing to make plans for a successful future, by providing analytical support and overcoming cognitive limits in processing and storing information. (iii) To Improve efficiency in the data model in order to manage a large quantity of data services by improving data management (Ariyachandra and Watson, 2005; Sharda et al., 2018).

The application of BI is a novel and still unexplored topic in safety management despite its leading role in modern enterprise management. In regard to safety domains, previous works exploited the potential of database management for the identification of behaviors or factors for a more efficient

risk prevention and safer environment system (Morgan, 2021). Moreover, BI allows to manage efficiently large sets of data, i.e., data collected from failures and undesired events in industrial processes (Jamshidi et al., 2017).

In addition to BI, Machine Learning (ML) can be further adopted to build algorithms that rely on a collection of examples of some phenomena. ML has several practical applications in modern industrial settings (De Felice et al., 2019; Zhu et al., 2021), and it can be the key to unlocking the value of safety data, fostering a proactive operational and occupational risk management (Edwards, 2016; Khan et al., 2019). Practical examples of ML adoption in safety management refer to predictions of system's losses and other risks in undesired cases (Paltrinieri et al., 2019).

Based on this stream of knowledge, this paper highlights both occupational and operational features of reported safety data, as available in MHIDAS. To this extent, this paper aims to investigate how BI tools can be used to support the analysis of reported industrial incidents and disclose occupational/operational data. This aim is then complemented by ML solutions with the purpose of investigating potential clustering of diverse events into a manageable set of categories that may economically represent occupational/operational key variables. The presented paper provides a methodological discussion and a roadmap for BI and ML techniques adoption in safety domain. Hence, it supports a dynamic multivariate analysis based on an underlying structured data model.

The remainder of the paper is structured as follows. Section 2 provides a brief summary of the database used in this research, including its history and structure. Section 3 explains the methodology developed for the analysis, with specific references to the BI and ML solutions. Section 4 presents the analysis applied on MHIDAS, from the creation of the BI data model, the corresponding descriptive analysis, and the subsequent integration with ML-driven clustering algorithms. Finally, Section 5 provides managerial implications, and Section 6 summarizes weaknesses and strengths of the proposed approach.

2. Exploring MHIDAS database

The manuscript is grounded on data available in the MHIDAS database, a repository that includes reports of all types of accidents in industrial settings related to hazardous materials (Harding, 1997).

The MHIDAS database was created following the 1970's investigation by the United Kingdom (UK) Health and Safety Executive (HSE) on operational hazards. The study was carried out by the Safety and Reliability Directorate (SRD) of the UK Atomic Energy Authority (AEA) and became the most comprehensive non-nuclear application of risk assessment techniques at the time, revealing several areas where reliable data were not available (Harding, 1997).

A study was then commissioned by the HSE to collect information. The scope of the study was widened to include toxic releases, and, eventually, to cover "those incidents involving hazardous materials that resulted in or had the potential to produce an off-site impact". The importance of this study was reinforced by the occurrence of several major accidents (Seveso, 1976, Mexico City and Bhopal, 1984).

The operating version of MHIDAS was launched in 1986 by UKAEA (as SRD) and the UK HSE. The database draws on public domain information sources (press cuttings, magazine articles, journals, published reports) to ensure that the information can be widely disseminated, and it has been continuously updated since its inception. There are currently records of over 9000 incidents worldwide, including information on incidents that occurred prior to the launch of the database. The database was intended to provide a twofold usage:

- *Learning from past incidents*: to see what has happened, how it happened and what consequences it brought. This first usage is meant both for designers, to avoid falling into replicating previous mistakes, and for emergency planners, to appreciate the type and scope of incident that they may be expected to cope with.

- Where possible in terms of sample size, *developing metrics* for incident frequencies to use for reactive or even proactive risk assessment.

Regarding the information on MHIDAS, the database contains a detailed compilation of all the parameters necessary to know precisely what happened and how the accident can be described. Table 1 presents the main fields available in MHIDAS to collect information. From an operational safety perspective, one of the most notable features of MHIDAS is that the database registers one record for each substance involved in an accident. Therefore, the number of registrations for an accident equates to the number of substances involved in that accident (Llopart, 2001).

Table 1. Description of each parameter used on MHIDAS database (Llopart, 2001)

CODE	MEANING	DESCRIPTION
AB	Abstract	Brief summary of the incident, with detailed information.
AN	Record number	Registration number on the database.
CR	Contributor	Source of incident information.
DA	Date of incident	Date of the incident in the form DD/MM/YY or close to a previous or later date.
DG	Damage	Estimate (in dollars) of the material damage caused by the incident.
GC	General causes	General cause of incident e.g., Mechanical, Impact, etc.
IS	Ignition source	Code associated with the ignition source, which in its case, activates the fire/explosion, such as hot surface, cigarette, etc.
IT	Incident type	Code associated with the actual incidents that occurred, with historical evolution in case there is more than one (fire/explosion), such as fire, pool fire, etc.
KW	Key words	Indication, by means of a series of codes, of whether there is additional information available on some additional aspects.

LO	Location of incident	Indicates the location of the incident by three positions: City / Region / Country.
MC	Material code	Code used to reference the material name, i.e., UN Numbers.
ME	Major event	N: major accidents involving death and/or damage. Y: accidents that do not specify death and/or damage
MH	Material hazard	Field used to associate the most likely risk for each material or situation, whether it occurs, such as TO (toxic), FI (combustible), EX (explosive), etc.
MN	Material name	Substance name involved in the accident.
NP	People affected	Estimate number of fatalities, injured, or evacuated for consequences of the incident.
OG	Origin	Where the incident originated, such as, transport, process plant, etc.
PD	Population density	Indicates the population density of the affected area.
QY	Quantity of material	Estimation of the amount of material involved in the accident.
RA	References available	Amount of documentation such as articles or texts available about the incident for review.
SC	Specific causes	Specific cause of the incident, such as "overheat", "overload", etc.

3. Methods

This section introduces the concepts of descriptive Business Analytics and the nature of data, along with strategies to manage information through Extraction-Transformation-Loading processes. These processes represent the preliminary steps for the construction of data models that allow to perform descriptive analyses, while also constituting the basis for Machine Learning algorithms, i.e. hierarchical clustering on occupational/operational data.

3.1. Business Analytics

The word "analytics" has largely replaced the previous individual components of computerized decision support technologies that have been available under various labels in the past. Analysis combines computer technology, management science techniques, and statistics to solve real problems (Sharda, 2020). For this reason, the Business Analytics (BA) process concerns the ability to use information in ways that can improve the way the organization functions. Moreover, BA aims to boost the ability to manage the access and the availability of information to assess business

needs, identify data sources, and effectively manage the flow of information within an appropriate framework (Loshin, 2013).

Safety data is usually defined as the capability of going beyond raw data to extract information, i.e., referring to the notion of safety intelligence (Patriarca et al., 2019). All these applications are made possible by the analysis and interrogation of the data gathered by an organization. The level of analysis can involve statistical analysis to better understand the patterns. Moreover, this process can be followed by a further step to develop forecasts or models to predict customer response to a specific marketing campaign or ongoing service or product offerings. When an organization has a comprehensive vision of ongoing situations and possible future ones, it can also use other techniques to make the best decisions under specific circumstances (Sharda, 2020). Figure 1 represents a graphic overview of these three levels of analysis. By observing the presented figure, it is possible to notice that the three levels can be considered as inter-dependent steps, since one type of analytical application leads to another. Moreover, the figure suggests that there is some overlap between the three types of analysis, where the descriptive side relies on finding out the structure of past data (i.e., what happened), while the predictive (i.e., what will happen) and prescriptive (i.e., what should I do) sides seek to give a sense of prevention and/or improvement to the analyzed process data (Sharda et al., 2018).

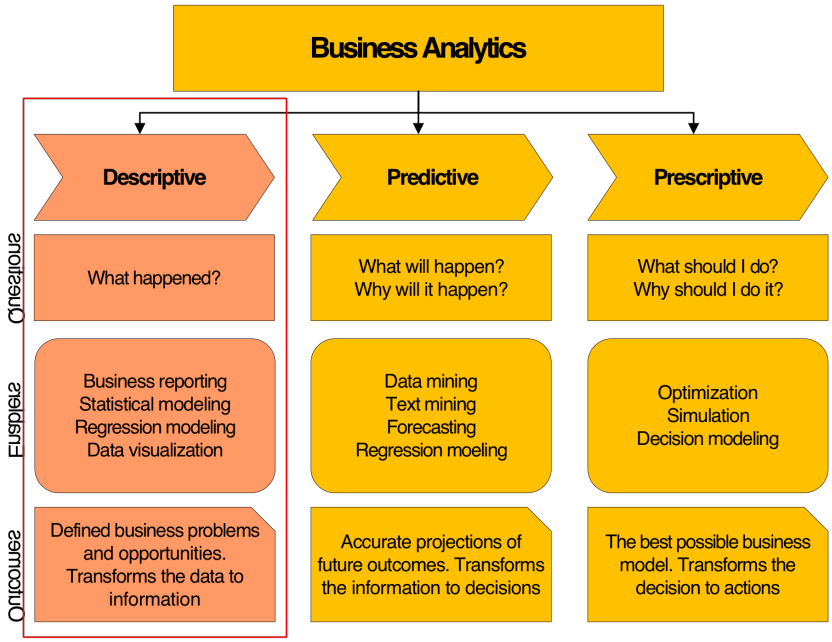


Figure 1. Types of analytics, and scope of the current manuscript (orange boxes).

In the figure, the orange boxes highlight the BA analyses used for the subsequent analyses of the incidents studied in the MHIDAS database, which provides a description of the incidents reports data by means of the business reports.

3.2. Business Intelligence

There is significant value embedded within the collective of sets of information available for safety analysts, and waiting to be discovered and exploited. However, to access this hidden treasure,

analysts must first adjust their way of thinking about data, information, and ultimately, actionable knowledge (Loshin, 2013). In this context, Business Intelligence (BI) can be a solution. BI is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies (Chaudhuri et al., 2011). Historically, data is the raw material that fuels operational activities and transaction systems. Therefore, limiting the use of those data sets to their original purposes is now an obsolete approach. Today, and for the foreseeable future, data utility is expanding to support operational activities, as well as tactical and strategic decisions (Chen et al., 2012). The process of BI is based on the transformation of data to information, then to decisions, and finally to actions.

Descriptive analytics

To reach a robust BI analysis, it is firstly necessary to know what is happening in an organization and to understand underlying trends and behaviors. The first step of this process involves the consolidation of data sources and the availability of all relevant data in a form that enables appropriate reporting and analysis. From this data infrastructure, it is then possible to develop appropriate queries, reports, and alerts using various reporting tools and techniques.

The nature of data

Data is the main ingredient for any BI, data science, and business analytics initiative. It can be defined as the raw material used from decision technologies to produce information, insight, and knowledge. Although data were once perceived as a big challenge to collect, store, and manage, it is now widely considered as one of the most valuable assets of an organization, with the potential to create invaluable insight to better understand customers, competitors, and business processes (Al-Aqrabi et al., 2015; Watson and Wixom, 2007).

Before proceeding with the safety analysis of the MHIDAS database, it is therefore necessary to explore data in a structured way. At the highest level of abstraction, data can be classified as structured and unstructured (or semi-structured). Unstructured/semi-structured data consists of any combination of text, images, voice, and Web content. These data will be discussed in more detail in the chapter on text mining and web mining. Structured data refers to data that uses data mining algorithms and can be classified as either categorical or numeric. Figure 2 shows the representation of the data taxonomy for the database MHIDAS.

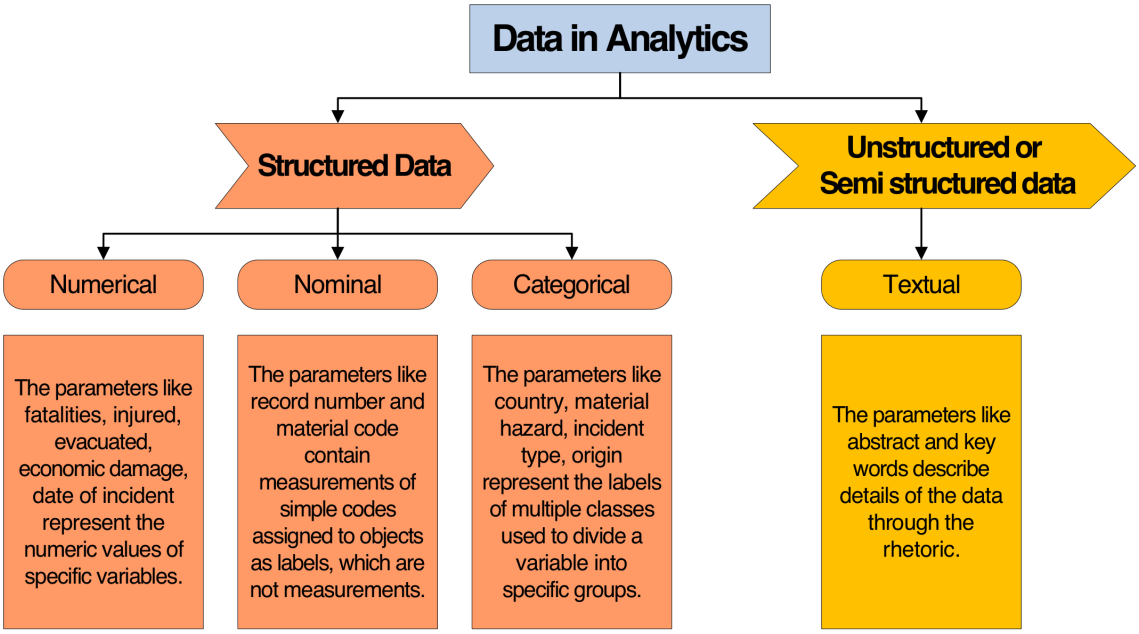


Figure 2. Taxonomy of relevant data types for the MHIDAS database

Data Warehousing

Data warehouse (DW) is a subject-oriented, integrated, nonvolatile, time-variant collection of data supporting management decisions. In other words, a DW is a pool of data produced to support the BI process. Data is usually structured to be readily available in the right form for analytical processing activities (online analytical processing, data mining, querying, reporting, and other decision support applications). DWs provide access to data for complex analysis, knowledge discovery, and decision making. They support high-performance demands on an organization's data and information (Elmasri and Navathe, 2011; Inmon, 2005).

MHIDAS reflects some feature of a DW (Inmon, 2014):

- **Subject oriented:** Data are organized by detailed subject, which remains relevant for safety analysts.
- **Integrated:** Integration is closely related to subject orientation. MHIDAS is an integrated DW because incorporated the information of many contributors.
- **Time variant (time series):** The warehouse supports historical data.
- **Nonvolatile:** Users cannot change or update the data.
- **Client/server:** The DW uses the client/server architecture to provide easy access for end users. Namely, MHIDAS provided information on Occupational Safety and Health on a CD-ROM (OSH-ROM), which back in the 1990s represented a state-of-the-art solution.
- **Include metadata:** The DW contains metadata about data organization and information on how to effectively use those data.

Whereas a DW is a repository of data, data warehousing includes the entire process. Data warehousing is a discipline that results in applications providing decision support capability, allowing ready access to business information, and creating business insight.

3.2.1.Extraction-Transformation-Loading process

The Extraction-Transformation-Loading (ETL) process allows to obtain quality data from DW information, hence facilitating the decision-making process. According to (Wang and Strong, 1996), quality dimensions are organized into four categories, namely (Souibgui et al., 2019):

- **Intrinsic** quality dimensions, which include accuracy, reputation, believability, and provenance. They rely on internal characteristics of the data during evaluation.
- **Contextual** quality is more information than data oriented, since it refers to attributes that are dependent to the context in which data is produced or used. It comprises: amount of data, relevance, completeness, and timeliness quality dimensions.
- **Representational** quality is related to the way data is perceived by its users and it relies on understandability, consistency, and conciseness quality dimensions.
- **Accessibility** allows measuring the ease of access to data and it covers accessibility and security dimensions.

Many reasons stand behind the need of a data integration phase within the decision system: heterogeneous formats; ambiguous or difficultly interpretable data formats; obsolete databases used by legacy systems; and ever-changing data source's structure. These characteristics of data sources make data quality uncertain. Several studies have been conducted with the purpose of identifying different quality issues within the data integration process (Souibgui et al., 2019). Most of these studies agree that data quality faces different challenges. Indeed, ETL is a crucial part in the data warehousing process where most of the data cleansing and curation are carried out. However, the ETL process is not a one-time event. In fact, since data sources change over time, the data warehouse have to be periodically updated. Moreover, the constant change of business implies the need to change the DW system in order to maintain its value as a tool for decision makers. As a result, the ETL changes and evolves as well, and it should be therefore designed for ease of modification. A solid, well-designed, and documented ETL system is the foundation of a successful data warehouse project (El-Sappagh et al., 2011).

A well designed ETL process extracts data from data sources and enforces data quality standards to allow developers and end-users to use the extracted data for applications and to make strategic decisions, respectively (Kimball and Caserta, 2011). In other words, the data is extracted from the source systems and it undergo a sequence of transformations before being loaded into the DW. The repository of the systems containing the sources of data for a DW can vary from spreadsheets to mainframe systems. The complex transformations are usually implemented in procedural routines. The design of an ETL process generally consists of six tasks (Trujillo et al., 2003):

1. Selecting the sources for extraction: the data sources (usually several different and heterogeneous data sources) to be used in the ETL process are defined.

2. Transforming the sources: once the data have been extracted from the data sources, they can be transformed, or new data can be derived. Some of the common tasks of this step are: filtering data, converting codes, calculating derived values, changing the different data formats, generating automatically sequences of numbers (surrogate keys), etc.
3. Joining the sources: different sources can be joined in order to load together all the data into a unique target.
4. Selecting the target to load: the target (or targets) to be loaded is selected.
5. Mapping source attributes to target attributes: the attributes (fields) to be extracted from the data sources are mapped to the corresponding target attributes.
6. Loading the data: the target is populated with the transformed data.

3.2.2.Constructing the BI model

After providing access to the DW, and processing data through an ETL process, additional operations might be conducted, e.g.,filtering, joining, and aggregation, to create a model where BA is applicable (Chaudhuri et al., 2011). Implementing a BI system requires careful planning to assure that the system meets the users' expectations. This process usually includes the following basic steps (Oracle, 2004):

I. **Identifying End-User requirements:** For this purpose, end-users are identified as safety and risk management researchers who will analyze the data. The questions that the BI system needs to answer can be summarized as: What information is currently available? What additional information is required? How should the information be presented? These answers must refer to MHIDAS incidents and should be explored both at the individual and aggregated level.

II. **Identifying the Data Source:** MHIDAS database is available in a .txt file on the Occupational Safety and Health CD-ROM (OSH-ROM), which was built in the early 2000s and historically distributed and developed by the National Institute for Occupational Safety and Health (NIOSH), the UK HSE, and the UK AEA.

III. **Designing the data model**: The data model was developed through an ad hoc ETL process (fully detailed in the 'Extraction-Transformation-Loading process' section), involving the following sub-steps:

a. Creating the Data Store: data are extracted from the source and imported into the software to create the data warehouse included in the workspace. Different data sources can be connected through specific data relationships (e.g., One to one, One to many, Many to one, Many to many).

b. Generating the Summary Data: Some of the data (ideally the most frequently queried data) is summarized and stored following a data maintenance procedure. This is done by creating and managing the required queries with a specific programming language.

c. Preparing the data for client access and granting access to end-users: Users should have database access rights in order to view and manipulate the data. After the datastore is ready for client access, it is possible to distribute the software and provide documentation to the end-users, where needed. For the sake of simplicity, the client's tools are not present in this contribution but they are freely accessible at the following link: http://bit.ly/MHIDAS2021

IV. **Creating and Distributing Reports:** It is possible to develop reports. The reports created for this research show the parameters employed to describe what occurred in the incidents and share them with the subject community, i.e., safety analysts.

The data obtained from the MHIDAS database was managed as a snowflake/hub and spoke BI model to deal with more than 9000 industrial incident reports worldwide. Each reported event in MHIDAS had maximum 21 parameters (either textual, categorical, or numerical) used to describe the respective event in a structured and systematic way. Following the data model development, a set of BI/ML solutions was proposed in dedicated reports.

3.3. Data clustering through Machine Learning

Clustering is an important unsupervised learning task aimed at investigating a collection of items by grouping them into subsets (clusters). For instance, items within a same cluster are more closely related (similar) to each other than items contained in different clusters (Kingrani et al., 2017; Mur et al., 2016).
Before delving into pragmatical clustering applications, it is important to describe some debugging processes, often referred to as outlier identification, i.e., an approach used to separate isolated points from the more representative one.

3.3.1.Outliers

The outlier detection algorithms are supervised learning methods and they are particularly effective for applications in which label information is either hard to obtain or unreliable. In fact, outlier detection identifies data points that are different from the remaining data, i.e., the algorithm identifies anomalies in the dataset (Aggarwal, 2017; Kriegel et al., 2011). Most of the approaches to the problem of outlier detection identified in the existing literature are based on density estimation methods or on nearest neighbour methods (Abe et al., 2006; Bouguessa, 2015; Breunig et al., 2000).

An important aspect of an outlier detection technique is the nature of the desired outlier, i.e., Point Outliers, Contextual Outliers and Collective Outliers (Chandola et al., 2009). This research focused on Collective Outliers in which it was possible to separate the most extreme points of the dataset. Collective anomalies have been explored in literature for sequence data (Chawla and Sun, 2006; Warrender et al., 1999), graph data, and spatial data (Shekhar et al., 2001). It should be noted that while point anomalies can occur in any dataset, collective anomalies can occur only in datasets containing related data instances (Lazarevic and Kumar, 2005). The techniques used to detect collective anomalies are different than those used for point and contextual anomaly detection, and they require a separate detailed discussion (Singh and Upadhyaya, 2012).

Isolation Forest

Isolation Forest (iForest) (Liu et al., 2008) is an advantageous outlier identification algorithm since it does not rely on building a profile for data in order to find non-conforming samples (Hariri et al., 2019). In an Isolation Forest algorithm, data is sub-sampled and processed in a tree structure based on random cuts in the values of randomly selected features in the dataset (Ding and Fei, 2013; Ramaswamy et al., 2000). iForest is a method inspired by Random Forest (Menze et al., 2011). It was proven that iForest can outperform current state-of-the-art outlier detection approaches in several applications, relying on a mechanism called isolation (Susto et al., 2017): a procedure that through iterative partitioning of the input space aims to separate a new observation

from the rest of the data at hand. The iForest algorithm is therefore applied in this research since it is very accurate and it also has several advantages compared to other methodologies (Ding and Fei, 2013; Hariri et al., 2019):

- It does not require a model to describe the input output relationship of the monitored process.
- It is computationally efficient with respect to common density or distance-based monitoring approaches.
- Low memory requirement.
- Natural parallel computing implementation.
- It identifies the anomalies by isolating outliers in the data.

The iForrest procedure is represented by an ensemble of t binary trees (random partition). The anomalies produce mean paths (from root to leaves) which are longer normal attributes. Given a dataset $X = \{x_1, \ldots, x_n\}, x \in \mathbb{R}^p$, each Isolation Tree (iT) is obtained by selecting a random subset $X' \subset X (\psi = |X'|)$ of attributes and by dividing X' by randomly selecting a feature q and a split value q until the node has only one instance, where ψ is the given dataset. This characteristic of iTs enables to implement subsampling, thus making this model capable of scaling up so as to handle extremely large sets of data and high dimensional problems. Furthermore, the iForest defines an Anomaly Score (AS) s, i.e., a quantitative index that defines the "outlierness" degree of an observations The AS is defined for an observation x by (Ding and Fei, 2013; Xu et al., 2017):

$$s(x, \psi) = 2^{\left(-\frac{E(h(x))}{c(\psi)}\right)} \in [0,1] \quad (1)$$

$E(h(x))$ is the average of $h(x)$ over the t.

$c(\psi) = E(h(x)|\psi)$ is an adjustment factor that considers the cardinality of the subsampled dataset.

3.3.2.Hierarchical clustering

This paper employs hierarchical clustering, which is currently adopted in various settings (Bouguettaya et al., 2015; Murtagh and Contreras, 2012). In fact, instead of building cluster hierarchies based on raw data points, the developed hierarchical clustering establishes a hierarchy based on a group of centroids. Such hierarchical algorithms can be easily divided into two groups of methods. The first group contains linkage methods, which were employed to obtain the results presented in this paper. The second group includes hierarchical clustering methods that ensure the specification of cluster centers (as an average or a weighted average of the member vectors of the cluster) (Murtagh and Contreras, 2012).

Distance

In clustering algorithms, distance metrics play an important role since they are the basic element to compute the similarity between two objects in a specific domain. All of the supervised and unsupervised algorithms use distance metrics to understand the patterns in the input data (Aggarwal et al., 2001; Reddy et al., 2018). In general, distance metrics employ functions describing the distance between sets of elements in the dataset. Several distance metrics can be used in a clustering algorithm, and they should be therefore carefully chosen to avoid introducing

errors or misinterpretations. The most common metric is the Euclidean distance, as it flexible and usable in different contexts. Moreover, this metric is significant for hierarchical clustering, since it reflects a low-dimensional space (Milligan and Cooper, 1985), and it is therefore applied in the present research.

The Euclidean distance is a basic distance metric used to identify the first and second nearest neighbor for each of the obtained descriptors. If the distance is small, the elements are presumably similar, and vice versa. The formula for the calculation of the Euclidean distance is defined by (Burkov, 2019; Reddy et al., 2018):

$$d(a,b) = \|a-b\| = \sqrt{(a-b)'(a-b)} = \sqrt{\sum_{i=1}^{\psi}(a_i - b_i)} \tag{2}$$

where a, b are the two different descriptors vectors, and ψ is the dataset given.

Ward's minimum variance method

In the following analyses, the Ward's minimum variance method was applied, as it tends to create compact, uniformly sized clusters. Although Ward's method is much less computationally intensive than other methods, it is still appropriate for most purposes (Bouguettaya et al., 2015).

Ward's method can be defined and implemented recursively by means of the Lance–Williams algorithm. Therefore, if points i and j are agglomerated into cluster $i \cup j$, it is sufficient to specify the new dissimilarity between the cluster and all other points (clusters). The formula explaining this procedure is defined by (Murtagh and Contreras, 2012):

$$d(i \cup j, k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)| \tag{3}$$

Where $\alpha_i, \alpha_j, \beta$ and γ define the agglomerative criterion

The cluster size and distance function refer to the cluster algorithm's definition. For that purpose, the Ward's Minimum Variance method from the Lance-Williams formula has been implemented:

$$\alpha_i = \frac{|i| + |k|}{|i| + |j| + |k|} \tag{4}$$

$$\alpha_j = \frac{|j| + |k|}{|i| + |j| + |k|} \tag{5}$$

$$\beta = \frac{-|k|}{|i| + |j| + |k|} \tag{6}$$

$$\gamma = 0 \tag{7}$$

Where | . | is the absolute value of the points.

Other update formulas exist that allow the implementation of agglomerative methods, e.g., complete link, single link or median method. The Euclidean distance should be used for equivalence between the approaches. For example, having a and b be two points (m-dimensional vectors: objects or cluster centers) which have been agglomerated, and c another point, it is possible to use squared Euclidean distances to update Lance–Williams dissimilarity formula.

$$d^2(a \cup b, c) = \frac{d^2(a,c)}{2} + \frac{d^2(b,c)}{2} - \frac{d^2(a,b)}{2} = \frac{\|a-c\|^2}{2} + \frac{\|b-c\|^2}{2} - \frac{\|a-b\|^2}{4} \quad (8)$$

It is possible to define the new cluster center $(a+b)/2$, thus obtaining a distance from point c

$$d(a \cup b, c) = \left\| c - \frac{a+b}{2} \right\|^2 \quad (9)$$

where $\| \, . \, \|$ is the norm in Euclidean metric.

4. Results

4.1. Data model

The Facts Table including all the parameters sharing a One-to-One relationship with the accident identification (ID in figure defined by Record) was crucial for the implementation of the model. In fact, in the branches, the data model presented several Many-to-One relationships, where the number of parameters was higher than the accident ID (hazardous, origin, causes, incident type, ignition source, specific causes, keywords, general causes, material code). Figure 3 presents a visual representation of the data model architecture realized using the Crow's Foot notation, in line with Table 1.

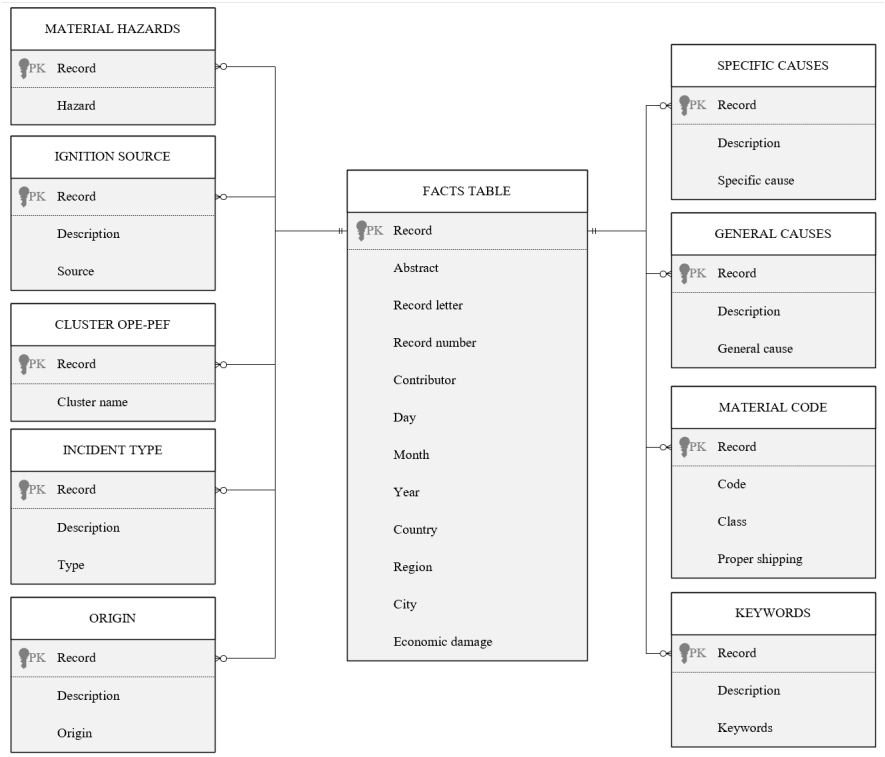


Figure 3. MHIDAS data model (PK = Primary Key)

4.2. General overview

Firstly, a general descriptive analysis of the incidents reported in the MHIDSA database was performed. Figure 4 shows a line chart presenting the annual occurrence of incidents identified through the above-mentioned analysis.

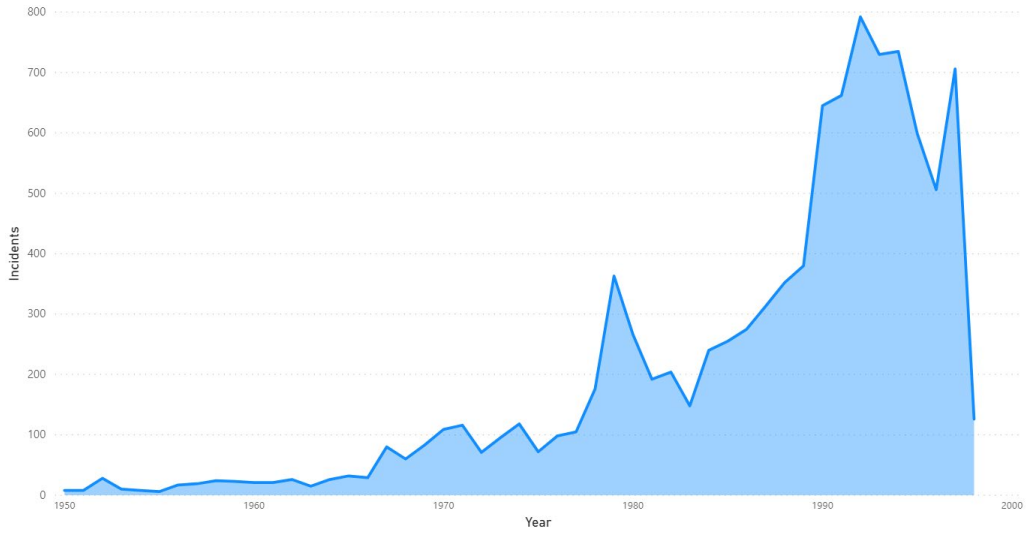


Figure 4. Development over the time of the collected incidents in the MHIDAS database

Furthermore, a geographical analysis was performed to complement the temporal analysis. The results are presented in Figure 5, where the color of the bubbles indicate the overall economic

damage resulting from the incidents, while the size of the bubble represents the total number of fatalities caused by the accidents. This map presents, albeit generally, some aggregated highly abstract result on occupational and operational hazardous events.

Figure 5. Worldwide distribution by Economic Damage [$] (bubbles' color) and number of fatalities (bubbles' size)

The temporal and geographical analyses were then followed by the analysis of the clustering of the incidents reported in the MHIDAS database. In fact, this analysis allows to investigate the hazardous events at more granular levels. Subsequently, occupational, operational, and aggregated data analyses were performed.

4.3. Occupational analyses

Firstly, a scatter plot was implemented to identify the incidents related to occupational issues: the number of fatalities was plotted on the x-axis and the number of injuries was plotted on the y-axis (cf. Figure 6). Furthermore, since the size of the bubble represents the number of incidents related by each type of hazard, this graph allowed to represent the occupational impact of different hazards in terms of fatalities and injuries. On this basis, it was possible to isolate the most serious hazards, i.e., Fire (50.426 injured and 13.349 fatalities), Toxic (58.712 injured and 3.173 fatalities) and Explosive (11.734 injured and 4.618 fatalities).

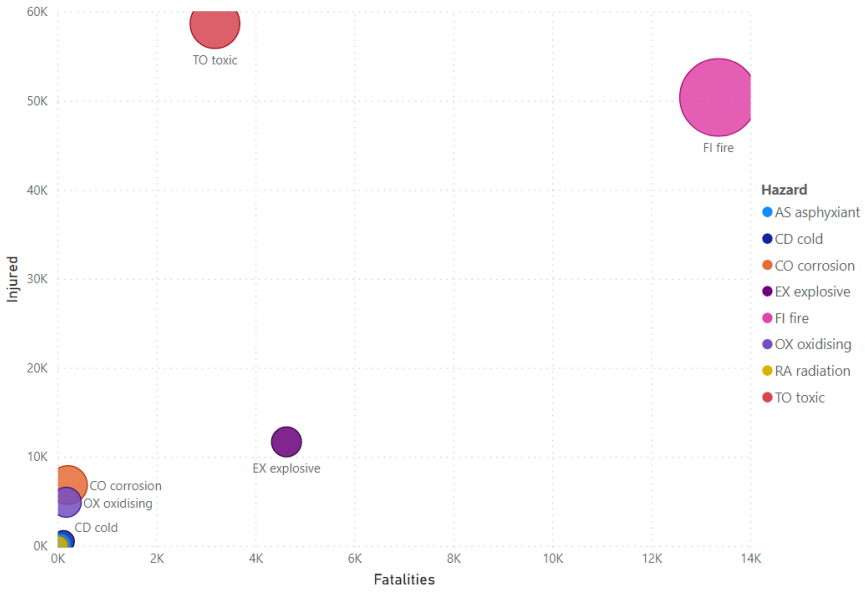


Figure 6. Classification of incidents by hazard.

Table 2 represents the data obtained in the scatter plot, sorted by the number of incidents related to each of the hazards found in the MHIDAS database.

Table 2. Classification by hazard of the incidents

Hazard	Number of Incident	Injured	Fatalities
FI, fire	5.801	50.426	13.349
TO, toxic	2.174	58.712	3.173
CO, corrosion	1.128	6.888	202
EX, explosive	539	11.734	4.618
OX, oxidizing	501	4.949	168
CD, cold	88	540	109
AS, asphyxiant	56	267	56
RA, radiation	6	30	1

By observing the presented table, it is possible to note that the ratio between the number of fatalities and the number of injured is not uniform across hazardous substance. In fact, the most serious hazards in terms of incident occurrence are Fire (5.801), Toxic (2.174) and Corrosion (1.128). This new prioritization shows that Corrosion-related events are more frequent than Explosion-related ones, but they have minor occupational consequences.

To further deepen this occupational-oriented analysis, Figure 7 presents a stacked bar chart where each bar represents the ratio percentage of fatalities and injured workers for relevant parameters. By observing Figure 7, it is possible to make some considerations: (i) the higher relative percentage of fatalities is connected to explosives (28,24%) and the lowest relative percentage of fatalities is connected to corrosion (2,85%) (cf. Figure 7A); (ii) External causes are associated with the highest relative percentage of fatalities (21,49%) and Instrumental causes are associated with the lowest relative percentage of fatalities (10,07%) (cf. Figure 7B).

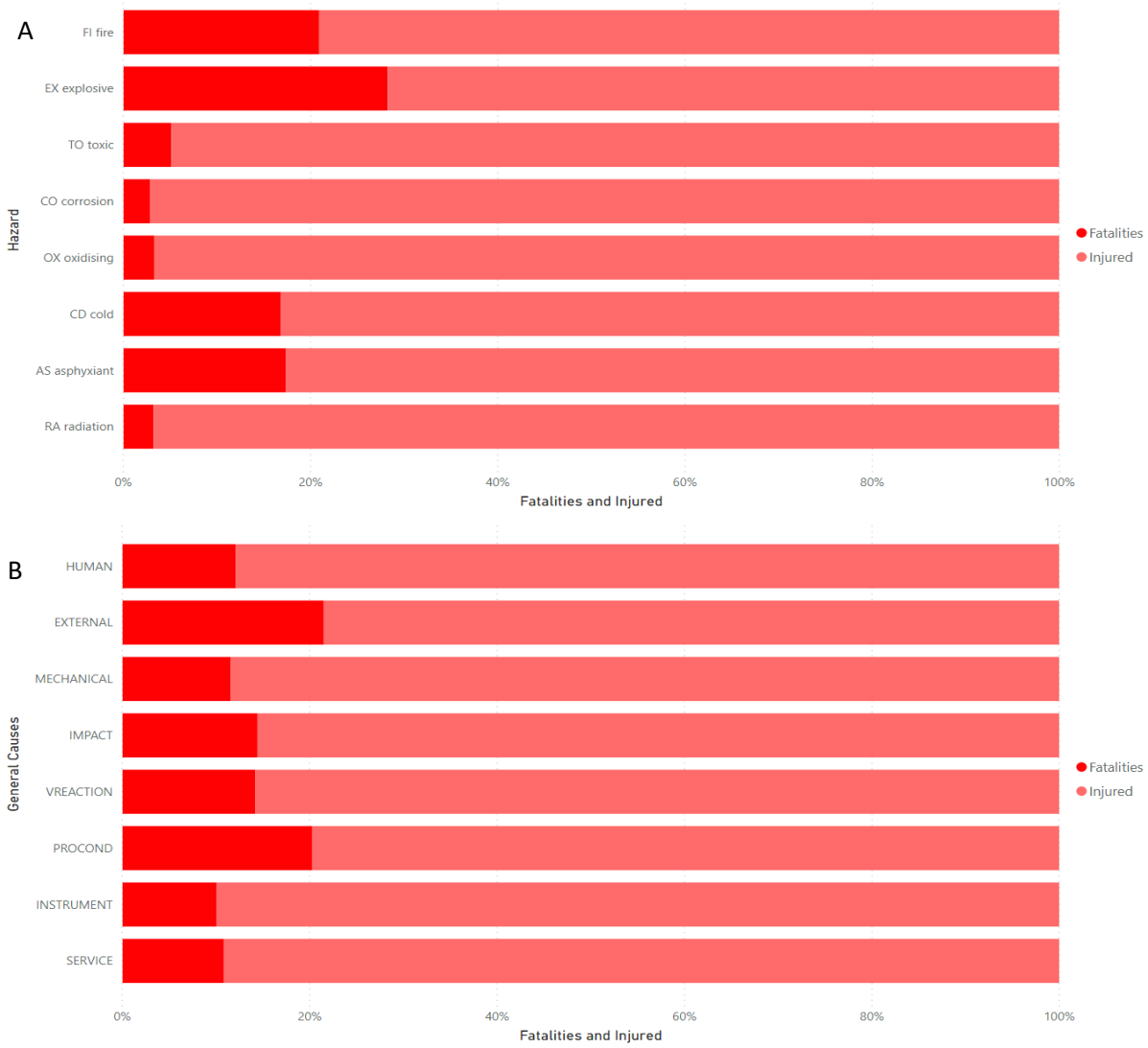


Figure 7. Proportion ratio by hazard and general causes of the fatalities and injured.

Furthermore, a classification of different materials was performed to emphasize their occupational impact on fatalities and injuries in the reported accidents. The results of this analysis are presented in Figure 8. The x-axis of the scatter plot shows the number of fatalities, the y-axis shows the number of injuries, and the size of the bubble represents the number of incidents related to a specific material. As a result of the previous observations on the relation between material and economic damage, it is possible to highlight that the most critical materials are Chlorine (12.039

injured and 131 fatalities), LPG (4.504 injured and 1.948 fatalities), and Crude oil (1.247 injured and 1.259 fatalities)

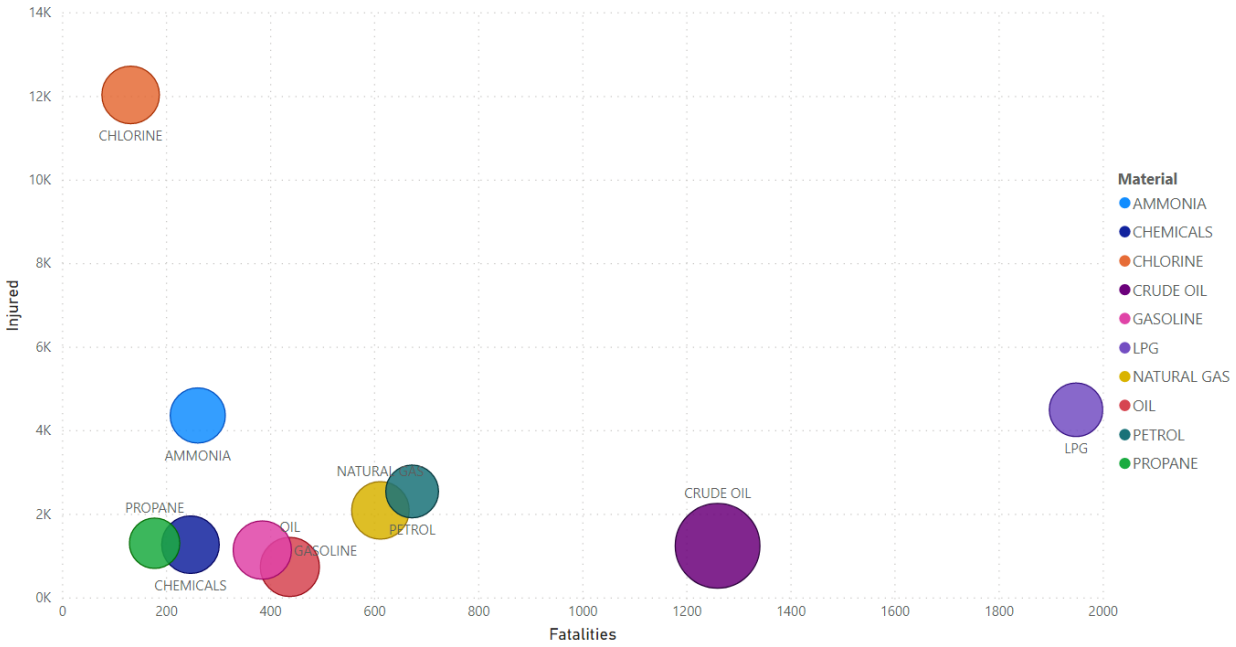


Figure 8. Classification the incidents by material.

Table 3 represents the data sorted by the number of incidents related to each material.

Table 3. Classification by material of the incidents

Material	Number of Incidents	Fatalities	Injured
CRUDE OIL	754	1.259	1.247
OIL	345	437	742
GASOLINE	326	384	1.146
CHLORINE	324	131	12.039
CHEMICALS	315	246	1.275
NATURAL GAS	315	611	2.096
AMMONIA	290	260	4.368
LPG	271	1.948	4.504
PETROL	262	672	2.549
PROPANE	235	177	1.309

4.4. Operational analyses

Within the MHIDAS database, 339 materials were related to industrial incidents. Subsequently, a Pareto analysis was performed to examine which of the identified materials had a greater economic impact. Figure 9, supplemented by Table 4, represents the materials with higher economic damage reported on MHIDAS. In this context, Crude oil represents the 46,59% of the total of economic damage and, in conjunction with Oil, Butane and Propane, it accounts for 69,08% of the total of economic damage.

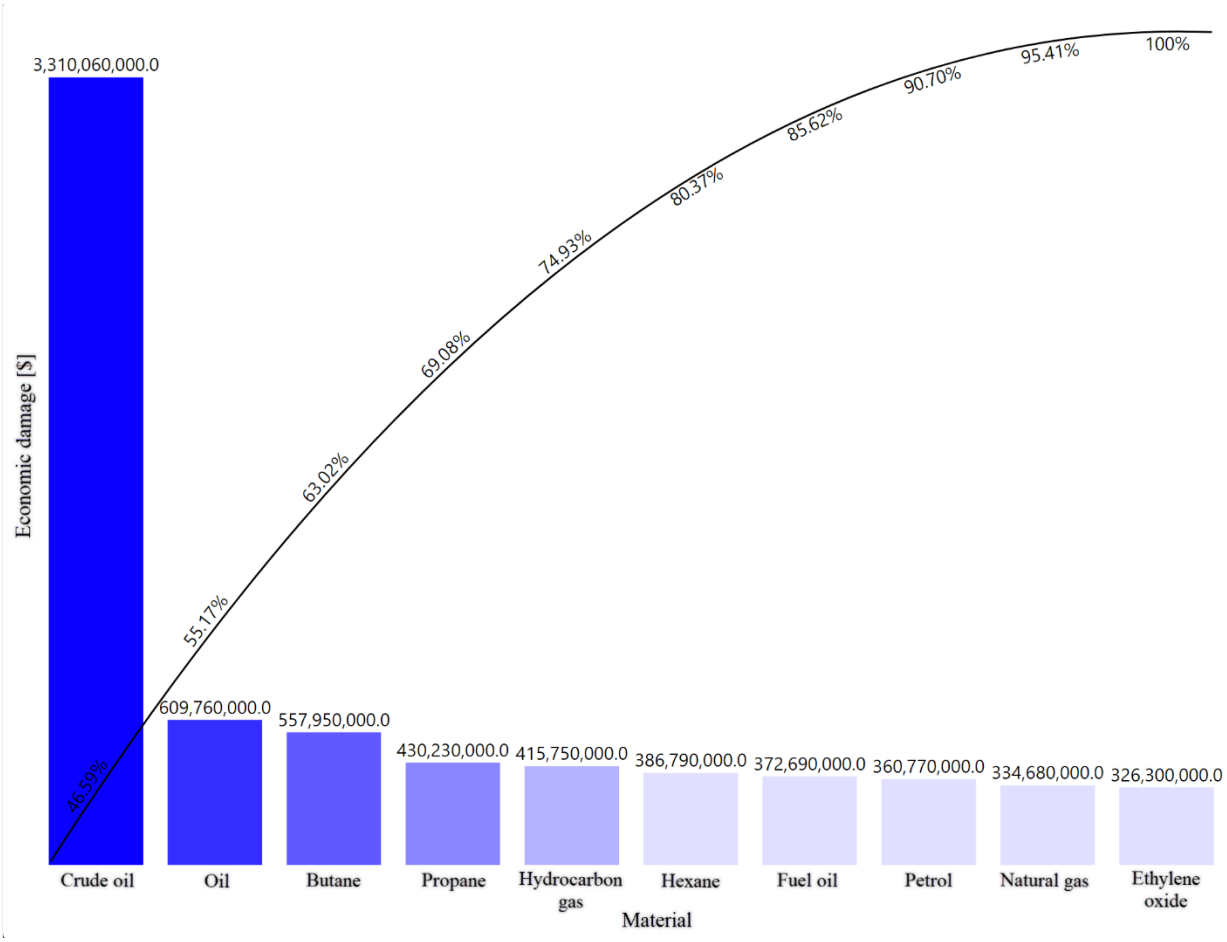


Figure 9. Pareto classification of materials sorted by economic damage [$]

Table 4. Classification of materials with higher economic damage

Material	Economic damage
CRUDE OIL	$3.310.060.000,00
OIL	$609.760.000,00
BUTANE	$557.950.000,00

PROPANE	$430.230.000,00
HYDROCARBON GAS	$415.750.000,00
HEXANE	$386.790.000,00
FUEL OIL	$372.690.000,00
PETROL	$360.770.000,00
NATURAL GAS	$334.680.000,00
ETHYLENE OXIDE	$326.300.000,00

Furthermore, it is possible to extract additional information about the economic damage with respect to materials. In fact, by isolating the first four materials identified by Pareto analysis (cf. Figure 9), it is possible to identify the amount of discharges for each of the identified materials: 2.572.253,15 Long Tons of Crude oil discharges (50,35% of all discharges, i.e., 5.108.287,10 Long Tons) posing Fire, Explosive and Toxic hazards; 297.150,69 Long Tons of Oil discharges (5,81% of the total releases) posing Fire and Toxic hazards;15.062,00 Long Tons of Butane discharges (0,29% of the total) posing Fire hazards; 32.756,71 Long Tons of Propane discharges (0,64% of the total) posing Corrosion, Fire and Explosive hazards. Subsequently, a further analysis was performed to prioritize critical materials based on their root origin (see Figure 10): the x-axis of the chart presents the economic damage [$], the y-axis presents the quantity of material [Long Tons], the size of the bubbles defines the number of incidents triggered by group of material, and the color of the bubbles represents the group of material (Top 5 materials classified with regard to the economic damage [$]). For demonstration purposes, such analysis was restricted to records involving a specific quantity of material (1000 Long Tons or less) and a specific economic damage (100.000$ or less).

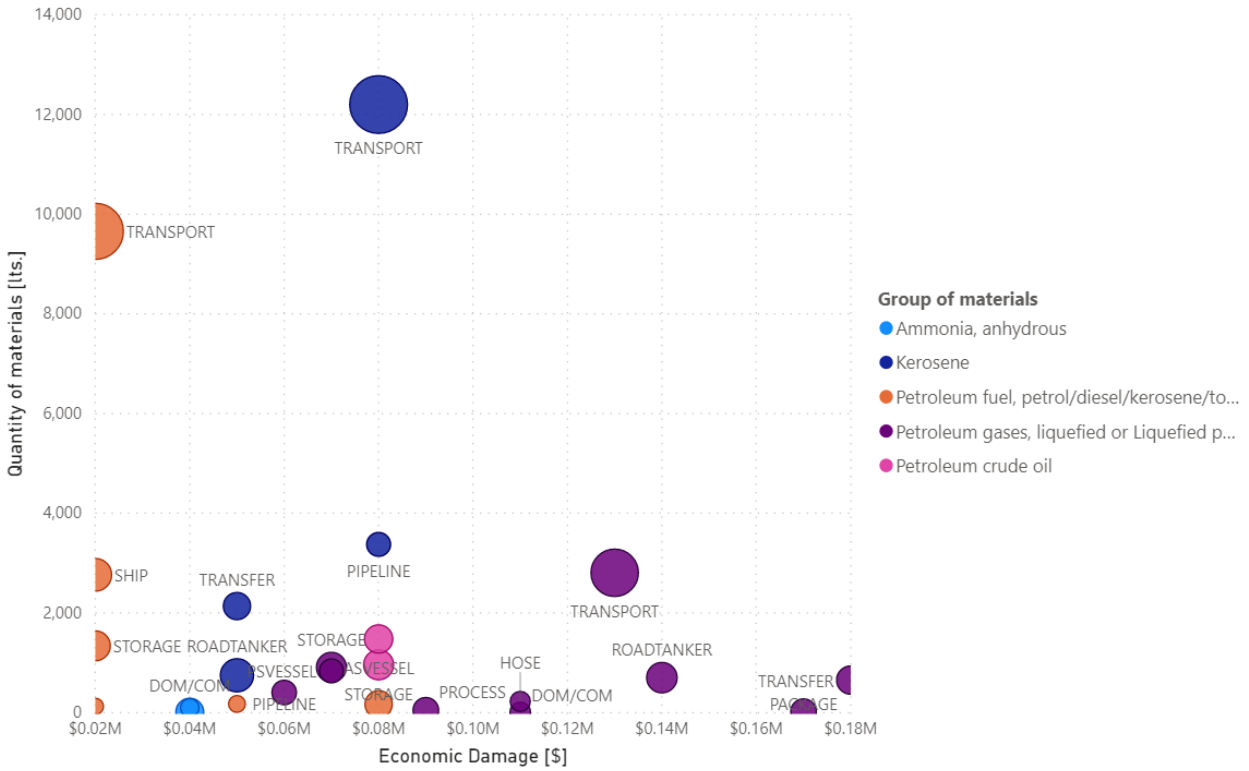


Figure 10.Classification the incidents by group material and origin of those

Although the two parameters, i.e., economic damage and quantity of material, do not have a linear relationship, it is possible to observe that many incidents involve minor material discharges (i.e., Petroleum Gases or Petroleum Crude Oil).

4.5. Integrated Occupational/Operational analyses

To perform the integrated occupational and operational analyses, a metric parameter called Potential Equivalent Fatality (PEF) was used to integrate the computations regarding fatalities and injured workers. A complete description of PEF is outlined in (Edwin et al., 2016; Paltrinieri and Khan, 2016). However, for the purpose of the analysis, the PEF number was here modified according to the data available in MHIDAS, since the database does not include exposure data or death ratio due to hazards, nor information on different severity levels related to injuries. For the sake of simplicity, all the injuries found in MHIDAS were considered as major injuries. The resulting simplified formulation was:

$$PEF = Ft + 0{,}1 * Ij \tag{10}$$

$Ft = Fatalitiesreportedintheincident$

$Ij = Injuredpeoplereportedintheincident$

Moreover, a standardization algorithm was implemented within the economic damage parameter because of the significant differences between their ranges, which caused trouble computing a good cluster in the dataset. Furthermore, the z-scores algorithm was used. This paradigm allowed

to rescale the parameter values so as to provide them with the properties of a standard normal distribution, with $\mu = 0$and $\sigma = 1$, where μ is the sample mean (the average value of the feature, averaged over all the examples in the training data) and σ is the standard deviation from the sample mean.

Z-scores of the features were calculated as follows:

$$\hat{x}^{(j)} \leftarrow \frac{x^{(j)} - \mu^{(j)}}{\sigma^{(j)}} \tag{11}$$

where $\mu^{(j)}$ is the sample mean of the values of feature j, and $\sigma^{(j)}$ is the standard deviation of the values of feature j from the sample mean.

After defining the metrics used for the comprehensive analysis, the clustering algorithm was developed to link Economic Damage [$] and PEF, respectively, in order to find potential classification strategies. Following the theoretical description proposed in the 'Outliers' section, an outlier algorithm was implemented to isolate the points called disasters.

The results of the analysis performed with the iForest algorithm suggested that incidents with a PEF greater than 100 or an economic damage greater than $50.000.000,00 should be considered outliers, and thus studied separately. Figure 11 summarizes the results of iForest, where a 1% contamination parameter was set to isolate the previously mentioned incidents. After the application of this algorithm, the number of accidents identified in the resulting dataset decreased from 9993 to 9893. In fact, the outliers isolated 100 incidents that where included in the 1% contamination parameter established. These 100 incidents represent extreme events, namely disasters in either occupational or operational terms (or both). Some notable examples are:

- Record 6017: economic damage of $2.000.000.000 and 0 PEF in a rural area. Disaster caused by Human and Impact factor caused by Fire hazard combined with Crude Oil releases. Exxon Valdez company located in Valdez, Alaska (USA).
- Record 3454: economic damage of $370.000.000 and 7 PEF in an unknown population density area. Disaster caused by Fire hazard and hydrocarbon gas releases. Plant located in La Mede (France).
- Record 9098: 3000 PEF in an urban area. Disaster involving Fire and Toxic releases, caused by human components and reactions with the hydrocarbon gas involved. Plant located in Bhopal, Madhya Pradesh (India).
- Record 7536: 1300 PEF in an urban area. Disaster involving Explosive hazard with dynamite. Plant located in Cali (Colombia).

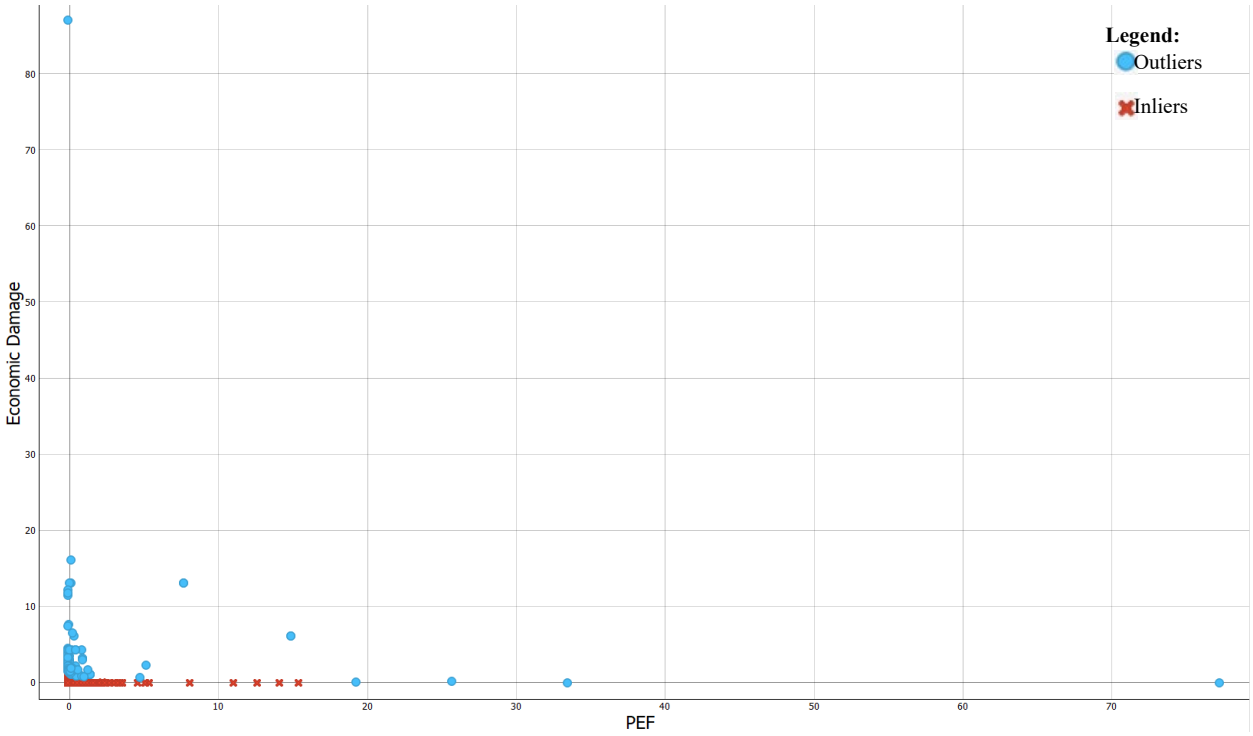


Figure 11. Differentiation of the incidents and disasters in the MHIDAS database.

Furthermore, a hierarchical cluster algorithm was applied with Ward's method (8) for centroids based on Euclidean distance (2). Subsequently, combinations were defined and sorted according to their silhouette scores, following the methodology described in the 'Constructing the BI model' section. Figure 12 shows the results of the silhouette values that can be used to determine the optimal number of clusters by selecting the combination with the highest average silhouette, i.e., cluster 4 (cf. Figure 13).

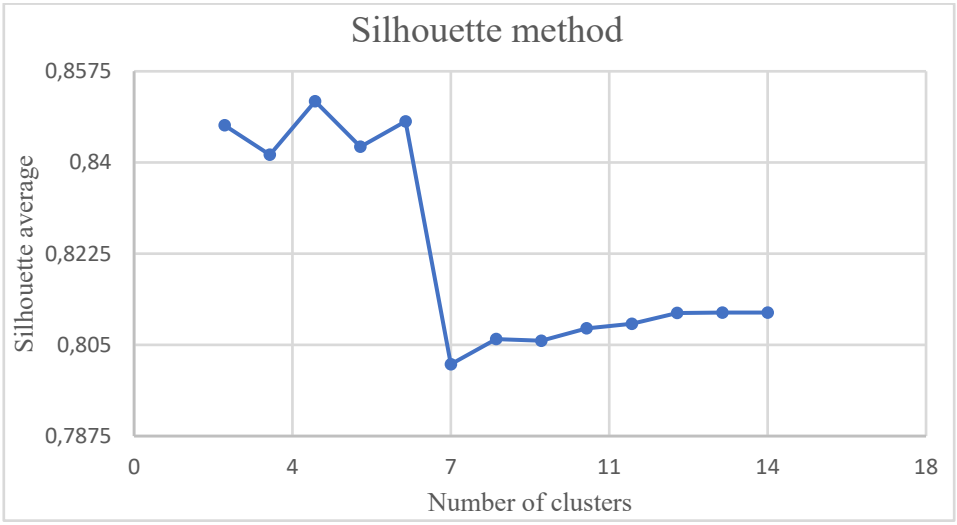


Figure 12. Silhouette values to determine the optimal number of the cluster (4).

After determining the optimal number of clusters and verifying that the silhouettes of each group were feasible (average silhouette scores C1:0,93; C2:0,016; C3:0,67; C4:0,19), the four obtained clusters were reported in a two-dimensional occupational/operational graph. The cluster is thoroughly described in Figure 13 and Table 5.

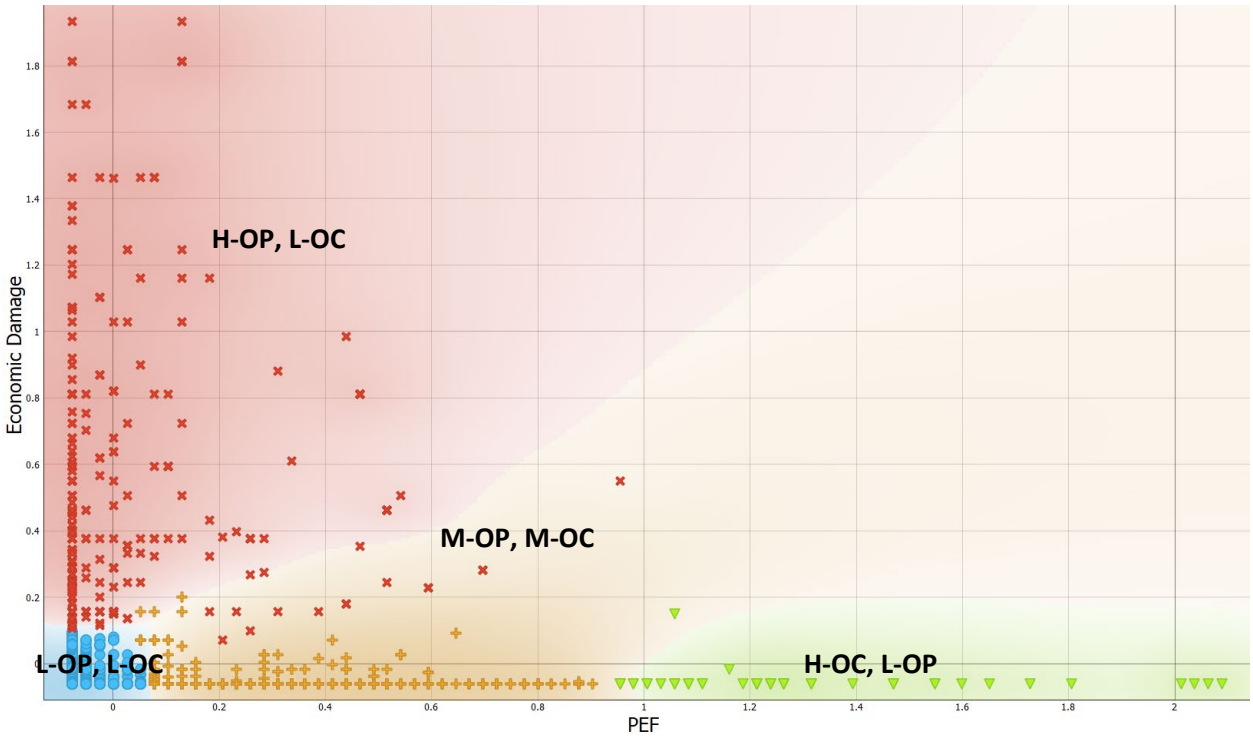


Figure 13. Cluster with the four groups selected for the analysis.

Table 5. Description of the group of clusters

Color	Name	Description
Red	H-OP, L-OC	High operational, Low occupational
Green	H-OC, L-OP	High occupational, Low operational
Orange	M-OP, M-OC	Medium operational, Medium occupational
Blue	L-OP, L-OC	Low operational, Low occupational

The red zone represents incidents with high operational damage and low occupational damage, i.e., incident with a high economic damage and low PEF number. On the other hand, the green zone represents a high PEF number (occupational damage) and low economic damage (operational damage). Finally, the orange and blue zones represent incidents related with both issues, i.e., incidents with different values of occupational and operational damage.

This analysis allows to make further observations. Namely, Figure 14 shows a timeline illustrating the proportion ratio of the type of incident and it specifies the name of the clusters for each year. By observing this figure, it is possible to notice a decrease of with H-OP, L-OC and M-OP, M-OC incidents over the time. In fact, in the 1960s the proportion ratio of the above-mentioned incidents was the highest: M-OP, M-OC incidents had a peak in 1961. i.e., 7 incidents, accounting for the 33,33% of the total amount of incidents (21 incidents), while the peak of H-OP, L-OC incidents was recorded in 1963, when the 3 incidents accounted for the 20% of the total number of accidents (over the 15). Contrastingly, L-OP, L-OC incidents increased over the time. In fact, in 1951 and 1954 only 8 L-OP, L-OC 8 incidents were recorded, while in 1998 the number of L-OP, L-OC incidents was 122 incidents, i.e., 97,6% of the 125 incidents.

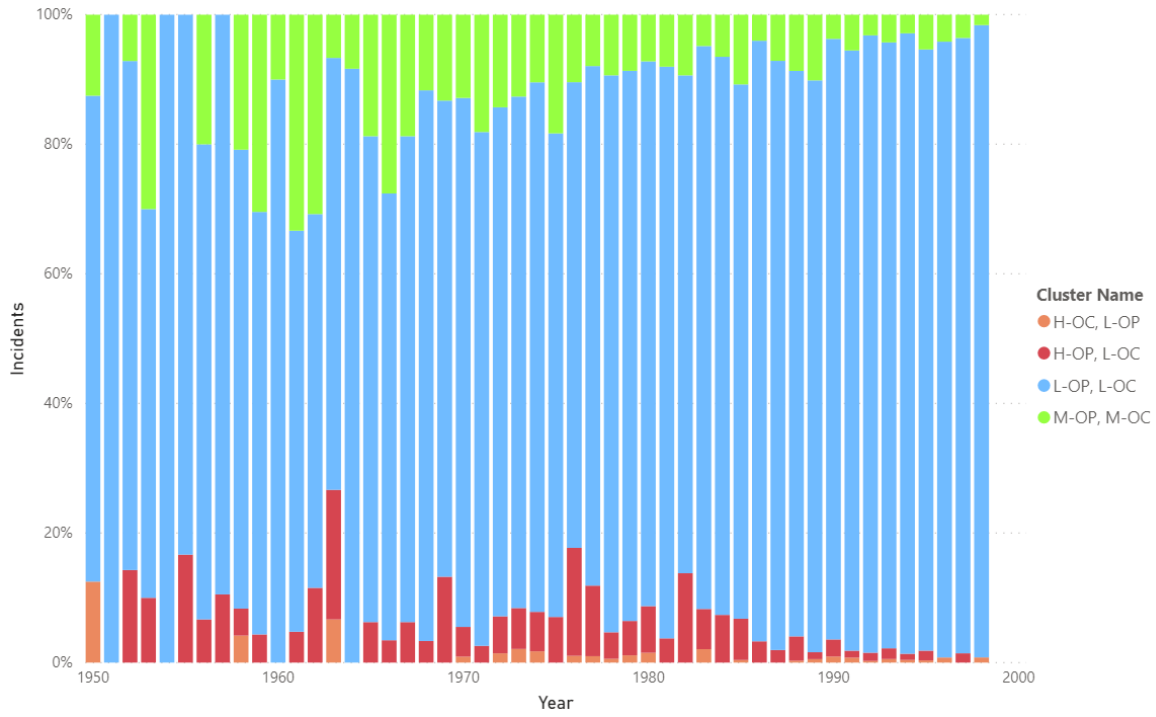
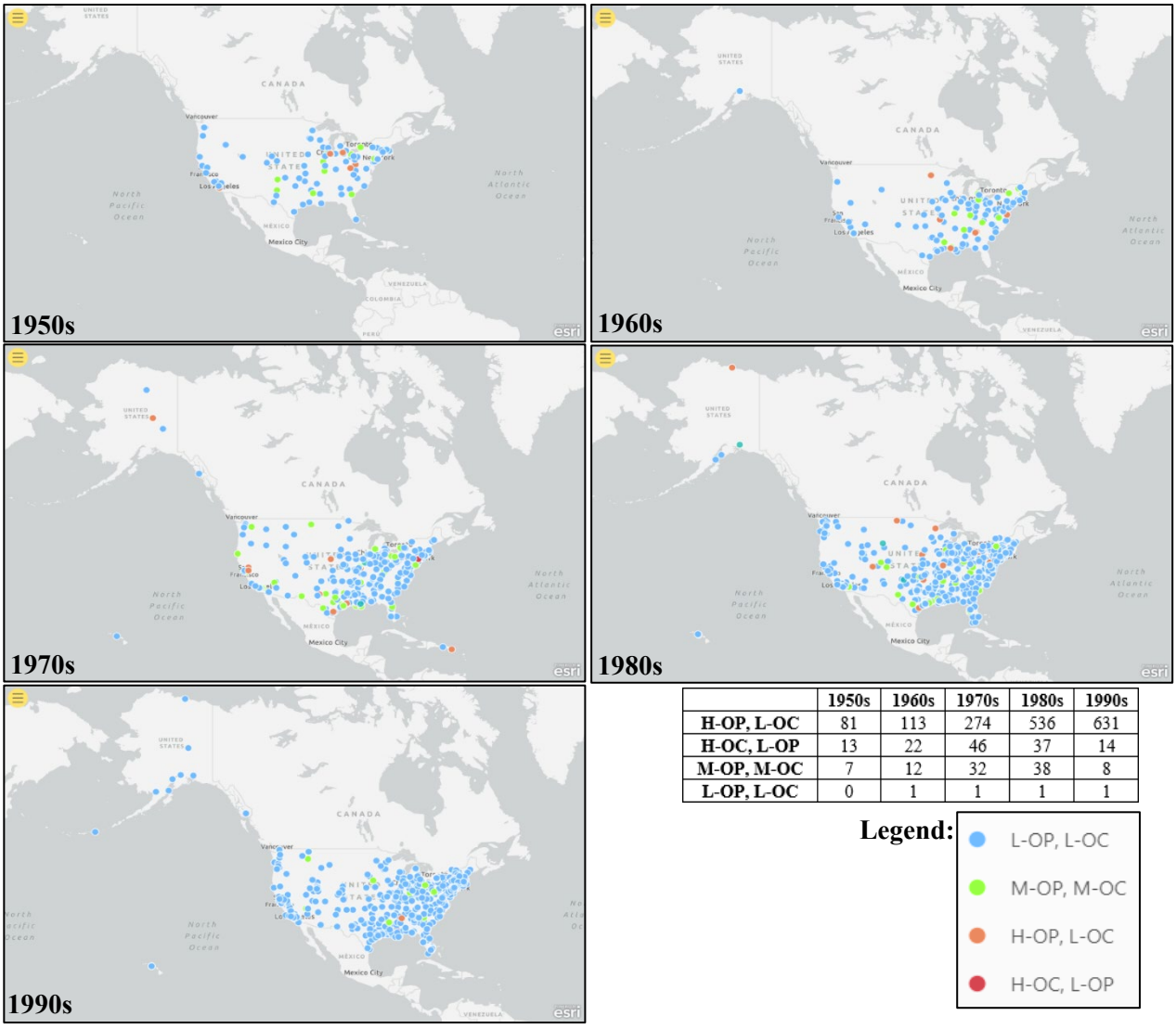


Figure 14. Behavior over the time by the incidents

The results of the clustering can be further extended to focus on a specific country, in the attempt to understand the evolution over time of incident criticalities and related reporting actions. For example, Figure 15 proposes an analysis of the United State of America (USA) over 5 decades (1950s-1990s). This analysis illustrates the higher number of L-OP, L-OC incidents over time. In the 1950s, the proportion ratio of L-OP, L-OC was the 80.19% , in the1960s the L-OP, L-OC incidents accounted for 76.35% of the total (113 incidents), in the 1970s the percentage of L-OP, L-OC incidents was 77.62% (274 incident), in the 1980s 536 L-OP, L-OC incidents were reported, i.e., 87.58% of the total, and in the 1990s the proportion ratio of L-OP, L-OC incidents was 96.48% (631 incidents). The remaining types of incidents do not have a particular behavior or evident trend to highlight.



	1950s	1960s	1970s	1980s	1990s
H-OP, L-OC	81	113	274	536	631
H-OC, L-OP	13	22	46	37	14
M-OP, M-OC	7	12	32	38	8
L-OP, L-OC	0	1	1	1	1

Figure 15. Incident occurrence in USA over the decades

5. Discussion

The development of a BI safety data model (as the one presented in Figure 3) can significantly contribute to manage safety data in large-scale industrial systems. BI safety models should in fact represent a data structure that allows a better documentation, so as to enhance dynamic analyses and reduced bias through faster and efficient querying systems supporting safety decision-making processes. The implementation of BI implies managerial implications for regulatory institutions and enterprises.

BI models might benefit from aggregated analysis to design safety guidelines and ensure their development over time following a data-driven perspective. An example of this data-intensive guideline design process can be identified in the analysis of the distribution specific accidents in diverse regions (e.g. Figure 5) which in turn supports the identification of main criticalities, and the detection of location-based patterns.

At enterprise level, BI implementation could spotlight relevant interpretations to improve the safety/risk management in industrial settings and support safety decision-making processes. For example, at occupational level, it is possible to detect underlying causal factors, as presented in Figure 6 and Table 2, where Fire were highlighted as the critical hazard involving the highest number of accidents and fatalities reported on MHIDAS. Furthermore, this analysis can be complemented with considerations on the materials involved, and/or the type of industry. Therefore, learning from other companies operating in similar conditions allows to identify and follow best practices in order to avoid recurring factors.
On other hand, these results can be interpreted in larger performance management contexts, by implementing Data Warehousing to collect multiple data sources, which can also be enterprise-based. For example, the results presented in Figure 15 can be complemented with metrics that extend the analysis into other areas, i.e., development index, economic metrics, demographic (by industry or population) density, seeking out a sustainable and safer resilient system. In this case, multiple models require robust ontological relationships to create a consistent data model (Mao et al., 2020).

From an analytical perspective, these positive results can be further discussed in terms of future research. On the one hand, it would be interesting to investigate the depth and numerosity of the parameters that should be studied and reported in case of industrial incidents. MHIDAS for example may be considered as an oversimplified reporting system in terms of human and organizational elements, thus requiring a more systemic approach in this field. This objective may be achieved by complementing the structure of the database with a logic relying on systemic methods, e.g., FRAM method (Patriarca et al., 2020), STAMP model (Leveson, 2011), and PRAF (Jain et al., 2018).

On the other hand, there are several additional algorithms that should be tested and validated. For instance, starting from the descriptive research dimension employed in this research, other predictive analyses can be developed to explain or forecast behaviors referred to specific parameters. Furthermore, text mining on incident narratives represents a promising area (Cheng et al., 2013), as well as the use of dedicated algorithms to understand patterns and support recommendations on the usage of specific materials under certain operating conditions (Anandarajan et al., 2019; Marshall and Wallace, 2019). These results are only the firsts step towards more complex IT applications for safety management, indicating a way forward for a wider industrial safety learning. In this regard, they also constitute the basis for other optimizations techniques, according to previous researches in this area (Paltrinieri et al., 2020).

The novel safety intelligence capacity strengthens cross-learning (Fruhen et al., 2014), i.e. BI is implemented to promote safety knowledge obtained from different companies, thus empowering intra- and inter-entities information sharing. We believe that the idea behind MHIDAS can be extended to other safety repositories, e.g., Analysis, Research and Information on Accidents (ARIA) or Major Accident Reporting System (eMARS). Moreover, a data-driven environment can also increase company awareness on industrial risk management, thus supporting safety culture measurements (Cooper Ph.D., 2000; Guldenmund, 2000).

6. Conclusion

The presented study provides a methodological discussion and a description of BI and ML techniques, which shows the potential for their large-scale adoption with regards to safety in industrial settings. Safety and loss prevention are multi-dimensional problems, involving a large set of variables that tightly interact to sustain the plant's functioning. When referring to reporting actions, those variables have to be studied jointly through informative data structures that permit to unleash the potential behind real data. BI was helpful in this regard, supporting a multi-variable dynamic analysis based on an underlying structured data model. BI represents an efficient way to provide answers, or even stimulate further questions relying on previous data. This structured analysis also allows a progressive enhancement of meta-knowledge to improve the quality of the investigations and data gathering. Although the case study was based on real events stored in the MHIDAS database and therefore outdated (1950s-1990s), it provided evidence on the significance of these results for modern incident reporting systems. Consequently, this analysis showed that with the help of BI tools it is possible to obtain a set of dashboards, thus providing a visual-descriptive analysis of extensive data information reported in a database and classifying/isolating relevant features (e.g., occupational, operational, mixed events). The developed analyses can provide useful information for different key users, supporting decision-making and cost-benefit analysis at different levels (Paltrinieri et al., 2012). We believe that these exemplary results may motivate organizational-wide adoptions of BI and ML within safety management systems, an area currently under-developed. The combination of BI with ML solutions provides a promising staging area for safety intelligence in industrial safety, thus paving the way for a bright research path in loss prevention.

Reference

Abdolhamidzadeh, B., Abbasi, T., Rashtchian, D., Abbasi, S.A., 2011. Domino effect in process-industry accidents – An inventory of past events and identification of some patterns. J. Loss Prev. Process Ind. 24, 575–593. https://doi.org/10.1016/j.jlp.2010.06.013

Abdulkhaleq, A., Wagner, S., Leveson, N., 2015. A Comprehensive Safety Engineering Approach for Software-Intensive Systems Based on STPA. Procedia Eng. 128, 2–11. https://doi.org/10.1016/j.proeng.2015.11.498

Abe, N., Zadrozny, B., Langford, J., 2006. Outlier detection by active learning, in: The 12th ACM SIGKDD International Conference. ACM Press, pp. 504–504. https://doi.org/10.1145/1150402.1150459

Adaku, E., Ankrah, N.A., Ndekugri, I.E., 2021. Design for occupational safety and health: A theoretical framework for organisational capability. Saf. Sci. 133, 105005–105005. https://doi.org/10.1016/j.ssci.2020.105005

Aggarwal, C.C., 2017. Outlier Analysis, Second. ed. Springer International Publishing, Cham, Switzerland.

Aggarwal, C.C., Hinneburg, A., Keim, D.A., 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space, in: Van den Bussche, J., Vianu, V. (Eds.), Database Theory — ICDT 2001, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 420–434. https://doi.org/10.1007/3-540-44503-X_27

Al-Aqrabi, H., Liu, L., Hill, R., Antonopoulos, N., 2015. Cloud BI: Future of business intelligence in the Cloud. J. Comput. Syst. Sci. 81, 85–96. https://doi.org/10.1016/j.jcss.2014.06.013

Anandarajan, M., Hill, C., Nolan, T., 2019. Practical Text Analytics: Maximizing the Value of Text Data, Advances in Analytics and Data Science. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-95663-3

Ariyachandra, T., Watson, H., 2005. Key factors in selecting a data warehouse architecture. Bus. Intell. J. 10 (3).

Blanc, F., Escobar Pereira, M.M., 2020. Risks, Circumstances and Regulation – Historical development, diversity of structures and practices in Occupational Safety and Health inspections. Saf. Sci. 130, 104850–104850. https://doi.org/10.1016/j.ssci.2020.104850

Bouguessa, M., 2015. A practical outlier detection approach for mixed-attribute data. Expert Syst. Appl. 42, 8637–8649. https://doi.org/10.1016/j.eswa.2015.07.018

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., Song, A., 2015. Efficient agglomerative hierarchical clustering. Expert Syst. Appl. 42, 2785–2797. https://doi.org/10.1016/j.eswa.2014.09.054

Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers, in: The 2000 ACM SIGMOD International Conference. ACM Press, pp. 93–104. https://doi.org/10.1145/342009.335388

Burkov, A., 2019. The hundred page machine learning book. Andriy Burkov, Quebec, Canada.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 1–58. https://doi.org/10.1145/1541880.1541882

Chaudhuri, S., Dayal, U., Narasayya, V., 2011. An overview of business intelligence technology. Commun. ACM 54, 88–98. https://doi.org/10.1145/1978542.1978562

Chawla, S., Sun, P., 2006. SLOM: a new measure for local spatial outliers. Knowl. Inf. Syst. 9, 412–429. https://doi.org/10.1007/s10115-005-0200-2

Chen, Chiang, Storey, 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Q. 36, 1165–1165. https://doi.org/10.2307/41703503

Cheng, C.-W., Yao, H.-Q., Wu, T.-C., 2013. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. J. Loss Prev. Process Ind. 26, 1269–1278. https://doi.org/10.1016/j.jlp.2013.07.002

Ciampi, F., Demi, S., Magrini, A., Marzi, G., Papa, A., 2021. Exploring the impact of big data analytics capabilities on business model innovation: The mediating role of entrepreneurial orientation. J. Bus. Res. 123, 1–13. https://doi.org/10.1016/j.jbusres.2020.09.023

Cooper Ph.D., M.D., 2000. Towards a model of safety culture. Saf. Sci. 36, 111–136. https://doi.org/10.1016/S0925-7535(00)00035-7

De Felice, F., Travaglioni, M., Piscitelli, G., Cioffi, R., Petrillo, A., 2019. Machine learning techniques applied to industrial engineering: A multi criteria approach, in: 18th International Conference on Modeling and Applied Simulation, MAS 2019. Dime University of Genoa, pp. 44–54. https://doi.org/10.46354/i3m.2019.mas.007

Dekker, S., 2019. Foundations of safety science: a century of understanding accidents and disasters. Taylor & Francis Group, CRC Press, Boca Raton.

Ding, Z., Fei, M., 2013. An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window. IFAC Proc. Vol. 46, 12–17. https://doi.org/10.3182/20130902-3-CN-3020.00044

Dorsey, L.T.C., Wang, B., Grabowski, M., Merrick, J., Harrald, J.R., 2020. Self healing databases for predictive risk analytics in safety-critical systems. J. Loss Prev. Process Ind. 63, 104014–104014. https://doi.org/10.1016/j.jlp.2019.104014

Edwards, D., 2016. How Machines Learn. Intelligentsia Research.

Edwin, N.J., Paltrinieri, N., Østerlie, T., 2016. Risk Metrics and Dynamic Risk Visualization, in: Dynamic Risk Analysis in the Chemical and Petroleum Industry. Butterworth-Heinemann, pp. 151–165.

Elmasri, R., Navathe, S., 2011. Fundamentals of database systems, 6th ed. ed. Addison-Wesley, Boston.

El-Sappagh, S.H.A., Hendawi, A.M.A., El Bastawissy, A.H., 2011. A proposed model for data warehouse ETL processes. J. King Saud Univ. - Comput. Inf. Sci. 23, 91–104. https://doi.org/10.1016/j.jksuci.2011.05.005

Fruhen, L.S., Mearns, K.J., Flin, R., Kirwan, B., 2014. Safety intelligence: An exploration of senior managers' characteristics. Appl. Ergon. 45, 967–975. https://doi.org/10.1016/j.apergo.2013.11.012

Guldenmund, F.W., 2000. The nature of safety culture: a review of theory and research. Saf. Sci. 34, 215–257. https://doi.org/10.1016/S0925-7535(00)00014-X

Harding, A.B., 1997. MHIDAS: the first ten years. 141 12–12.

Hariri, S., Carrasco Kind, M., Brunner, R.J., 2019. Extended Isolation Forest. IEEE Trans. Knowl. Data Eng. 1–1. https://doi.org/10.1109/TKDE.2019.2947676

Inmon, W.H., 2014. Data architecture: a primer for the data scientist, 1st editio. ed. Elsevier, Waltham, MA.

Inmon, W.H., 2005. Building the data warehouse, 4th ed. Wiley, New York.

Jain, P., Pasman, H.J., Waldram, S., Pistikopoulos, E.N., Mannan, M.S., 2018. Process Resilience Analysis Framework (PRAF): A systems approach for improved risk and safety management. J. Loss Prev. Process Ind. 53, 61–73. https://doi.org/10.1016/j.jlp.2017.08.006

Jamshidi, A., Faghih-Roohi, S., Hajizadeh, S., Babuska, R., Dollevoet, R., Li, Z., Schutter, B.D., 2017. A Big Data Analysis Approach for Rail Failure Risk Assessment 13.

Khan, W.A., Chung, S.H., Awan, M.U., Wen, X., 2019. Machine learning facilitated business intelligence (Part I): Neural networks learning algorithms and applications. Ind. Manag. Data Syst. 120, 164–195. https://doi.org/10.1108/IMDS-07-2019-0361

Kimball, R., Caserta, J., 2011. The Data WarehouseETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, Inc., Hoboken.

Kingrani, S.K., Levene, M., Zhang, D., 2017. Estimating the number of clusters using diversity. Artif. Intell. Res. 7, 15–15. https://doi.org/10.5430/air.v7n1p15

Kriegel, H.-P., Kroger, P., Schubert, E., Zimek, A., 2011. Interpreting and Unifying Outlier Scores, in: Proceedings of the 2011 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 13–24. https://doi.org/10.1137/1.9781611972818.2

Lazarevic, A., Kumar, V., 2005. Feature bagging for outlier detection, in: Proceeding of the Eleventh ACM SIGKDD International Conference. ACM Press, pp. 157–157. https://doi.org/10.1145/1081870.1081891

Lees', 2012. Lees' Loss Prevention in the Process Industries, in: Lees' Loss Prevention in the Process Industries. Elsevier, pp. 3661–3661.

Leveson, N., 2011. Applying systems thinking to analyze and learn from events. Saf. Sci. 49, 55–64. https://doi.org/10.1016/j.ssci.2009.12.021

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation Forest, in: 2008 Eighth IEEE International Conference on Data Mining (ICDM). IEEE, pp. 413–422. https://doi.org/10.1109/ICDM.2008.17

Llopart, S.C., 2001. BASES DE DATOS sobre accidentes industriales. MAPFRE N. 155, 47–56.

Loshin, D., 2013. Business Intelligence, The Savvy Manager's Guide, Second. ed. Morgan Kaufmann, Elseiver, Boston.

Manuele, F.A., 2008. Advanced safety management focusing on Z10 and serious injury prevention. Wiley-Interscience, Hoboken, N.J.

Mao, S., Zhao, Y., Chen, J., Wang, B., Tang, Y., 2020. Development of process safety knowledge graph: A Case study on delayed coking process. Comput. Chem. Eng. 143, 107094. https://doi.org/10.1016/j.compchemeng.2020.107094

Marle, F., Vidal, L.-A., 2016. Managing Complex, High Risk Projects. Springer London, London.

Marshall, I.J., Wallace, B.C., 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst. Rev. 8, 163, s13643-019-1074–9. https://doi.org/10.1186/s13643-019-1074-9

Menze, B.H., Kelm, B.M., Splitthoff, D.N., Koethe, U., Hamprecht, F.A., 2011. On Oblique Random Forests, in: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 453–469.

Micán, C., Fernandes, G., Araújo, M., Ares, E., 2019. Operational risk categorization in project-based organizations: A theoretical perspective from a project portfolio risk lens. Procedia Manuf. 41, 771–778. https://doi.org/10.1016/j.promfg.2019.09.069

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set 21–21.

Monteiro, G.P., 2020. How do organizational structures impact operational safety? Part 1 – Understanding the dangers of decentralization. Saf. Sci. 13–13.

Monteiro, G.P., Hopkins, A., Frutuoso e Melo, P.F., 2020. How do organizational structures impact operational safety? Part 2 – Designing structures that strengthen safety. Saf. Sci. 123, 104534–104534. https://doi.org/10.1016/j.ssci.2019.104534

Morgan, J.I., 2021. Implementing the theoretical domains framework in occupational safety: Development of the safety behaviour change questionnaire. Saf. Sci. 14.

Mur, A., Dormido, R., Duro, N., Dormido-Canto, S., Vega, J., 2016. Determination of the optimal number of clusters using a spectral clustering optimization. Expert Syst. Appl. 65, 304–314. https://doi.org/10.1016/j.eswa.2016.08.059

Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. WIREs Data Min. Knowl. Discov. 2, 86–97. https://doi.org/10.1002/widm.53

Oracle, 2004. Oracle Business Intelligence. Concept guide. Oracle.

Paltrinieri, N., Bonvicini, S., Spadoni, G., Cozzani, V., 2012. Cost-Benefit Analysis of Passive Fire Protections in Road LPG Transportation: Cost-Benefit Analysis of Passive Fire Protections. Risk Anal. 32, 200–219. https://doi.org/10.1111/j.1539-6924.2011.01654.x

Paltrinieri, N., Comfort, L., Reniers, G., 2019. Learning about risk: Machine learning for risk assessment. Saf. Sci. 118, 475–486. https://doi.org/10.1016/j.ssci.2019.06.001

Paltrinieri, N., Khan, F., 2016. Dynamic Risk Analysis in the Chemical and Petroleum Industry. Butterworth-Heinemann.

Paltrinieri, N., Patriarca, R., Pacevicius, M., Rossi, P.S., 2020. Lessons from past hazardous events: data analytics for severity prediction 8–8.

Patriarca, R., Di Gravio, G., Cioponea, R., Licu, A., 2019. Safety intelligence: Incremental proactive risk management for holistic aviation safety performance. Saf. Sci. 118, 551–567. https://doi.org/10.1016/j.ssci.2019.05.040

Patriarca, R., Di Gravio, G., Mancini, M., Costantino, F., 2016. Change management in the ATM system: Integrating information in the preliminary system safety assessment. Int. J. Appl. Decis. Sci. 9, 121–138. https://doi.org/10.1504/IJADS.2016.080123

Patriarca, R., Di Gravio, G., Woltjer, R., Costantino, F., Praetorius, G., Ferreira, P., Hollnagel, E., 2020. Framing the FRAM: A literature review on the functional resonance analysis method. Saf. Sci. 129. https://doi.org/10.1016/j.ssci.2020.104827

Ramaswamy, S., Rastogi, R., Shim, K., 2000. Efficient Algorithms for Mining Outliers from Large Data Sets 12–12.

Reddy, D., Lingras, P., Venkatanareshbabu, K. (Eds.), 2018. Advances in Machine Learning and Data Science: Recent Achievements and Research Directives, Advances in Intelligent Systems and Computing. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-10-8569-7

Sarkar, S., Maiti, J., 2020. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. Saf. Sci. 131, 104900–104900. https://doi.org/10.1016/j.ssci.2020.104900

Sharda, R., 2020. Analytics, data science, & artificial intelligence, Eleventh e. ed. Pearson, Hoboken, NJ.

Sharda, R., Delen, D., Turban, E., 2018. Business intelligence, analytics, and data science: a managerial perspective, Fourth edi. ed. Pearson, New York, NY.

Shekhar, S., Lu, C.-T., Zhang, P., 2001. Detecting Graph-Based Spatial Outliers: Algorithms and Applications(A Summary of Results)* t 6–6.

Singh, K., Upadhyaya, D.S., 2012. Outlier Detection: Applications And Techniques 9, 17–17.

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., Yahia, S.B., 2019. Data quality in ETL process: A preliminary study. Procedia Comput. Sci. 159, 676–687. https://doi.org/10.1016/j.procs.2019.09.223

Stefana, E., Marciano, F., Cocca, P., Rossi, D., Tomasoni, G., 2019. Oxygen deficiency hazard in confined spaces in the steel industry: assessment through predictive models. Int. J. Occup. Saf. Ergon. 1–15. https://doi.org/10.1080/10803548.2019.1669954

Susto, G.A., Beghi, A., McLoone, S., 2017. Anomaly Detection through on-line Isolation Forest: an Application to Plasma Etching 6–6.

Swuste, P., van Nunen, K., Reniers, G., Khakzad, N., 2019. Domino effects in chemical factories and clusters: An historical perspective and discussion. Process Saf. Environ. Prot. 124, 18–30. https://doi.org/10.1016/j.psep.2019.01.015

Tauseef, S.M., Abbasi, T., Abbasi, S.A., 2011. Development of a new chemical process-industry accident database to assist in past accident analysis. J. Loss Prev. Process Ind. 24, 426–431. https://doi.org/10.1016/j.jlp.2011.03.005

Trujillo, J., Luján-Mora, S., Goos, G., Hartmanis, J., van Leeuwen, J., 2003. A UML Based Approach for Modeling ETL Processes in Data Warehouses, in: Song, I.-Y., Liddle, S.W., Ling, T.-W., Scheuermann, P. (Eds.), Conceptual Modeling - ER 2003. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 307–320.

Uhrenholdt Madsen, C., Kirkegaard, M.L., Dyreborg, J., Hasle, P., 2020. Making occupational health and safety management systems 'work': A realist review of the OHSAS 18001 standard. Saf. Sci. 129, 104843–104843. https://doi.org/10.1016/j.ssci.2020.104843

Väyrynen, S., Häkkinen, K., Niskanen, T., 2015. Integrated Occupational Safety and Health Management. Springer International Publishing, Cham.

Villafañe, D., Darbra, R.M., Casal, J., 2011. Flash fire: Historical analysis and modeling, in: Chemical Engineering Transactions. Italian Association of Chemical Engineering - AIDIC, pp. 1111–1116. https://doi.org/10.3303/CET1124186

Wang, R.Y., Strong, D.M., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manag. Inf. Syst. 12, 5–33. https://doi.org/10.1080/07421222.1996.11518099

Warrender, C., Forrest, S., Pearlmutter, B., 1999. Detecting intrusions using system calls: alternative data models, in: 1999 IEEE Symposium on Security and Privacy. IEEE Comput. Soc, pp. 133–145. https://doi.org/10.1109/SECPRI.1999.766910

Watson, H.J., Wixom, B.H., 2007. The Current State of Business Intelligence. Computer 40, 96–99. https://doi.org/10.1109/MC.2007.331

Xu, D., Wang, Y., Meng, Y., Zhang, Z., 2017. An Improved Data Anomaly Detection Method Based on Isolation Forest, in: 2017 10th International Symposium on Computational Intelligence and Design (ISCID). IEEE, pp. 287–291. https://doi.org/10.1109/ISCID.2017.202

Yang, X., Haugen, S., Paltrinieri, N., 2018. Clarifying the concept of operational risk assessment in the oil and gas industry 259–268. https://doi.org/10.1016/j.ssci.2017.12.019

Zarei, E., Yazdi, M., Abbassi, R., Khan, F., 2019. A hybrid model for human factor analysis in process accidents: FBN-HFACS. J. Loss Prev. Process Ind. 57, 142–155. https://doi.org/10.1016/j.jlp.2018.11.015

Zhang, Z., Wu, Z., Rincon, D., Garcia, C., Christofides, P.D., 2019. Operational safety of chemical processes via Safeness-Index based MPC: Two large-scale case studies. Comput. Chem. Eng. 125, 204–215. https://doi.org/10.1016/j.compchemeng.2019.03.003

Zhu, R., Hu, X., Hou, J., Li, X., 2021. Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Saf. Environ. Prot. 145, 293–302. https://doi.org/10.1016/j.psep.2020.08.006