Exploratory Analysis of the HIAD Database: A Machine Learning Approach to Support Hydrogen Risk Assessment

Jairo Andrés Meneses-Gelves1, Rafael Amaya-Gómez2, Pascale Vacca1, Elsa Pastor1,\*

1Centre for Technological Risk Studies (CERTEC), Universitat Politècnica de Catalunya ⋅ Barcelona Tech, Barcelona East School of Engineering (EEBE). Avinguda Eduard Maristany, 16, 08019 Barcelona, Spain;

 2Department of Industrial Engineering, Universidad de los Andes, Cra 1E No. 19A-40, Bogotá, Colombia

\*elsa.pastor@upc.edu

1. Introduction

The growing interest in hydrogen technologies due to the global energy transition introduces new safety challenges arising from the unique physicochemical properties of hydrogen. Although previous studies have explored the application of machine learning to extract lessons learned from historical incidents and accident databases from the chemical industry, research specifically focused on hydrogen-related events remains limited. Tamascelli et al. (2022) utilized classification models trained on the Major Hazard Incident Data Service (MHIDAS) to estimate the severity of chemical accidents. Tamascelli et al. (2023) extended this approach using meta-learning and transfer-learning techniques in the context of ammonia. Similarly, Kurian et al. (2020) applied supervised learning and keyword analysis to a database of oil sands incident reports, developing trend analyses, risk matrices, and prevention strategies.

The present study explores the application of machine learning techniques to historical hydrogen-related accidents and incident databases. It presents a structured data pre-processing framework tailored to machine learning applications to gain insight on hydrogen-related events. It includes an exploratory and unsupervised learning assessment to detect hidden relationships between input features and accident outcomes (e.g., fatalities or injuries). This work is aligned with the implementation of Safety 5.0 principles, which aims to incorporate artificial intelligence, digital technologies, and real-time analytics into risk management practices (Pasman & Behie, 2024). Post-accident analysis and safety-informed decision making can benefit from identifying how machine learning approaches can extract meaningful patterns from structured hydrogen-related events.

2. Methods

The overall workflow is outlined in Figure 1. The methodology develops in three steps. First, data extraction from the Hydrogen Incident and Accident Database (HIAD) database, which comprises hydrogen-related events reports. Second, data pre-processing encompasses two sub-steps: (i) feature selection, whereby a set of accident descriptors is chosen based on literature review, and (ii) data cleaning, which standardizes categorical labels and filters out records missing any of the key outcome variables (in this study defined as number of fatalities and number of injuries). Finally, accident-based analysis applies Multiple Correspondence Analysis (MCA) to the resulting dataset, transforming the categorical feature set into a low-dimensional factor space. Detailed explanations of each stage are given in the following sections.



Figure 1. Methodology workflow.

2.1 Accident database: HIAD

The accident data are extracted from the Hydrogen Incident and Accident Database (HIAD) version 2.1, provided as an Excel file (Joint Research Centre, 2025). This dataset comprises 954 unique reports from 1785 to 2025 around the world. related to hydrogen incidents and accidents. Maintained by the European Commission’s Joint Research Centre (JRC), the database is specifically dedicated to documenting hydrogen-related events.

Data collection in HIAD 2.1 is based on a systematic approach that harvests information from publicly available primary and secondary sources, including peer-reviewed literature, government investigation reports, and open-access event summaries. To support the standardized information in the database, HIAD 2.1 incorporates three integrated modules: the Data Entry Module (DEM), which uses structured templates and controlled vocabularies to capture event descriptors; the Data Retrieval Module (DRM), which enables multi-criteria filtering and data export; and the Data Analysis Module (DAM), which provides both pre-configured statistical overviews and customizable analytical outputs. This modular architecture enhances the consistency and efficiency of the entire data lifecycle.

HIAD 2.1 employs a five-level labelling system to define the data quality. Unvalidated submissions begin at Quality 1 (not publicly released), while validated records advance through successive levels up to Quality 5, representing full data richness, including original investigation files and detailed quantitative information. JRC analysts review each entry for completeness, consistency, and traceability before a quality label is assigned.

2.2 Data pre-processing: Features selection and data cleaning

The original HIAD dataset was consolidated into a single table by merging columns from multiple worksheets based on their Event ID, thereby unifying core metadata, facility characteristics, consequence metrics, and quality labels into one record per event. Subsequently, a subset of 20 accident features was selected according to the framework proposed by Tamascelli et al. (2022), as shown in Table 1.

To develop the Machine Learning approach, the Country (CO) feature was excluded in favour of broader regional grouping to capture essential spatial variability while avoiding the high-cardinality and sparse encoding issues that can degrade classifier performance. Likewise, the Year (YE) feature was omitted since temporal trends did not enhance the causal–consequence relationships of interest. Finally, only descriptors present in at least 70 % of the 954 reports were retained, ensuring adequate data coverage for a robust machine learning analysis.

Table 1. Accident features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Code | Description | Data type | Missing values (%)  | Selected feature |
| ID | Event ID | Numerical | 0.0% | No |
| Q | Quality of the data | Categorical | 0.0% | Yes |
| NA | Nature of the consequences | Categorical | 0.0% | Yes |
| RE | Region (continent) | Categorical | 0.0% | Yes |
| YE | Year | Numerical | 0.0% | No |
| CA | Causes | Categorical | 0.0% | Yes |
| EV | Event Initiating system | Categorical | 0.9% | Yes |
| CL | Classification of the physical effects | Categorical | 1.0% | Yes |
| CO | Country | Categorical | 1.3% | No |
| AP | Application type | Categorical | 1.9% | Yes |
| NUI | Number of injures | Numerical | 5.9% | Yes\* |
| OP | Operational condition | Categorical | 11.7% | Yes |
| NUF | Number of fatalities | Numerical | 12.4% | Yes\* |
| LOD | Location description | Categorical | 13.8% | Yes |
| ST | Storage/process medium | Categorical | 15.1% | Yes |
| RET | Release type | Categorical | 17.9% | Yes |
| LOT | Location type | Categorical | 27.5% | Yes |
| IG | Ignition source | Categorical | 77.2% | No |
| REA | Release amount | Numerical | 82.6% | No |
| FL | Flame type | Categorical | 89.9% | No |

\*Proposed categorical variable

Regarding the data cleaning step, categorical label unification was done by standardizing text entries to a consistent format and correcting typographical variations, thereby reducing noise and ensuring uniform category encoding. Since the subsequent machine learning approach requires at least one observed outcome per record, any accident report lacking fatality and injury information was excluded. However, reports that included at least one of these target variables were retained to preserve partially observed cases and maintain sample diversity. This filtering procedure removed 260 records from the original 954, leaving 694 data events available.

The numerical variables (NUI and NUF) were modified into categorical variables following the CCPS Tier 1 Process Safety Event (PSE1) criteria, whereby any event with one or more fatalities qualifies as a catastrophic event (Center for Chemical Process Safety, 2018). Based on these standards, both NUI and NUF were converted as shown in Table 2.

Table 2. Accident consequences categories. People harmed refers to injuries or fatalities.

|  |  |
| --- | --- |
| Category  | Description |
| 0 | No people harmed |
| >1 | More than 1 people harmed |
| Unknown | Not reported |

2.3 Accident-based analysis: Multiple Correspondence Analysis (MCA)

The Multiple Correspondence Analysis (MCA) approach was selected because the features are categorical or have been categorized (as shown in Table 2). Unlike Principal Component Analysis (PCA), which optimally represents continuous variables in a low-dimensional Euclidean space, MCA can accommodate any number of categorical variables by transforming them into a complete indicator matrix (Greenacre, 2010). This transformation preserves the $χ^{2}$ distances between category profiles following the PCA framework, so they are defined as the difference between categorical observations to quantify the degree of similarity among individuals or categories. This approach enables the simultaneous exploration of associations among variable categories. MCA itself is an extension of classical correspondence analysis to more than two categorical variables.

The MCA begins by constructing an indicator matrix in which each row represents an event and each column represents one category of one feature. From this matrix, a correspondence table of relative frequencies is computed, and a singular value decomposition (SVD) yields principal axes (dimensions) that capture the largest share of the total inertia of the categorical cloud. Both individuals (events) and categories (accident features) are in two clouds that their dimensions come from the number of categories and registers and they are used to evaluate similarities depending on the closeness. Further details on this approach can be found in Amaya-Gómez et al. (2021) and Greenacre (2010).

Each event and each category are then projected onto these new axes, producing factor coordinates that reveal the clustering of similar events and co-occurrence patterns among categories. By retaining the first few dimensions, MCA reduces data complexity while maintaining the interpretability of categorical relationships. In this work, the package of “FactoMineR”, “FactoExtra” and “Factoshiny” in RStudio 2024.12.1 is implemented to evaluate the MCA results in terms of the inertia (variability) described, and the obtained individuals and categories clouds.

3. Results and discussion

3.1 Exploratory analysis of the HIAD 2.1

The distribution of the fatalities and injuries grouped by decade is shown in Figure 2. Both fatalities and injuries increased steadily from the 1960–1970 decade through 2000–2010. However, both curves turn downward from 2010–2020 to 2025, showing an apparent decline. This post-2010 decrease likely stems from significant advances in hydrogen safety practices. For instance, the first edition of NFPA 2, Hydrogen Technologies Code, was published in 2011.

Regarding the events occurring prior to 1960, the most significant contribution to the high‐fatality value corresponds to the LZ-129 Hindenburg disaster in 1937, which resulted in 35 fatalities. In contrast, the main injury value arises from the 1959 explosion at a Japanese chemical plant, which resulted in 11 fatalities and 44 injuries.



Figure 2. Number of fatalities and injuries per decade.

Figure 3 illustrates the proportional distribution of hydrogen‐related accidents by region and the total of events across decade. For the period before 1960, North America, Europe and Asia contributed exactly with one event, yielding equal proportions. From 1960 onwards, Europe’s share of proportion of the events increased steadily, whereas North America’s share declined. Event proportions in Oceania and South America remain negligible, reflecting the developing adoption of hydrogen technologies in those markets.



Figure 3. Distribution of hydrogen-related events (bar) and total events (line) per decade.

3.2 Machine Learning approach: Multiple Correspondence Analysis

The scree plot of principal inertia (Figure 4) shows the associative structure of the categorical variables and guides the selection of dimensions for further analysis. The first two dimensions account for 6.85% of the total inertia (3.87% and 2.98%, respectively), indicating that this two-dimensional plane explains a modest share of the event–feature cloud variability. Although this proportion appears small, it exceeds the baseline reference of 2.9% (equivalent to the average inertia per dimension under an independent model), confirming that the first plane captures meaningful structure rather than noise.

The observed inertias were compared against the 0.95-quantile of inertias derived from randomly permuted indicator matrices to determine the number of dimensions carrying substantive information. Fourteen dimensions shown inertias greater than this permutation-based threshold (23.79% cumulative inertia versus 18.19% expected at the 95% quantile), suggesting that only these axes represent genuine associations among accident descriptors. Consequently, subsequent clustering and typology development draw exclusively on the factor coordinates of these significant axes, ensuring that derived event classes reflect true categorical co-occurrence rather than sampling artefacts. However, the analysis presented in this study focuses on the first two dimensions. This choice reflects a deliberate focus on the most interpretable plane of variation, even when it accounts for a modest proportion of the categorical cloud.



Figure 4. Decomposition of the total inertial.

Figure 5 shows the correlation ratio for the first two MCA dimensions. Each axis quantifies the correlation ratio between a variable and its respective component. If this correlation ratio is close to unity for a given component, individuals in the same category have similar coordinates (Husson et al., 2017). This visualization, therefore, highlights which variables co‐vary most tightly in the reduced space. In this study, the supplementary elements (NUI and NUF) lie in close proximity to three variables: Operational Condition (OP), Location Type (LOT), Quality of the data (Q), and Region (RE). The OP variable distinguishes whether the process state at the time of the accident was “normal,” “abnormal,” or “unknown,” and its strong alignment with the first two dimensions suggests that deviations from standard operating regimes are highly predictive of human‐impact severity. Similarly, the LOT variable, which categorizes the event environment as “open,” “confined,” “semi‐confined,” or “unknown,” exhibits a high correlation ratio, indicating that spatial enclosure characteristics materially influence casualty outcomes. The Quality of the data (Q) correlates with the dimensions in a manner that underscores the importance of data completeness. Finally, the RE variable confirms the information discussed in Figure 3. These correlations affirm that process state, spatial confinement, and regional factors underpin the principal patterns of human harm captured by the MCA.



Figure 5. Variable correlation with the two principal MCA components.

Figure 6 presents the individual factor map defined by the first two MCA dimensions. The Wilks test was conducted to identify which categorical variables best discriminate among event points on this plane, with lower p-values indicating stronger separation. Among the two factors (NUI and NUF), the NUI variable yielded the most significant p-value, confirming it as the primary driver of distance between individual accidents on these axes. In the map, each point represents a single hydrogen-event report. For the black points, related with NUI>1, they are predominantly on the positive side of Dim 1, indicating that high-injury events share similar categorical profiles. The sharing labels had high frequency for events that occurred in industrial applications (AP = “Industry”), were driven by explosive failure modes (NA = “Explosion”), took place within industrial facility types (LOD = “Industrial”), and coincided with abnormal operational conditions (OP = “ABNORMAL”).



Figure 6. Individual factor map.

Figure 7 illustrates the MCA factor results, highlighting those with an associated squared‐cosine (cos²) values greater than 0.2. The events in which both the NUI and NUF fall into the “zero” category tightly around the “normal” operational condition and the “open” location type. This alignment intuitively reflects that events without human casualties tend to occur under routine process states and in unconfined environments. The NUI and NUF in the category ">1" cluster with the following categories: CL= "Hydrogen release and ignition", NA= “Explosion”, AP= “Industry”, and LOD= “Industrial”. This means that events with more than one fatality or injury are most likely to involve hydrogen releases and ignition with an explosion as a consequence in industrial settings. However, these categories associated squared‐cosine values, which measure the level of association among categories, remain uniformly low, indicating that neither dimension captures a large proportion of their variance. This outcome demonstrates a substantial level of noise in the dataset. It suggests that the current categorical descriptors and their granularity may insufficiently isolate the principal accident severity drivers.

Several strategies merit consideration to reduce data noise and enhance signal clarity in future analyses. First, refining or consolidating low‐frequency categories will prevent sparse levels from disproportionately inflating dimensionality. Second, exploring alternative accident descriptors could yield more discriminative feature sets. Finally, applying supervised feature‐selection techniques or regularized embedding methods before MCA can help identify the most informative variables and suppress random variation. Together, these refinements promise to strengthen the robustness of MCA‐derived typologies and improve the interpretability of hydrogen‐event analyses.



Figure 7. Categories factor results with a Cos2=0.2.

4. Conclusions

This study examined a Machine Learning exploration of the Hydrogen Incident and Accident Database (HIAD 2.1). Integrating a pre-processing framework with a Multiple Correspondence Analysis, this work demonstrates that low-dimensional representations of 694 hydrogen-related events can be obtained despite the database’s high categorical complexity and missing-value patterns.

The analysis identifies operational conditions, spatial confinement, and geographic regions as the categorical triad most strongly associated with the severity of human impact. Events occurring under abnormal process states and confined environments cluster in the same factor space as events with more than one injury, whereas casualty-free events align with normal operations in open settings.

This work establishes a transferable workflow aligned with Safety 5.0 principles, illustrating how unsupervised learning can increase post-event investigations, guide the prioritisation of safeguards and improve the risk assessments for emerging hydrogen infrastructures. Nevertheless, the modest inertia captured by the first MCA plane signals residual noise arising from sparse or low-frequency categories. Future research should therefore refine label taxonomies and incorporate higher-fidelity process descriptors.

Acknowledgments

The first author works within the framework of Project 101119358, ‘PROSAFE’, funded by the Marie Skłodowska-Curie Actions programme, HORIZON-MSCA-2022-DN-01. This work was also partly funded by Grant PID2023-150607OB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU.

European Hydrogen Incidents and Accidents database HIAD 2.1, European Commission, Joint Research Centre

References

Amaya-Gómez, R., Bastidas-Arteaga, E., Muñoz, F., & Sánchez-Silva, M. (2021). Statistical Soil Characterization of an Underground Corroded Pipeline Using In-Line Inspections. Metals, 11(2), 292. https://doi.org/10.3390/met11020292

Center for Chemical Process Safety. (2018). Process Safety Metrics: Guide for Selecting Leading and Lagging Metrics. AIChE. https://www.aiche.org/sites/default/files/docs/pages/ccps\_process\_safety\_metrics\_-\_april\_2018.pdf

Greenacre, M. J. (2010). Correspondence analysis. WIREs Computational Statistics, 2(5), 613–619. https://doi.org/10.1002/wics.114

Husson, F., Lê, S., & Pagès, J. (2017). Exploratory multivariate analysis by example using R (Second edition). CRC Press Taylor & Francis Group.

Joint Research Centre. (2025). European Hydrogen Incidents and Accidents database HIAD 2.1. European Commission.

Kurian, D., Sattari, F., Lefsrud, L., & Ma, Y. (2020). Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. Safety Science, 130, 104873. https://doi.org/10.1016/j.ssci.2020.104873

Pasman, H. J., & Behie, S. W. (2024). The evolution to Industry 5.0 / Safety 5.0, the developments in society, and implications for industry management. Journal of Safety and Sustainability, 1(4), 202–211. https://doi.org/10.1016/j.jsasus.2024.11.003

Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2023). Learning From Major Accidents: A Meta-Learning Perspective. Safety Science, 158, 105984. https://doi.org/10.1016/j.ssci.2022.105984

Tamascelli, N., Solini, R., Paltrinieri, N., & Cozzani, V. (2022). Learning from major accidents: A machine learning approach. Computers & Chemical Engineering, 162, 107786. https://doi.org/10.1016/j.compchemeng.2022.107786