

SIRIMA

SINKHOLE HAZARD AND RISK MANAGEMENT IN POST MINING AREAS
RFCS PROJECT NO 101157400



Co-funded by
the European Union

Operational Tool for Predicting of Sinkhole Hazard

WORK PACKAGE 4 - Task 4.1 DELIVERABLE 4.2

Responsible partner:

INERIS: Nathalie Conil (Nathalie.CONIL@ineris.fr), part 1

SYSTRA Subterra: Geovanny Benavides (gbenavides@systra.com); Guillermo Vaquero (gvaquero@systra.com), part 2

Contributing Partners:

1. Bureau de Recherches Géologiques et Minières - BRGM, France,
2. Główny Instytut Górnictwa – Państwowy Instytut Badawczy, Poland,
3. Instytut Mechaniki Górotworu Polskiej Akademii Nauk – IMG-PAN, Poland,
4. Université de Lorraine, France.

This Deliverable has been made as a result of the SIRIMA project – Sinkhole hazard and risk management in post mining areas.

September, 2026

Abstract

This document presents a methodology for the assessment and prediction of sinkhole hazards in European post-mining regions, developed under The SIRIMA: Sinkhole Hazard and Risk Management in Post-Mining Areas project carried out with co-financing from the Research Found for Coal and Steel of the European Union as the action RFCS-2023-01-RPJ Coal Research Projects under Grant Agreement No. 101157400. The study is structured into two main phases: Part 1 details the cleaning, harmonisation, and exploratory statistical analysis of a database comprising 1,868 historical sinkhole records associated with shallow abandoned coal mines across France, Germany, and Poland. This foundational analysis identifies critical geomechanical parameters while addressing the inherent challenges of historical data incompleteness and heterogeneity.

Building upon this dataset, Part 2 introduces a data-driven Operational Hazard Tool based on a dual-modelling approach. First, a Susceptibility Model evaluates the probability of sinkhole occurrence using a weighted sum algorithm governed by working depth, extracted seam thickness, and overburden geomechanical resistance. To overcome data gaps, the model integrates advanced Machine Learning techniques, including Natural Language Processing (NLP) for lithological classification and K-Nearest Neighbors for spatial imputation. Concurrently, a Magnitude Model employs a conditional probability framework to predict the expected physical severity (Low, Medium, or High) of the collapse based on historical combinations of working thickness and the coal seam dip angle.

The outputs of both models are integrated into a unified Hazard Matrix that classifies the terrain into discrete operational risk levels. Finally, this predictive framework is deployed as an interactive web-based application, providing land-management authorities and stakeholders with an automated decision-support system to transition from reactive emergency responses to proactive geohazard management.

While this tool provides an operational framework for hazard assessment, its outputs must be interpreted with engineering caution. The predictive accuracy is inherently constrained by the historical training database, incompleteness of the sinkhole catalog, geomechanical assumptions and statistical imputations. Consequently, this tool should serve as a complementary screening instrument rather than an absolute geotechnical verdict, recognizing that localized, high-resolution data and advanced field testing remain necessary to fully validate site-specific hazard mitigation strategies.

Table of Content

1.	Introduction.....	9
2.	Data processing and preparation for hazard modelling	10
2.1	Scope and structure of the input dataset	10
2.2	Data cleaning and harmonisation	10
2.3	Definition of analytical populations.....	11
3.	Overview of data completeness and exploitable variables	12
3.1	Completeness of the main variables.....	12
3.2	Heterogeneous completeness by country: example with the working characteristics.....	15
3.3	Selection of variables for the analysis	17
4.	Spatial and geographical analysis of sinkhole records	18
4.1	Distribution by country	18
4.2	Spatial distribution of sinkholes.....	20
4.3	Representativeness and limitations of the inventory	20
5.	Temporal analysis of sinkhole occurrence and mining periods	21
5.1	Availability and reliability of temporal information	21
5.2	Distribution of sinkhole occurrence over time	23
5.3	Comparison between occurrence dates and mining periods.....	28
5.4	Interpretation and limitations.....	30
6.	Analysis of sinkhole characteristics.....	31
6.1	Sinkhole dimensions	31
6.2	Origin of collapse	35
7.	Analysis of mining-related characteristics	37
7.1	Working depth	37
7.2	Working thickness	38
7.3	Working-depth to working-thickness ratio.....	39
7.4	Coal seam dip angle	40
7.5	Relationships between mining variables and sinkhole characteristics	41
7.6	Water conditions and roof conditions.....	44
7.7	Type of mining structure: underground workings versus shafts	45
8.	Analysis of geological and overburden conditions	45
8.1	Geological descriptions of the overburden	45
8.2	Presence of an aquifer in the overburden.....	47
8.3	Links between overburden characteristics and observed sinkholes.....	48
9.	Statistical relationships between variables.....	51
9.1	Scope of the statistical analysis.....	51
9.2	Relationships with sinkhole dimensions	51
9.3	Relationships with post-closure delay.....	52
9.4	Limits.....	52

10.	Main findings from the database analysis	52
11.	General overview	55
12.	The Susceptibility Model	56
12.1	Selection of variables of interest	56
12.2	Material overburden resistance	56
12.3	Imputation of material resistance variable	57
12.4	Depth of the drift relative to the surface	59
12.5	Coal working thickness	61
12.6	Multicollinearity analysis	62
12.7	Balance of susceptibility variables	64
12.8	Normalization process	66
12.9	Susceptibility Score	67
12.10	Susceptibility Model Results	69
13.	The Magnitude Model	70
13.1	Sinkhole characteristics variables	70
13.2	Depth of the sinkhole	71
13.3	Diameter of the sinkhole	71
13.4	Origin of Collapse	72
13.5	Number of sinkholes	73
13.6	Magnitude variable derivation	74
13.7	Selection of conditioning variables	75
13.8	Conditional probability model	76
14.	Hazard Integration Framework	77
15.	Operational Deployment	78
16.	Conclusions and perspectives	81
17.	References	82

List of Figures

Figure 1	Completeness of mining-characteristic variables by country.	15
Figure 2	Median working thickness by country.	16
Figure 3	Median working depth by country.	16
Figure 4	Working accessibility by country.	17
Figure 5	Spatial distribution of sinkholes related to underground workings in the SIRIMA inventory.	18
Figure 6	Distribution of underground-working sinkhole records by country and municipality. ..	19
Figure 7	Ten municipalities with the highest recorded sinkhole density.	19
Figure 8	Ten highest sinkhole densities observed within 1 km ² grid squares.	20
Figure 9	Distribution of mining start dates by historical period for the underground-working sinkhole dataset.	22
Figure 10	Distribution of mining end dates by historical period for the underground-working sinkhole dataset.	23
Figure 11	Decadal distribution of recorded sinkhole occurrences with usable calendar dates ...	24
Figure 12	Annual distribution of recorded sinkhole occurrences. Only dates with sufficient temporal precision are included.	25
Figure 13	Decadal distribution of recorded sinkhole occurrences with usable calendar dates in Germany	26
Figure 14	Decadal distribution of recorded sinkhole occurrences with usable calendar dates in Poland	26
Figure 15	Decadal distribution of recorded sinkhole occurrences with usable calendar dates in France	27
Figure 16	Seasonal distribution of recorded sinkhole occurrences. Only dates with sufficient temporal precision are included.	28
Figure 17	Distribution of sinkhole records according to reported “before” or “after” temporal thresholds (n = 404).	28
Figure 18	Distribution of operational duration for the 123 mines included in the inventory.	29
Figure 19	Sinkhole occurrence relative to mine closure (n = 666).	30
Figure 20	Distribution of length or diameter for sinkholes related to underground workings.	31
Figure 21	Distribution of sinkholes according to the French intensity classes based on collapse diameter.	32
Figure 22	Distribution of width for sinkholes related to underground workings.	33
Figure 23	Distribution of impacted surface area for sinkholes related to underground workings.	34
Figure 24	Distribution of sinkhole depth for sinkholes related to underground workings.	35
Figure 25	Distribution of reported collapse origins for sinkholes related to underground workings.	36
Figure 26	Simplified distribution of reported collapse origins after grouping roof-room and roof-gallery failures.	37
Figure 27	Distribution of underground-working depth below ground level for sinkholes related to underground workings for the different countries.	38
Figure 28	Distribution of working thickness for sinkholes related to underground workings.	39
Figure 29	Distribution of the ratio between working depth and working thickness for documented sinkhole records.	40
Figure 30	Distribution of the dip of the exploited coal layer.	41
Figure 31	Median sinkhole horizontal dimensions by underground-working depth class.	42
Figure 32	Median impacted surface area by underground-working depth class.	43
Figure 33	Median sinkhole depth by underground-working depth class.	43
Figure 34	Median sinkhole length and depth by working-thickness class.	44
Figure 35	Distribution of sinkholes according to water conditions in the mining workings.	45
Figure 36	Distribution of general geological ages assigned to sinkholes related to underground workings.	46
Figure 37	Cross-distribution of cleaned overburden material resistance classes and simplified lithological classes for sinkholes related to underground workings. Only records with both variables documented are included.	47

Figure 38	Distribution of sinkholes according to the reported presence of an aquifer in the overburden.	48
Figure 39	Median impacted surface by overburden-thickness class.	49
Figure 40	Median impacted surface area by overburden resistance class. Only records with both resistance information and impacted surface area documented are included.	50
Figure 41	Median sinkhole depth by overburden resistance class. Only records with both resistance information and sinkhole depth documented are included.	51
Figure 42.	Hazard matrix derived from the integration of susceptibility and magnitude models	55
Figure 43.	Bar plot of discrete overburden material resistance variable	56
Figure 44.	Overburden material resistance distribution post-ML imputation.	59
Figure 45.	Depth of the drift relative to the surface frequency distribution	60
Figure 46.	Distribution and outliers of the working depth variable	60
Figure 47.	Comparison of boxplots for the coal working thickness variable before and after corrections. (a) Original boxplot, and (b) Corrected boxplot.	61
Figure 48.	Distribution of coal working thickness after correction	62
Figure 49.	Correlation matrix of conditioning variables of susceptibility model	63
Figure 50.	Representation of applied imputation strategy	65
Figure 51.	Point-to-point verification of imputed values	66
Figure 52.	Distribution of susceptibility score	67
Figure 53.	Thresholds definition for susceptibility score	68
Figure 54.	Model stability analysis. Thresholds variation across 100 random realizations	68
Figure 55.	Susceptibility model results compared to conditioning variables	69
Figure 56.	Workflow for magnitude model creation	70
Figure 57.	Distribution and outliers of the depth of the sinkhole variable	71
Figure 58.	Distribution and outliers of the diameter of sinkhole variable	72
Figure 59.	Swarm size compared to their diameter of sinkholes	73
Figure 60.	Swarm size compared to their depth of sinkholes	73
Figure 61.	Cramer's V analysis for predictive variables	76
Figure 62.	General workflow for the hazard operational tool deployment	79
Figure 63.	Workflow for parameter settings update	80

List of tables

Table 1	Average database completion rate by country.....	12
Table 2	Variable completion rates for sinkholes related to underground workings.....	13
Table 3	Availability and summary statistics of mining start dates for the underground-working sinkhole dataset.....	21
Table 4	Availability and summary statistics of mining end dates for the underground-working sinkhole dataset.....	22
Table 5	Availability and temporal precision of underground-working sinkhole occurrence dates.....	23
Table 6	Distribution of sinkhole occurrence relative to mine closure, based on 666 events with usable occurrence and closure dates.	30
Table 7	Sinkhole intensity classes based on collapse diameter, following Franck et al. (2019). .	32
Table 8	Median sinkhole characteristics by underground-working depth class.....	41
Table 9	Cross-distribution of general geological age and simplified lithological classes for sinkholes related to underground workings.....	46
Table 10	Distribution of sinkholes according to the reported presence or absence of an aquifer in the overburden, by country. Only records for which the aquifer field is documented as “Yes” or “No” are included. 48	
Table 11	Keywords count for every overburden material resistance level.	57
Table 12	Serial clustering keywords and average resistance calculations	58
Table 13	Criteria for the Variance Inflation Factor	63
Table 14	VIF results for the susceptibility variables	63
Table 15	Completeness of conditioning variables of the susceptibility model	64
Table 16	Completeness of conditioning variables after imputation.....	65
Table 17	Descriptive statistics for model stability analysis	69
Table 18	Descriptive statistics of the depth of the sinkhole variable.....	71
Table 19	Descriptive statistics of the diameter of the sinkhole variable.....	72
Table 20	Median value of diameter and depth of the sinkhole related to the origin of collapse	73
Table 21	Comparison of swarm of sinkholes with the origin of collapse.....	74
Table 22	Defined thresholds for the Magnitude of sinkhole variable	74
Table 23	Thresholds used to convert predictor variables into categorical classes.....	75
Table 24	Conditional probability matrix for coal working thickness and coal seam dip angle relative to magnitude classes.....	77

Part 1
Sinkhole Database Description and Analysis

1. Introduction

The SIRIMA project aims to reduce the risk associated with sudden discontinuous ground movements, particularly sinkholes, in European post-mining regions. Its purpose is to develop operational methods that can support the assessment and management of sinkhole hazards in areas affected by shallow abandoned coal mine workings.

Work Package 2 established the common data foundation for the project. It was dedicated to building a shared European database of sinkholes linked to shallow abandoned coal mine workings, including events associated with old mining shafts. The work covered the definition of the database structure, the selection of variables to be collected, the harmonisation of partner contributions, and the compilation of national datasets into a single operational inventory.

The objective of WP2 was therefore practical: to collect as many documented sinkhole cases as possible from the partner countries, while retaining the information needed to describe their location, geological and mining context, geometry, observed characteristics and consequences.

The first stage of the work consisted in defining the information to be recorded. INIERIS identified the parameters required to describe both the conditions that may favour sinkhole occurrence and the intensity of the resulting surface disorder. These parameters include quantitative information, such as dimensions and depths of the sinkhole and of the mining works, and qualitative information, such as lithology or overburden resistance. To support consistent data entry across partners, explanatory glossaries and guidance material were prepared. The final database structure was organised into five thematic groups and validated by the project partners in December 2024.

The second stage of the work was the construction of the database. Each partner provided the data available for its territory. By the time Deliverable 2.2 (Sinkhole database) was prepared in December 2025, the database within WP2 had been structured, documented, validated and populated with 1814 sinkhole records from three European countries (France, Poland and Germany). These records were organised into two main populations: sinkholes related to drifts or underground workings, and sinkholes related to shafts (WP2 Deliverable 2.2: Sinkhole database, December 2025).

This inventory forms the input dataset for the subsequent stages of the project, including the analyses presented in this report.

Within this framework, WP4 addresses the assessment and management of risks associated with ground movements in post-mining areas. It combines remote sensing, geotechnical interpretation, vulnerability analysis and risk modelling. Its main objectives are to develop an operational tool for sinkhole hazard prediction, perform InSAR analyses on selected test sites, assess the effects of ground movements on buildings and critical infrastructure, develop fragility functions, and implement risk assessment workflows within the VIGIRISKS platform developed by BRGM.

Task 4.1 focuses on the development of an operational methodology for predicting sinkhole hazards in European abandoned shallow coal mining areas. The task relies primarily on the database compiled during WP2, with particular attention to data quality, heterogeneity and missing information, since these factors directly control the robustness of any statistical interpretation.

The task examines how sinkhole events relate to their mining, geological and hydrogeological setting. The objective is intentionally modest. At this stage, the database can be used to identify variables that are sufficiently documented and relevant for hazard assessment, but it is not robust enough to establish causal mechanisms. This deliverable should therefore be considered as an upstream qualification step of the database.

It describes the preparation of the database, including cleaning, harmonisation and the definition of the populations used for analysis. It then assesses data completeness and identifies the variables that can be used with reasonable confidence. The report then examines the spatial, temporal, geometrical, mining-related and geological characteristics of the recorded sinkholes, with particular attention to the limits these data impose on operational hazard modelling.

2. Data processing and preparation for hazard modelling

2.1 Scope and structure of the input dataset

The input database is organised into five thematic groups that describe the event, its mining and geological context, and its observed consequences:

- Location includes the information required to identify and locate each event: identifiers, names, coordinate systems, mine name, country and municipality.
- Sinkhole characteristics describe the surface disorder itself, including the number of sinkholes, impacted surface area, dimensions, depth, date of appearance and presumed origin of collapse.
- Mining-work characteristics link the observed disorder to the underground context, including the type of structure, mining period, dip of the coal seam, working depth, support conditions.
- Overburden characteristics describe the geological layers between the mine and the surface, including lithology, roof fracturing, material resistance and water conditions.
- Surface consequences record the presence of buildings or infrastructure and any reported damage.

The description of the principal variables is provided in the appendix.

The database was implemented in MS Excel, with dedicated sheets for data entry, field definitions and explanatory diagrams. These support materials were intended to help partners complete the records in a consistent way.

The database includes records from three European countries with a long history of shallow coal mining and documented sinkhole events. It is not exhaustive. The inventory was shaped by the availability of historical archives, the quality of source documents, national reporting practices and restrictions on access to sensitive data. These constraints differ between countries and have to be kept in mind when interpreting the results of the analyses.

In Germany, the contribution was provided by DMT-THGA and focused on the Ruhr coal basin, particularly the Essen area. The main limitation was the restricted access to sinkhole data, which are considered sensitive, as well as the frequent difficulty in distinguishing between natural sinkholes and mining-induced collapses. As a result, 398 records were integrated into the database; however, most records contained only limited information, mainly the location of the sinkholes.

In Poland, the Central Mining Institute compiled more than 906 records from the Upper Silesian Coal Basin, primarily associated with the abandoned Siersza, Kazimierz-Juliusz and Jan Kanty mines. The dataset was built from archival mining documents and complemented by more recent inventories, including LiDAR and satellite-based observations. Nevertheless, significant historical gaps remain, particularly for the nineteenth and early twentieth centuries, and many older records lack detailed information on mine geometry or clear evidence linking the sinkhole to a specific underground void.

In France, 513 sinkhole records affecting approximately fifty coal mines were collected, with a high concentration of cases in the Messeix basin. The information was compiled from the GEODERIS database, hazard assessment studies, expert reports and former operators' archives. The main challenges were related to the date of many events, the heterogeneity of the available sources, and the frequent lack of detailed technical information regarding underground workings, geological conditions and overburden characteristics.

2.2 Data cleaning and harmonisation

The initial database contained 1815 sinkhole records and required substantial preparation before analysis. Original identifiers were preserved throughout the process to maintain traceability between the source files and the cleaned analytical dataset.

The first cleaning step standardised the structure and content of the records. Capitalisation, leading and trailing spaces, spelling variants and heterogeneous codes for missing values were harmonised.

In total, 9262 entries coded as “unknown”, “unknow”, “?”, or “non disponible” were reclassified as missing values, so that they would not be treated as valid analytical categories. Categorical values were standardised without deleting the original observations or their bibliographic references.

Although predefined drop-down menus were provided for most categorical variables to guide data entry and improve consistency, these controlled lists were not systematically used by the operators filling in the database. As a result, several fields contained free-text entries, spelling variants or additional categories that required manual review and harmonisation before analysis.

Date formatting required extensive processing due to high heterogeneity, including exact dates, single years, approximate periods, intervals, and multi-date entries (e.g., “1958, 1991”, “1905-1940”, and “08.2011-04.2019”). These entries were interpreted consistently while preserving their original meaning and uncertainty. When a source record indicated two distinct occurrence dates, it was split into two separate observations. Consequently, 54 identifiers appear twice, increasing the dataset from 1815 source records to 1868 observations without generating new source identifiers.

Mining structures were then classified to distinguish sinkholes associated with underground workings from those related to shafts. This distinction is essential, as both groups involve different collapse mechanisms and cannot be analysed together.

The final analytical datasets comprise 1725 observations for underground workings and 121 confirmed shaft-related events.

Numerical data standardisation involved decimal separators, units, inequalities, ranges and text strings across parameters such as sinkhole dimensions, impacted surface area, working thickness (coal seam), shaft depth and the dip of the exploited layer. For numerical variables provided as ranges or inequalities, the original information was preserved for traceability, while a standardised numerical value was derived for analysis. In most cases, ranges were converted to their mean value, whereas inequalities were retained as threshold information. These derived values should be interpreted as analytical conventions rather than reconstructed event dates. They allow grouped comparisons but do not remove the original temporal or numerical uncertainty.

Impacted surface areas were sparsely populated in the source database, with only 41 direct values reported. Values were recalculated using usable dimensions—multiplying length by width for rectangular features and applying circular formulas when only a diameter was available. This yielded 1239 usable surface values in the final datasets (454 rectangular and 785 circular), all cross-checked against their corresponding dimensions.

Geological descriptions were reviewed to correct spelling inconsistencies and standardise terminology. Original descriptions were retained, with a clear distinction maintained between missing data and documented geology. Geological information is available for 1338 underground-working sinkholes and 115 shaft-related events.

Finally, coordinate data were verified for consistency between declared projected coordinate systems and geographical coordinates, enabling reliable spatial comparisons without mixing incompatible national grids.

This harmonisation resolved the main formatting inconsistencies, clarified date information and reduced terminology noise while preserving the source data. The cleaned dataset can therefore be used for the spatial, temporal and statistical analyses presented in the following sections. Following the cleaning and harmonisation process, the final analytical database comprises 1 868 observations, including 516 French records, 397 German records and 955 Polish records.

2.3 Definition of analytical populations

After cleaning, three analytical datasets were defined. The complete database includes all recorded sinkhole observations. It was then divided into two main analytical populations according to the associated mining structure: sinkholes related to underground workings, including isolated galleries and room-and-pillar excavations, and sinkholes related to mine shafts. This distinction is essential

because the two populations involve different underground structures, failure mechanisms and surface expressions.

Depending on the variable considered, only records with sufficient and reliable information were retained. As a result, the number of observations varies from one analysis to another.

3. Overview of data completeness and exploitable variables

3.1 Completeness of the main variables

Table 2 summarises data availability for the 1725 sinkholes related to underground workings. Identification, location and mining-site variables are almost complete, with completion rates close to 100%. Bibliographic references are also well documented (96.1%), as are mining end dates (91.0%). Several variables used in the subsequent analyses show moderate coverage, generally between 60% and 80%. Other technical fields are much less complete, particularly roof fracture condition (1.0%), ground support (5.7%) and observed damage (0.2%).

Table 1 presents the average database completion rate across the 47 columns for each country. Poland has the highest average rate, at 67.1%, followed by France at 63.0%, whereas Germany has a much lower rate of 36.8%. These results indicate that the Polish and French records are generally more complete than the German records. However, this indicator measures overall data availability rather than data quality and may hide major differences in the completion of individual variables. For example, Poland has much better coverage for working depth (98.8%, compared with 84.0% in France) and working thickness (92.3%, compared with 58.3%). Conversely, France provides more information on void density (62.6%, compared with 0% in Poland) and also contains information on opening width, uncertainty and ground support, which are almost entirely absent from the Polish subset. The example of working thickness is discussed in the next section. These differences between countries may reflect different approaches to sinkhole description and management.

Table 1 Average database completion rate by country

Country	Average completion rate
Poland	67,1 %
France	63,0 %
Germany	36,8 %

Table 2 Variable completion rates for sinkholes related to underground workings.

Categories	Variable	Filled values	Missing values	Completion rate
Location (12 variables)	ID BD SINKHOLE	1725	1	99,9%
	Name	1335	391	77,3%
	ID Sinkhole	1313	413	76,1%
	Coordinate X	1725	1	99,9%
	Coordinate Y	1725	1	99,9%
	Projected Coordinate System	1725	1	99,9%
	X_WGS84	1725	1	99,9%
	Y_WGS84	1725	1	99,9%
	Measurement uncertainty (m)	304	1422	17,6%
	Name of the mine	1725	1	99,9%
	Country	1725	1	99,9%
	Municipality	1725	1	99,9%
Sinkhole characteristics (10 variables)	Number of sinkholes	1717	9	99,5%
	Impacted surface (m ²)	1126	600	65,2%
	Length / diameter (m)	1129	597	65,4%
	Width (m)	397	1329	23,0%
	Depth of the sinkhole (m)	1416	310	82,0%
	In-situ measurement or estimation	1170	556	67,8%
	Date of appearance of the sinkhole	1190	536	68,9%
	Date_of_appearance_of_the_sinkhole_precision	1189	537	68,9%
	Uncertainty of date (years)	231	1495	13,4%

	Origin of collapse	1211	515	70,2%
	Similar phenomena known nearby	1327	399	76,9%
<i>Characteristics of mining work (shaft/cavity-drift)</i> (12 variables)	Start date of the mining	1366	360	79,1%
	End date of the mining	1571	155	91,0%
	Accuracy of date	261	1465	15,1%
	Dip of the exploited layer (degree)	1182	544	68,5%
	Type of mining work (shaft/cavity-drift)	601	1125	34,8%
	Width / diameter	48	1678	2,8%
	Working thickness / Shaft depth (m)	1092	634	63,3%
	Void density	263	1463	15,2%
	Depth of the drift relative to the surface (m)	1260	466	73,0%
	Uncertainty of depth (m)	48	1678	2,8%
	Ground support	98	1628	5,7%
	Accessibility	1146	580	66,4%
<i>Overburden characteristics</i> (5 variables)	Geology of the overburden	1338	388	77,5%
	Roof fracture condition	17	1709	1,0%
	Material overburden resistance	384	1342	22,2%
	Water conditions in the mining workings	1186	540	68,7%
	Presence of an aquifer in the overburden	887	839	51,4%
<i>Surface occupation & consequences</i> (2 variables)	Presence of buildings or infrastructure	1153	573	66,8%
	Damage observed	3	1723	0,2%
<i>Others</i> (2 variables)	Observations	585	1141	33,9%
	Bibliographic references	1659	67	96,1%

3.2 Heterogeneous completeness by country: example with the working characteristics

The database excluding shafts contains 1725 sinkhole records distributed across three countries: Poland accounts for 917 records (53.2%), France for 420 (24.3%), and Germany for 388 (22.5%). However, the number of records alone does not reflect the analytical information available for each country. The completeness of mining-characteristic variables differs strongly among the national subsets (Figure 1), indicating that the database combines sources with different structures, reporting practices and levels of detail.

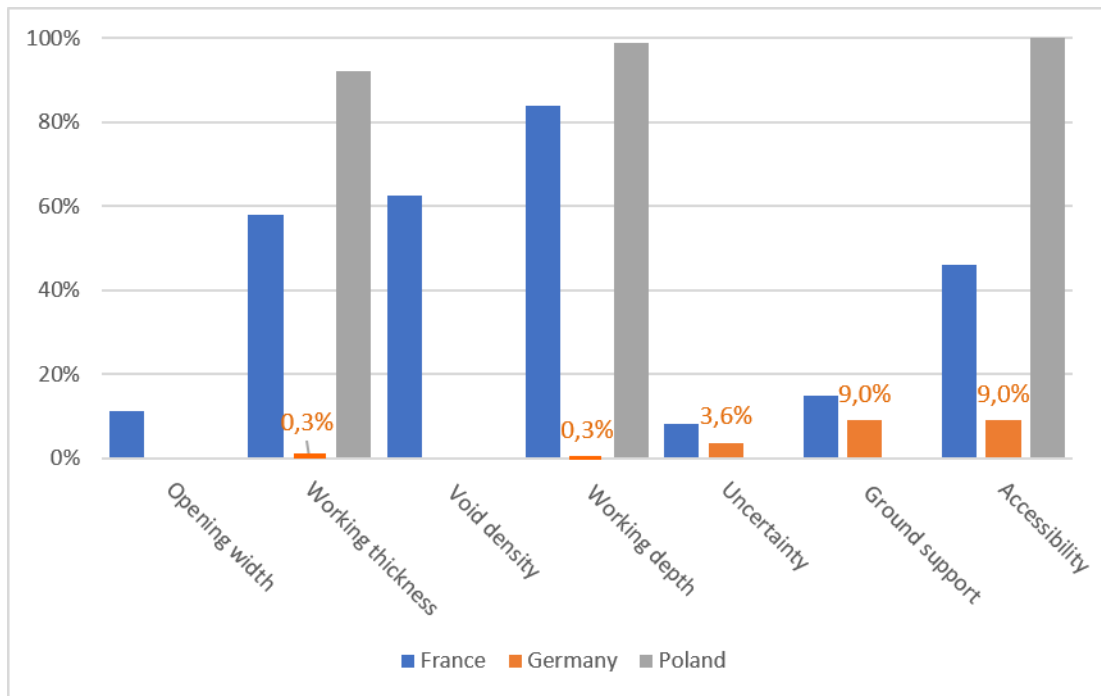


Figure 1 Completeness of mining-characteristic variables by country.

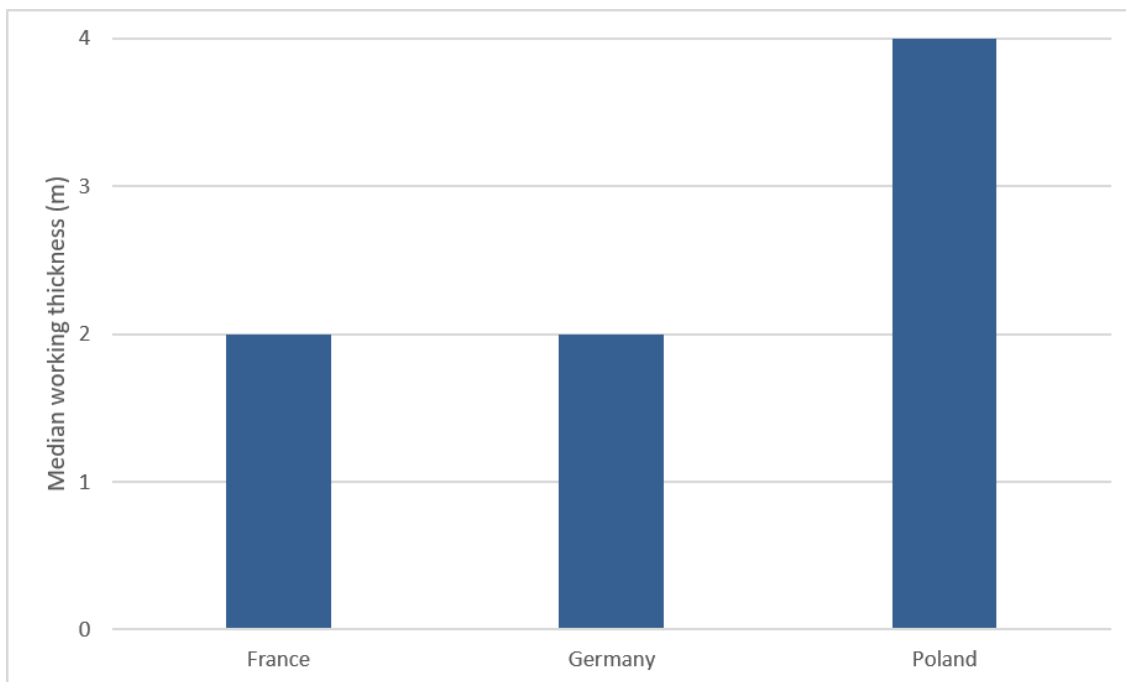


Figure 2 Median working thickness by country.

The Polish subset is highly complete for working thickness (92.3%), working depth (98.8%) and accessibility (100%). Among the documented observations, Polish workings have a median thickness of 4 m (Figure 2) and a median depth of 45 m (Figure 3). Accessibility is recorded as “no” for all Polish cases (Figure 4), meaning that the former workings are reported as not accessible. In contrast, the Polish records contain no usable information on opening width, void density, uncertainty or ground support. Poland therefore provides a strong basis for analysing thickness and depth, but not for comparing the full set of mining characteristics.

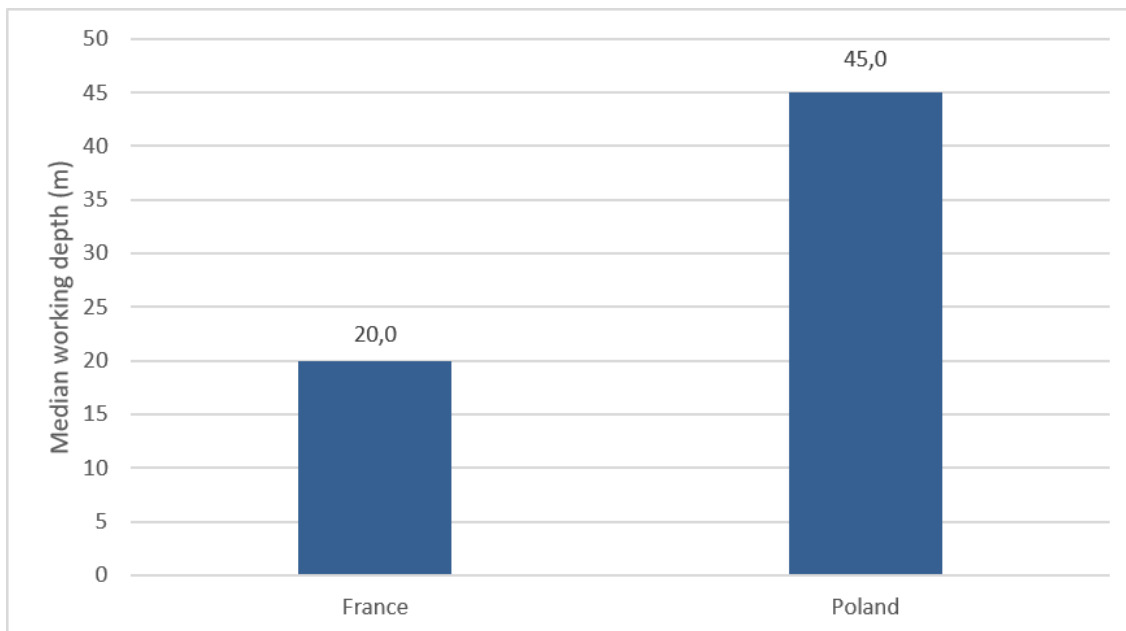


Figure 3 Median working depth by country.

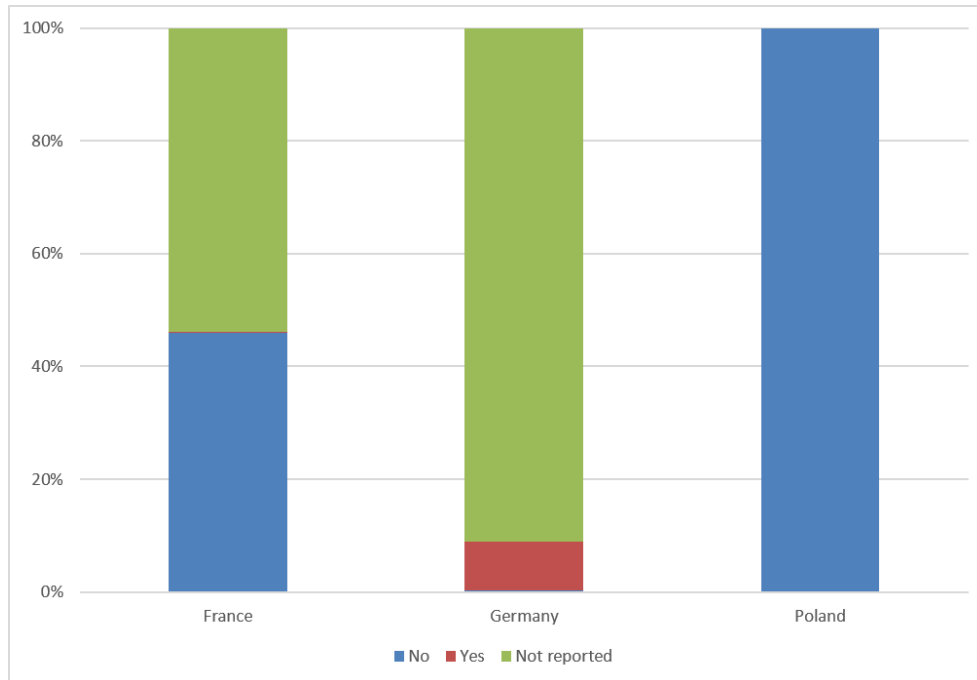


Figure 4 Working accessibility by country.

The French subset contains information for more types of mining-characteristic variables, although completeness remains heterogeneous across variables (Figure 1). Working depth is reported for 84.0% of French records and has a median value of 20 m (Figure 3), while working thickness is available for 58.3% and has a median of 2 m (Figure 2). France is also the only country for which void-density information is sufficiently documented to be described, including block caving, drifts, and room-and-pillar configurations. Accessibility is documented for 46.2% of French cases (Figure 4), whereas opening width, uncertainty and ground support are reported much less frequently. The French data therefore describe more types of characteristics, but missing values remain an important constraint.

The German subset is the least complete for these variables (Figure 1). Coverage reaches only 9.0% for ground support and accessibility, 3.6% for uncertainty, and approximately 0% for opening width, void density, working thickness and working depth. Although a median thickness of 2 m appears in Figure 2, it is based on only one documented observation and cannot be considered representative of German mines. The absence of a German median in Figure 3 similarly reflects the lack of usable depth information, not a depth value of zero. Quantitative comparisons involving Germany are therefore mainly limited by the sparse documentation of mining-characteristic variables.

Apparent differences between countries may reflect real differences in mining conditions, but they may also result from differences in data availability, especially for older mines where archival information is incomplete or missing. They are also strongly influenced by how each national inventory was compiled. Comparisons are therefore meaningful only for variables that are sufficiently documented in each country and should be reported together with the number and proportion of documented observations, as illustrated in Figure 1. Missing information must not be treated as the absence of a characteristic. For this reason, multivariable models, country rankings and causal interpretations would require specific control for source effects and unequal data completeness.

3.3 Selection of variables for the analysis

The completion rates show that the database is reliable for describing sinkhole locations and their general mining context. Coordinates, mine names, countries and municipalities are almost always available, which makes the dataset suitable for spatial analysis and for comparisons between mining basins. Several variables describing the sinkholes themselves are also sufficiently documented for

further analysis, including the number of sinkholes, sinkhole depth, impacted surface area, date of occurrence and reported origin of collapse. The level of detail, however, remains variable from one record to another.

By contrast, some variables are too poorly filled to support a robust statistical analysis. This is particularly true for roof fracture conditions, ground support, void density and observed damage. These fields may contain useful information for individual cases, but they cannot be treated as representative of the full population. Their low completion rates mainly reflect the nature of the available sources. Many records come from historical archives or former mining documents, where the description of the underground geometry and the condition of the overburden were often limited or not recorded at all.

The selection of variables used in the rest of the work was therefore based on both technical relevance and data availability. Variables with high or moderate completion rates were retained for quantitative or comparative analyses when their definitions were sufficiently consistent across countries. Variables with very low completion rates were kept only as descriptive information or excluded from analysis. This limits the influence of poorly documented data and avoids overinterpreting small or unrepresentative subsets.

4. Spatial and geographical analysis of sinkhole records

4.1 Distribution by country

The spatial distribution of the 1725 sinkholes recorded in the three countries covers 60 municipalities. It primarily reflects the current state of the available inventories rather than the actual geographical distribution of hazard. The dataset is strongly concentrated in Poland, which accounts for 917 records (53.2%), particularly in Trzebinia and Jaworzno. Germany represents 388 records, most of them located in Essen. This concentration needs to be read in relation to the size and definition of the mapped areas: Essen is a much larger municipality (210.37 km²) than Trzebinia, which covers 31.3 km². France shows a more spread-out pattern, with 420 records distributed over a larger number of sites. Figure 5 and Figure 6 illustrate this contrast between concentrated and more dispersed national inventories.

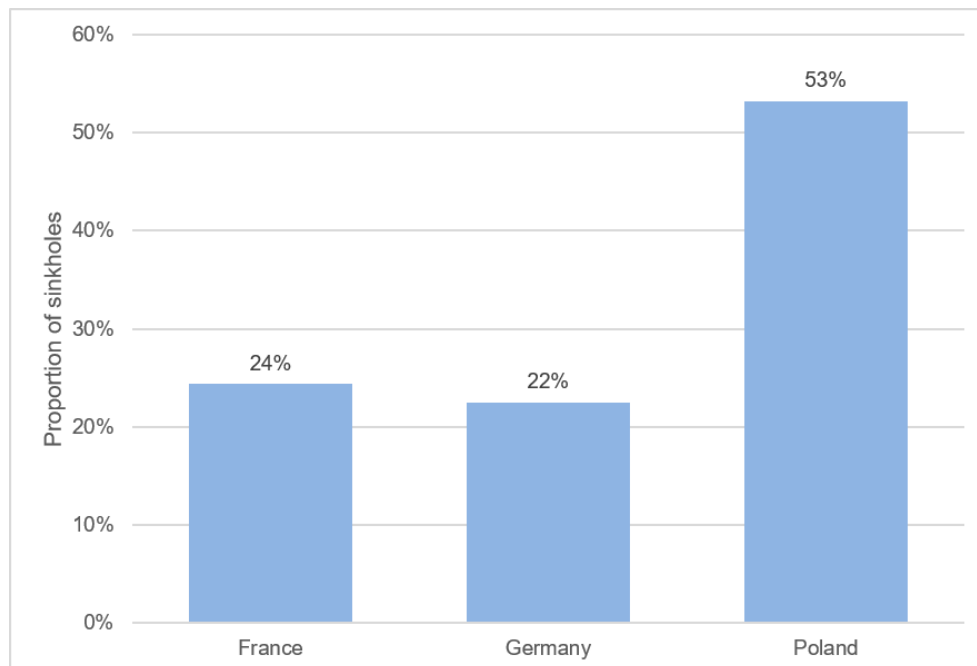


Figure 5 *Spatial distribution of sinkholes related to underground workings in the SIRIMA inventory.*

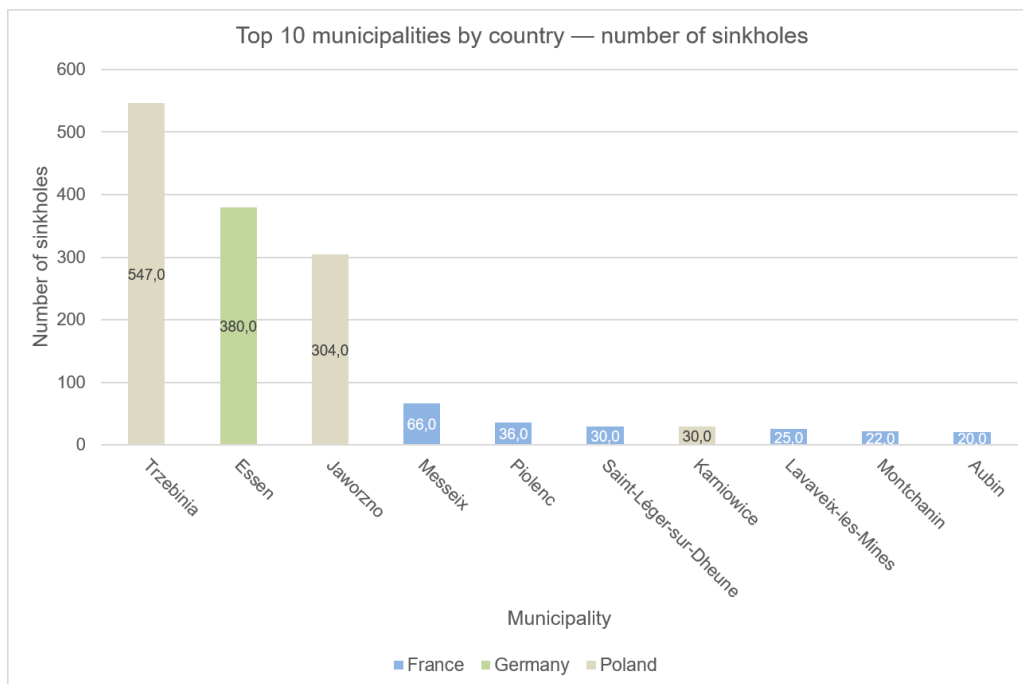


Figure 6 *Distribution of underground-working sinkhole records by country and municipality.*

When sinkhole density is examined at the municipal scale, the ranking changes, largely because the municipalities differ greatly in surface area (Figure 7). The density is much higher in a small municipality such as Lavaveix-les-Mines than in Essen, which covers a much larger area and has a density of less than two sinkholes per km².

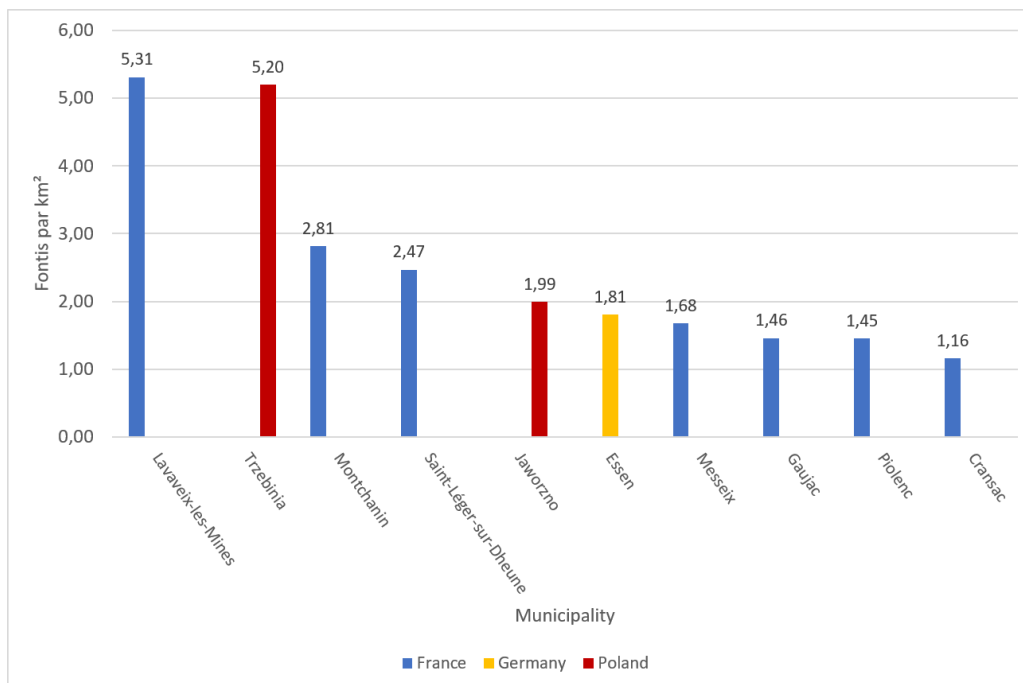


Figure 7 *Ten municipalities with the highest recorded sinkhole density.*

4.2 Spatial distribution of sinkholes

The spatial distribution of recorded sinkholes was analysed at a finer scale using a regular 1 km² grid. This approach makes it possible to quantify the degree of spatial clustering within the inventory and to identify areas where records are concentrated. The grid was generated in the metric coordinate reference system appropriate for each country: Lambert-93 for France, UTM 32N for Germany and UTM 34N for Poland. Two atypical Polish coordinates were identified and nevertheless retained in the calculation.

A total of 268 occupied grid squares, defined as squares containing at least one recorded sinkhole, were identified. The results show a strong spatial concentration of the inventory. Among these occupied squares, 37 contain at least ten recorded sinkholes. The ten densest squares account for 41% of all records, with a maximum of 263 sinkhole occurrences in a single 1 km² square (Figure 8). This high concentration identifies areas where the inventory is particularly dense and where the source records should be examined in more detail. This concentration may partly be related to differences in site and mining conditions.

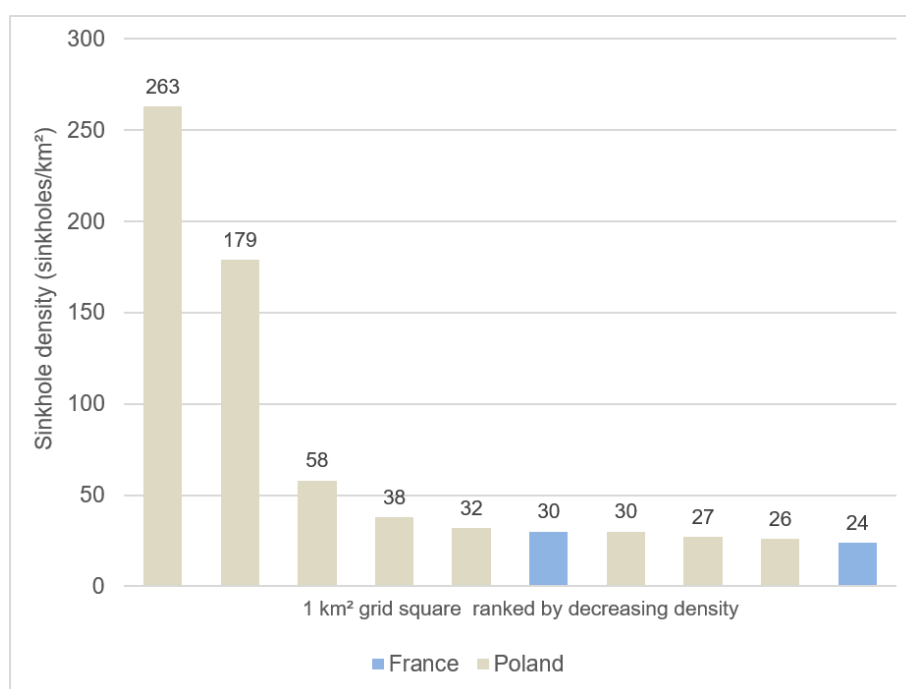


Figure 8 Ten highest sinkhole densities observed within 1 km² grid squares.

4.3 Representativeness and limitations of the inventory

The inventory provides a valuable basis for analysing sinkholes associated with shallow coal mining, but it is not an exhaustive record of all events. The availability and level of detail of the information vary strongly between countries and mining areas. These differences reflect archive quality, reporting practices, access to data and the age of the mining works. In some cases, information has been lost over time or was never recorded in a structured way.

The spatial distribution of the records is therefore shaped by the way the database was built. The Polish and German contributions mainly focus on selected mining basins where sinkholes are numerous, recurrent or well documented. The French records are more dispersed and were mainly retained where usable information was available. This introduces a selection bias: the best-documented or most problematic areas are likely to be overrepresented. The database should consequently be read as a structured inventory of available information, not as a complete map of sinkhole hazard. These aspects are discussed in detail in (Vuidart et al., 2026).

5. Temporal analysis of sinkhole occurrence and mining periods

Temporal information in the database was analysed to describe when sinkholes occurred and to compare those dates with the documented mining periods. The purpose was to distinguish, where possible, between events that occurred during active mining, shortly after mine closure, and several decades later. Particular attention was paid to uncertainty in the source data, since many records include approximate years, intervals, incomplete dates or relative indications such as “before” and “after”.

5.1 Availability and reliability of temporal information

Temporal information is not documented with the same level of completeness or precision across the dataset. The date of sinkhole occurrence is available for about 68.9% of records. Mining start and end dates are better documented, with completion rates of approximately 79.1% and 91.0%, respectively. Table 3 and Table 4 summarise the availability and main descriptive statistics for mining dates, while Figure 9 and Figure 10 show their distribution by historical period.

Table 3 Availability and summary statistics of mining start dates for the underground-working sinkhole dataset.

	Value	Part
Total sinkholes	1725	100 %
Rows with start date	1366	79.2%
Rows without start date	359	20.8%
Earliest mining start date	1575	
Latest mining start date	1921	
Average start year	1860	

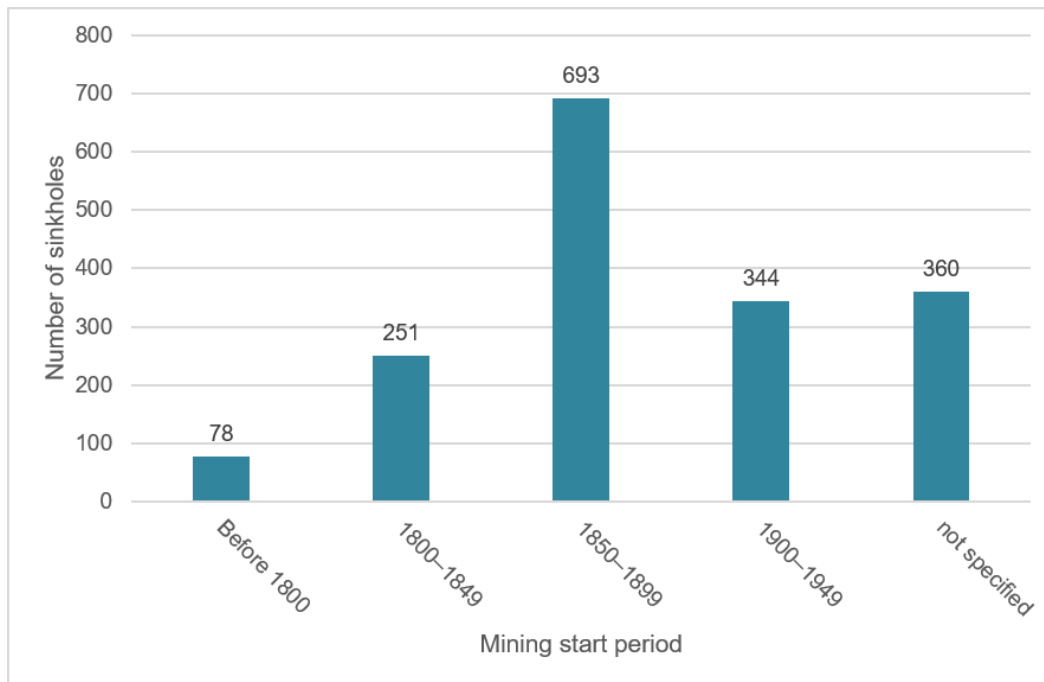


Figure 9 Distribution of mining start dates by historical period for the underground-working sinkhole dataset.

Table 4 Availability and summary statistics of mining end dates for the underground-working sinkhole dataset.

	Value	Part
Total sinkholes	1725	100 %
Rows with end date	1528	88.6%
Rows without end date	197	11.4%
Earliest mining end date	1844	
Latest mining end date	2015	
Average end year	1973	

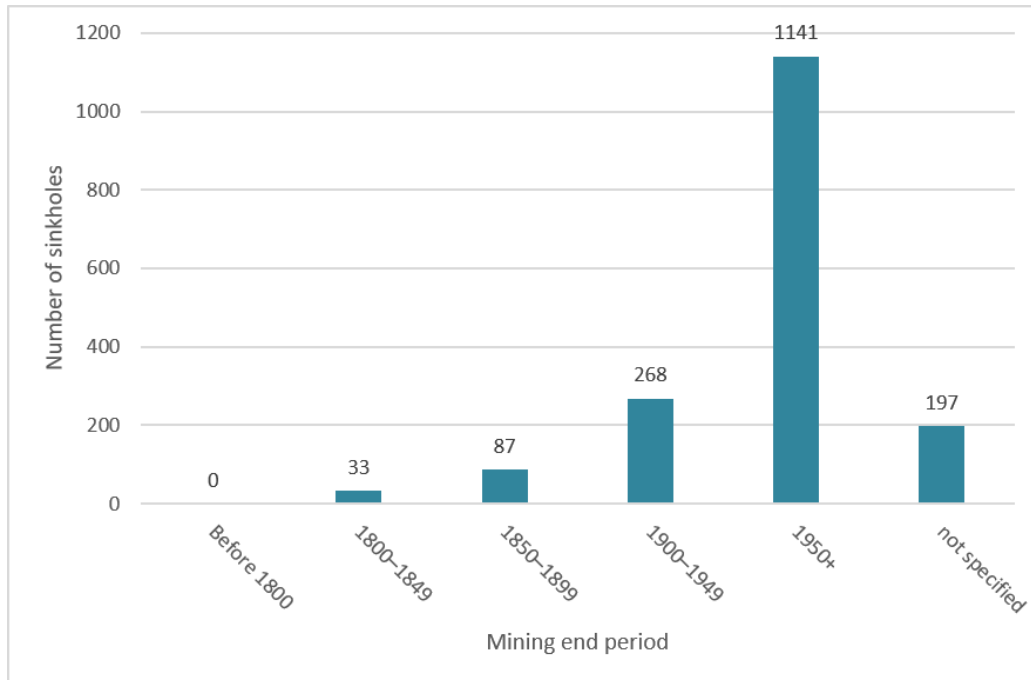


Figure 10 Distribution of mining end dates by historical period for the underground-working sinkhole dataset.

The heterogeneity of temporal information affects the level of analysis that can be performed. Exact dates, approximate years, intervals and relative indications do not carry the same precision. The dataset is therefore suitable for identifying broad temporal patterns and for comparing documented periods, but it cannot support a detailed year-by-year reconstruction of sinkhole occurrence. The comparison between sinkhole occurrence dates and mining periods is presented separately below.

5.2 Distribution of sinkhole occurrence over time

Occurrence timing is available for 1,190 of the 1,725 records, corresponding to 69.0% of the underground-working dataset. Only 785 records, however, can be assigned directly to a calendar year. The remaining documented dates include 404 relative indications, such as “before” or “after”, which were excluded from the annual, decadal and seasonal distributions. Table 5 summarises the availability and precision of occurrence dates.

Table 5 Availability and temporal precision of underground-working sinkhole occurrence dates.

	Count	Part	Interpretation
Total sinkholes	1725	100,0%	
Occurrence recorded	1190	69,0%	Includes exact and relative dates
Usable calendar date	785	45,5%	Precision: year, month or day
Day precision	358	20,8%	Highest temporal precision
Month precision	323	18,7%	Assigned to the corresponding year
Year precision	104	6,0%	Suitable for annual grouping

Relative (before/after)	date	404	23,4%	Excluded from year-based charts		
Missing or unusable		536	31,1%	No direct assignment	calendar-year	

For all the inventories at the decadal scale, recorded occurrences are mainly concentrated between 1980 and 2009. The 2000-2009 decade contains the highest number of cases, with 197 occurrences, or 25.1% of usable dates. It is followed by 1980-1989, with 174 cases (22.2%), and 1990-1999, with 115 cases (14.6%). Together, these three decades account for 61.9% of the occurrences that can be assigned to a calendar year (Figure 11). At the annual scale, the most visible peaks occur in 1985 (45 cases), 2022 (42 cases) and 2019 (37 cases) (Figure 12). These peaks should not be overinterpreted: the 1985 and 2019 peaks are entirely attributable to Polish records.

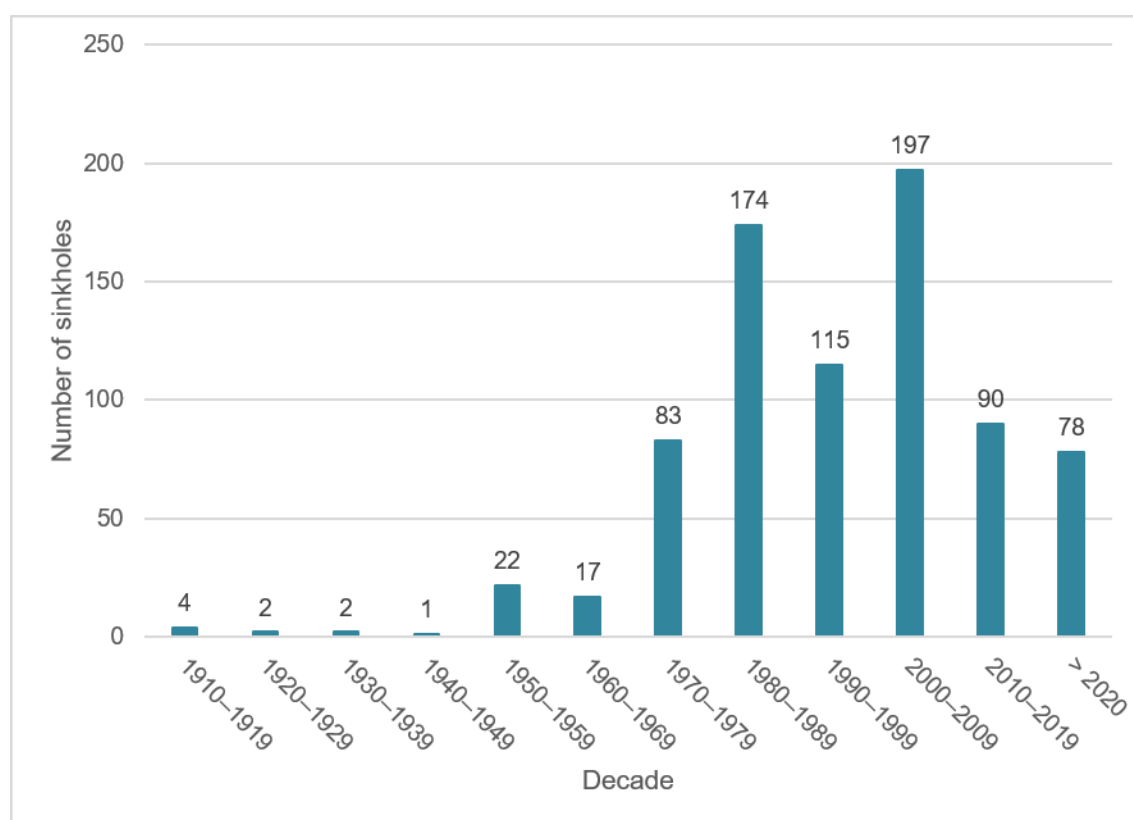


Figure 11 Decadal distribution of recorded sinkhole occurrences with usable calendar dates

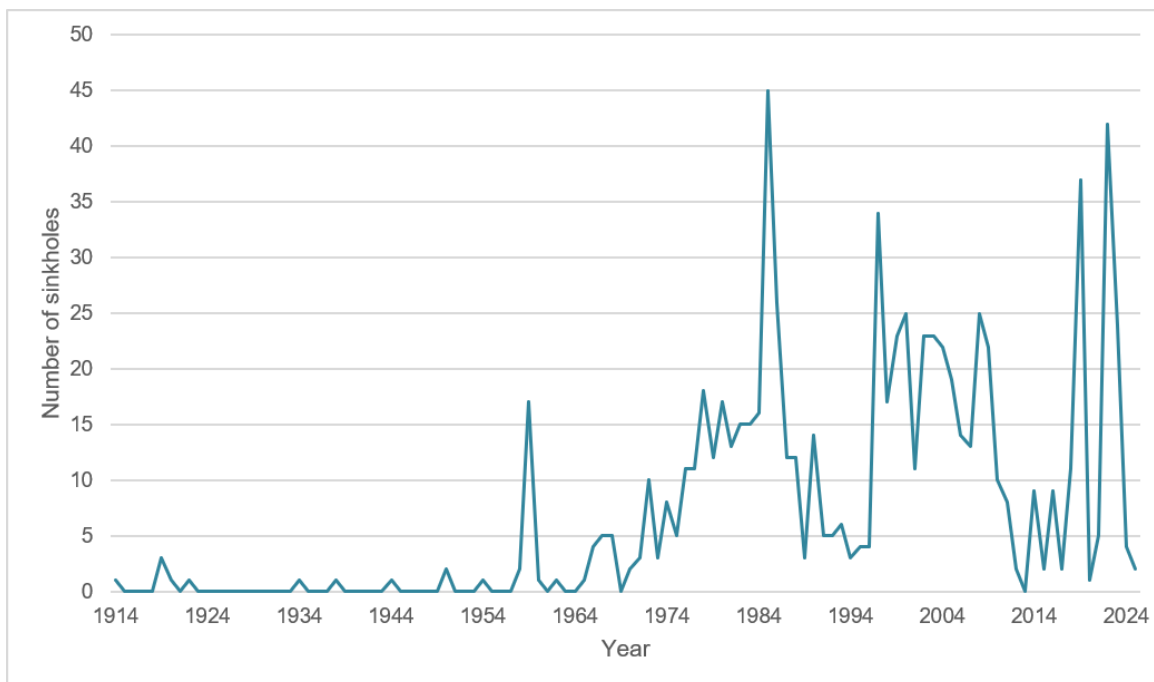


Figure 12 *Annual distribution of recorded sinkhole occurrences. Only dates with sufficient temporal precision are included.*

To check whether the overall temporal distribution is shared by the three countries or mainly reflects the weight of one national dataset, the analysis was repeated country by country. The decadal distribution of the 785 usable calendar dates shows clearly different temporal patterns between countries. In Germany, 174 of the 218 dated occurrences (79.8%) are concentrated in the 2000-2009 decade, indicating a strong temporal concentration (Figure 13). In Poland, the sample is larger ($n = 505$) and the distribution is more spread out: the highest number of cases is observed in 1980-1989, with 172 records (34.1%), while 24.0% of occurrences are recorded between 2010 and 2025 (Figure 14). In France, the number of usable dates is lower ($n = 62$) and the pattern is more recent, with a maximum in 2010-2019 (20 cases; 32.3%) and 40.3% of occurrences recorded between 2010 and 2025 (Figure 15). This is probably linked to the creation of GEODERIS, the increasing involvement of GEODERIS by the DREALs over time, and the growing number of hazard studies.

Overall, the country-by-country analysis shows that the apparent temporal pattern observed for the full dataset is not uniform across national subsets but results probably from the combined effect of different inventory sizes, documentation periods and reporting practices.

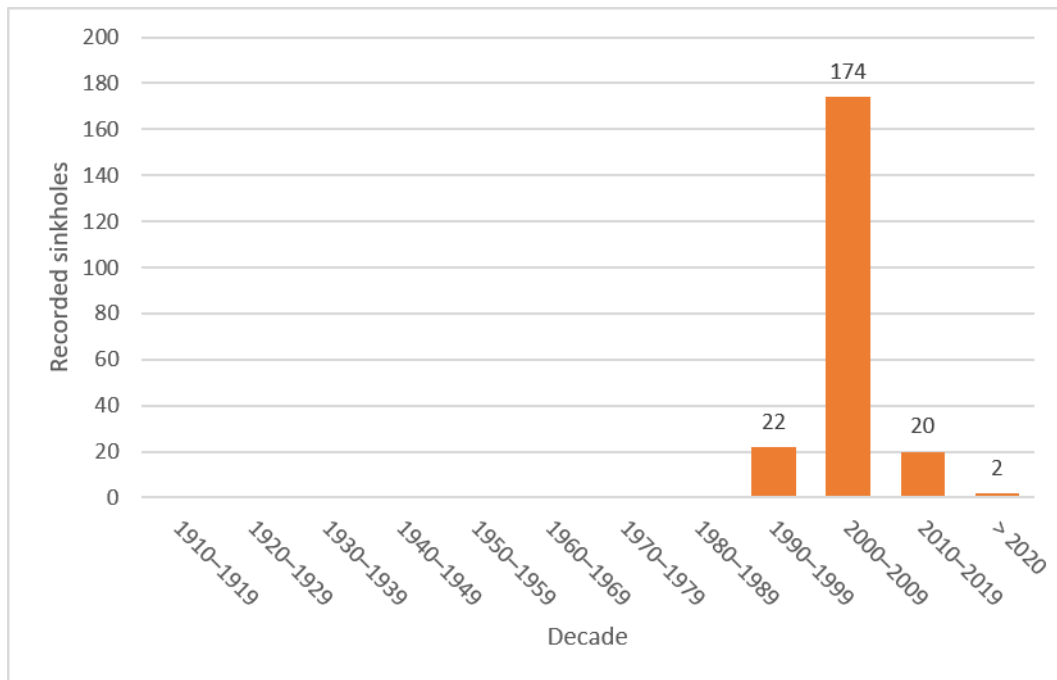


Figure 13 Decadal distribution of recorded sinkhole occurrences with usable calendar dates in Germany

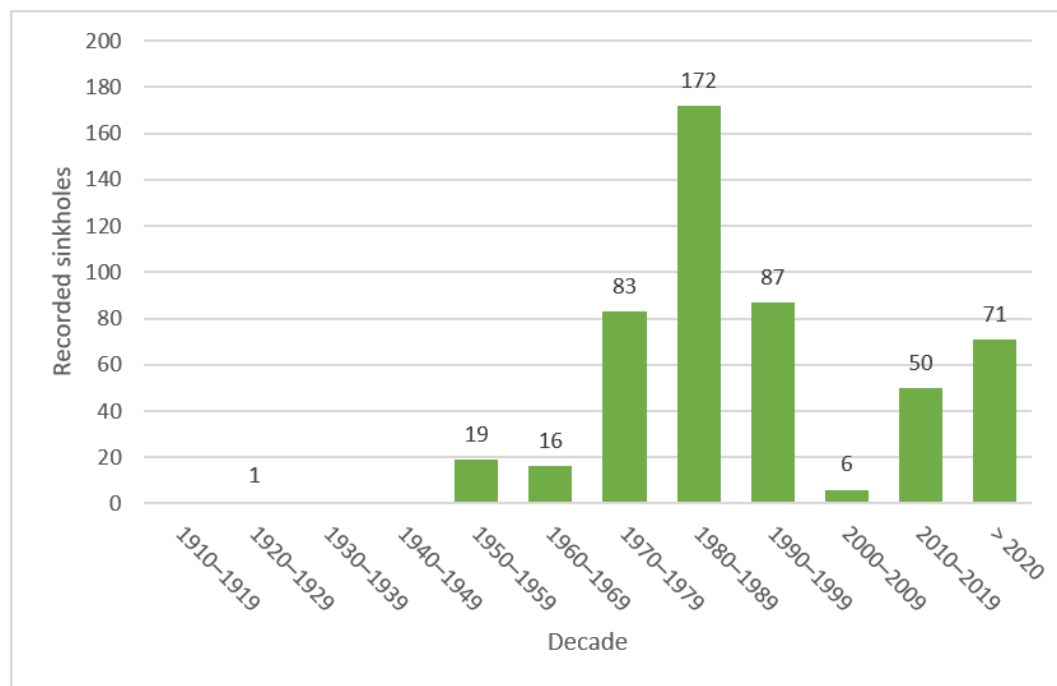


Figure 14 Decadal distribution of recorded sinkhole occurrences with usable calendar dates in Poland

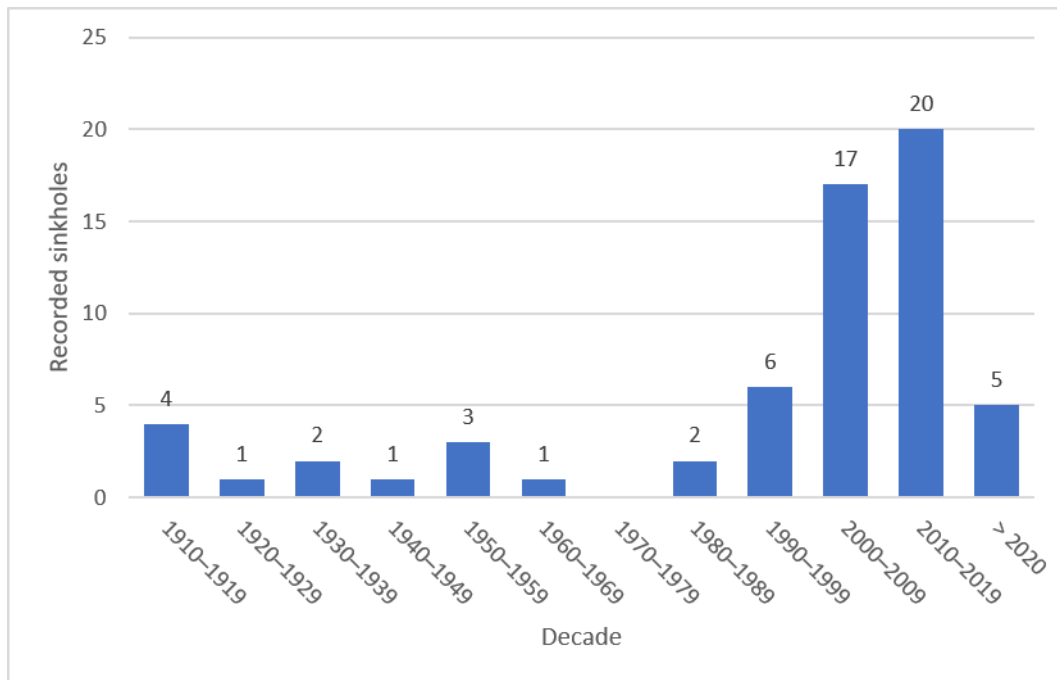


Figure 15 Decadal distribution of recorded sinkhole occurrences with usable calendar dates in France

The seasonal analysis was restricted to the 681 records documented at least to the month. Spring is the most represented season, with 251 cases (36.9%), followed by autumn (161 cases; 23.6%), summer (147 cases; 21.6%) and winter (122 cases; 17.9%) (Figure 16). These temporal patterns primarily reflect the structure of the inventory. They should not be used, without further control, to infer changes in sinkhole frequency or seasonal triggering. Archive availability, survey campaigns, retrospective data integration, reporting practices and seasonal detection conditions may all contribute to the observed distributions.

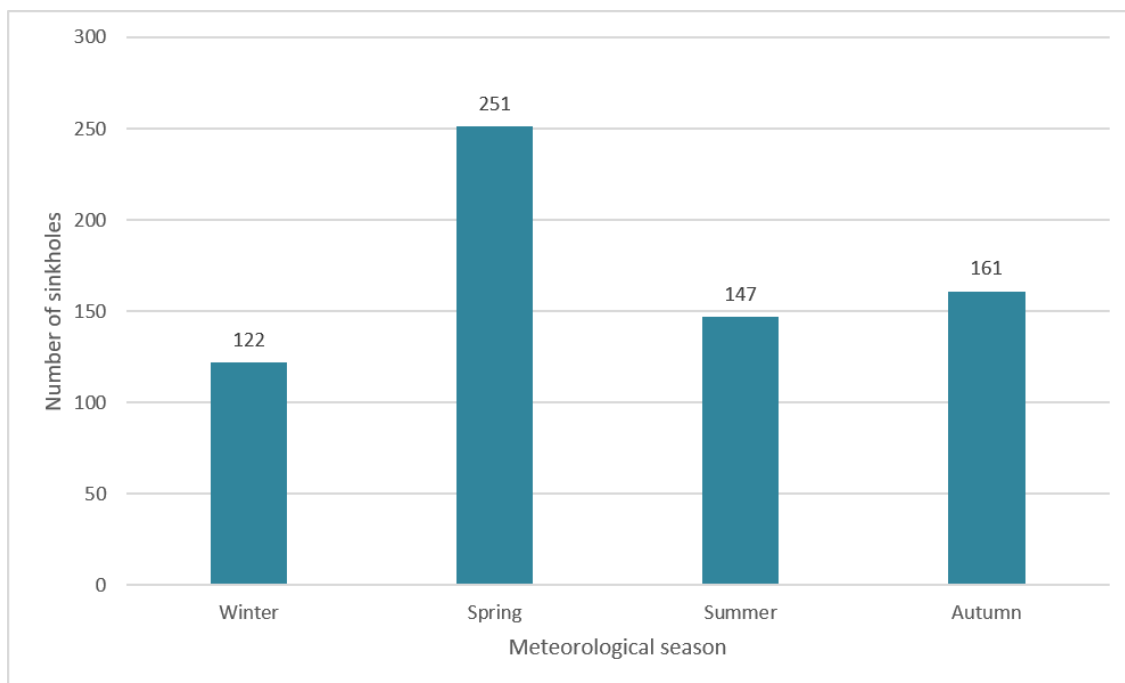


Figure 16 Seasonal distribution of recorded sinkhole occurrences. Only dates with sufficient temporal precision are included.

The large number of records labelled “before 2014” (306 cases) is largely related to the high proportion of Polish sinkholes in the dataset (Figure 17). Consistently, the annual distribution shows that 635 of the 785 precisely datable occurrences, or 80.9%, fall between 1954 and 2014. This concentration describes the temporal structure of the available records; it cannot, on its own, be used to identify a real change in sinkhole frequency.

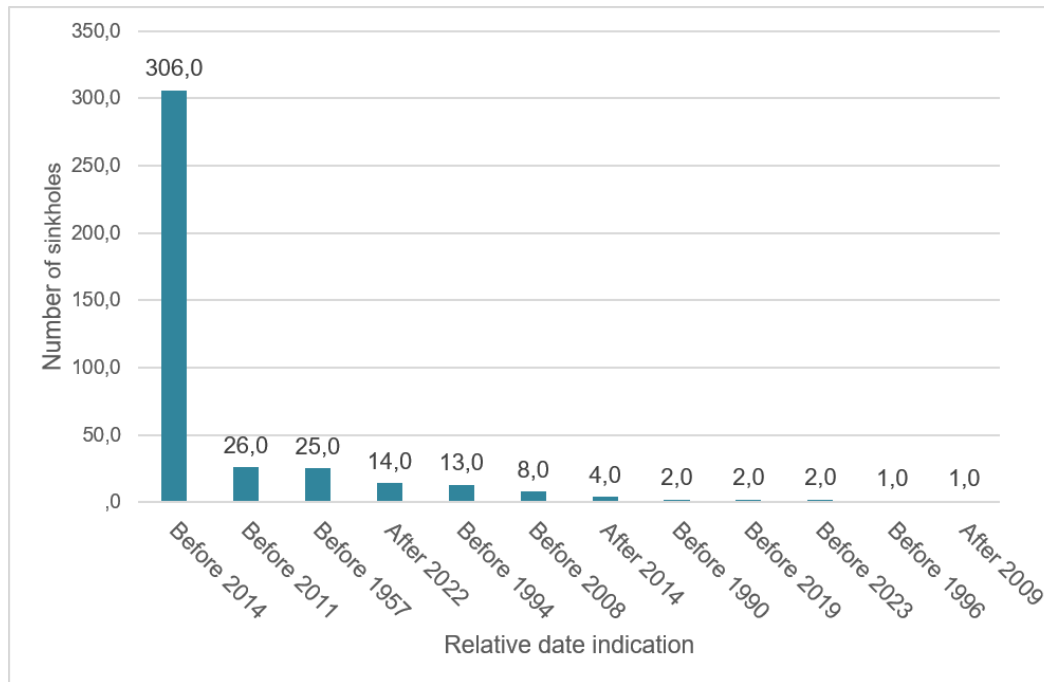


Figure 17 Distribution of sinkhole records according to reported “before” or “after” temporal thresholds (n = 404).

The temporal analysis mainly shows a concentration of recorded sinkhole occurrences from the 1980s onward. This result has to be read in the context of the available archives: older events are probably less accessible, less precisely documented or not preserved in the same way. Spring is the most represented season among dated records, but this pattern may also result from reporting and detection conditions. The distributions describe the structure of the inventory and cannot, on their own, demonstrate either a change in sinkhole frequency or a seasonal control.

5.3 Comparison between occurrence dates and mining periods

The database includes 123 distinct mines: 41 in France, 76 in Germany and 6 in Poland. The operational period could be calculated for 62 of these mines. As shown in Figure 18, mining durations range from 1 to 391 years, with an average of 127.5 years. The 100-149 year category is the most represented, with 13 mines, or 21 % of those for which an operational period could be calculated.

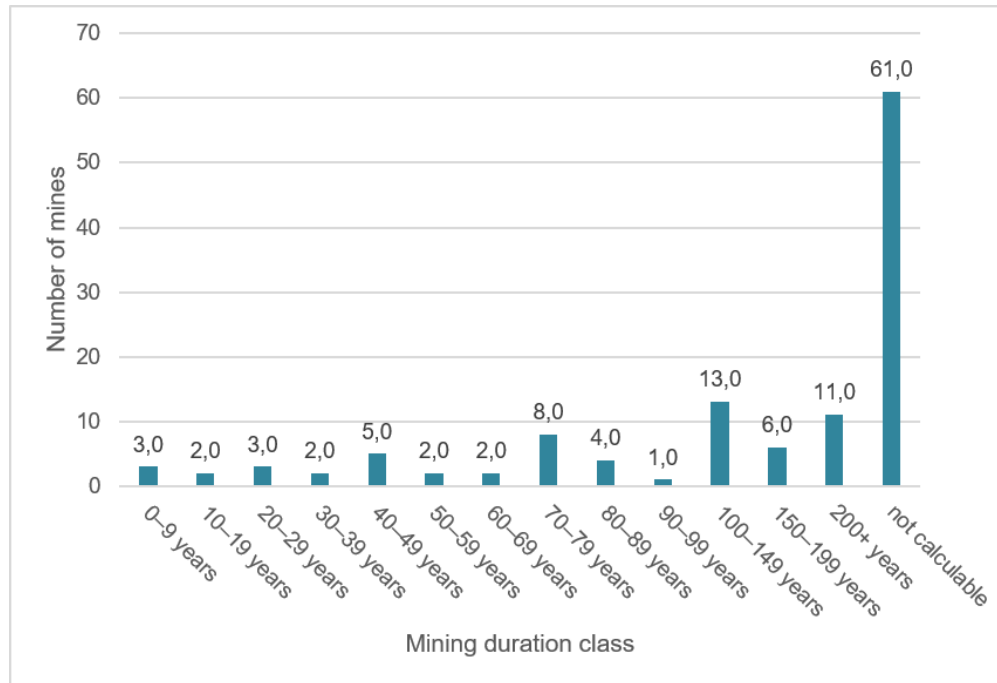


Figure 18 Distribution of operational duration for the 123 mines included in the inventory.

The time lag between sinkhole occurrence and mine closure could be calculated for 666 events (Figure 19, Table 6). In this subset, 385 sinkholes (57.8%) occurred before mine closure. These cases are almost entirely from Poland, with 379 records (98.4%), compared with six in France and none in Germany. Among them, 380 clearly occurred during the mining period, between the documented start and end dates. For the remaining five cases, the start date is missing, so the timing cannot be confirmed. The other 281 sinkholes (42.2%) occurred after mine closure. Post-closure events are not limited to the first years after mining stopped: 146 sinkholes, or about 22% of the full dated subset, were recorded between 1 and 30 years after closure, corresponding to just over half of the post-closure cases. Later occurrences are also documented, with 102 events more than 50 years after closure, including 51 more than a century after mining had ceased. This indicates that the first decades after closure are well represented in the inventory, but that recorded post-mining sinkholes may also occur over much longer timescales. Nevertheless, this interpretation remains limited by the use of the mine closure date as a proxy for the end of local mining activity. In large mines operated over long periods, individual sectors may have been abandoned well before official mine closure. District-level end dates would provide a more accurate reference, but this information is rarely available in the archives. The calculated time lag should therefore be considered an approximate indicator.

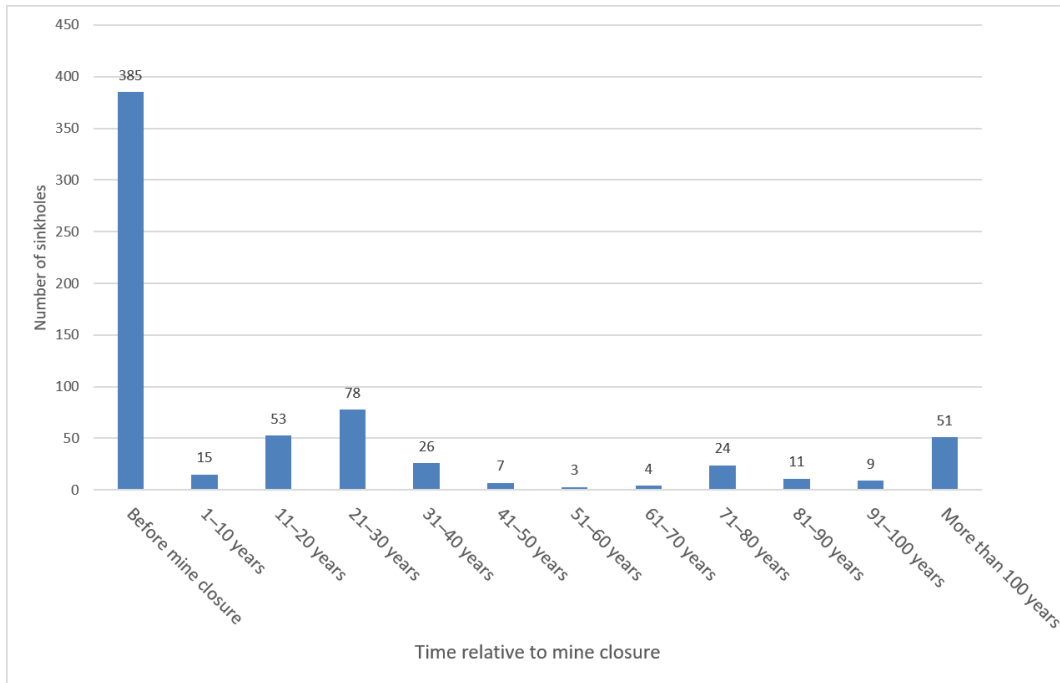


Figure 19 Sinkhole occurrence relative to mine closure ($n = 666$).

Table 6 Distribution of sinkhole occurrence relative to mine closure, based on 666 events with usable occurrence and closure dates.

Time lag between sinkhole occurrence and mine closure	Count	Share
Before mine closure	385	57,8%
1–10 years after closure	15	2,3%
11–20 years after closure	53	8,0%
21–30 years after closure	78	11,7%
31–40 years after closure	26	3,9%
41–50 years after closure	7	1,1%
51–60 years after closure	3	0,5%
61–70 years after closure	4	0,6%
71–80 years after closure	24	3,6%
81–90 years after closure	11	1,7%
91–100 years after closure	9	1,4%
More than 100 years after closure	51	7,7%
Total	666	100,0%

5.4 Interpretation and limitations

The comparison between sinkhole occurrence dates and mine closure dates highlights the long-term nature of post-mining instability. Events recorded several decades, and in some cases more than a century, after closure show that surface disorders can still appear long after mining has stopped. This is an important observation. Its interpretation is nevertheless constrained by the precision of

occurrence and closure dates, the incompleteness of older archives, uneven reporting and detection conditions, and differences in documentation between countries and mining basins. The result is best used as descriptive evidence of long-term instability; it is not sufficient, on its own, to estimate the timing, probability or causes of future sinkhole occurrence.

This long-term occurrence does not mean that all galleries or rooms will ultimately lead to sinkhole formation. Surface collapse depends on favourable local predispositions, including the geometry and depth of the workings, roof condition, the nature and thickness of the overburden, hydrogeological conditions and possible triggering factors.

6. Analysis of sinkhole characteristics

6.1 Sinkhole dimensions

Length or diameter is available for 1126 sinkholes, while 599 records (34.7%) have no usable value. The variable refers to length when the sinkhole is described as ellipsoidal; in that case, a width may also be recorded. For circular sinkholes, the same field corresponds to the diameter. The most frequent class is 2-5 m, with 410 cases (23.8% of the full dataset). A further 294 sinkholes (17.0%) are 2 m or less. Taken together, these two classes represent about 41% of the full population and almost two thirds of the documented cases.

Larger features are less common. The database includes 217 cases between 5 and 10 m and 163 between 10 and 20 m. Only 42 sinkholes exceed 20 m, including seven above 50 m. For these very large features, the term sinkhole should be checked against the original descriptions, as some records may correspond to broader collapse phenomena rather than individual localized sinkholes. The distribution, illustrated in Figure 15, therefore describes a population dominated by small to medium-sized surface features, while still retaining a limited number of large events that require separate attention.

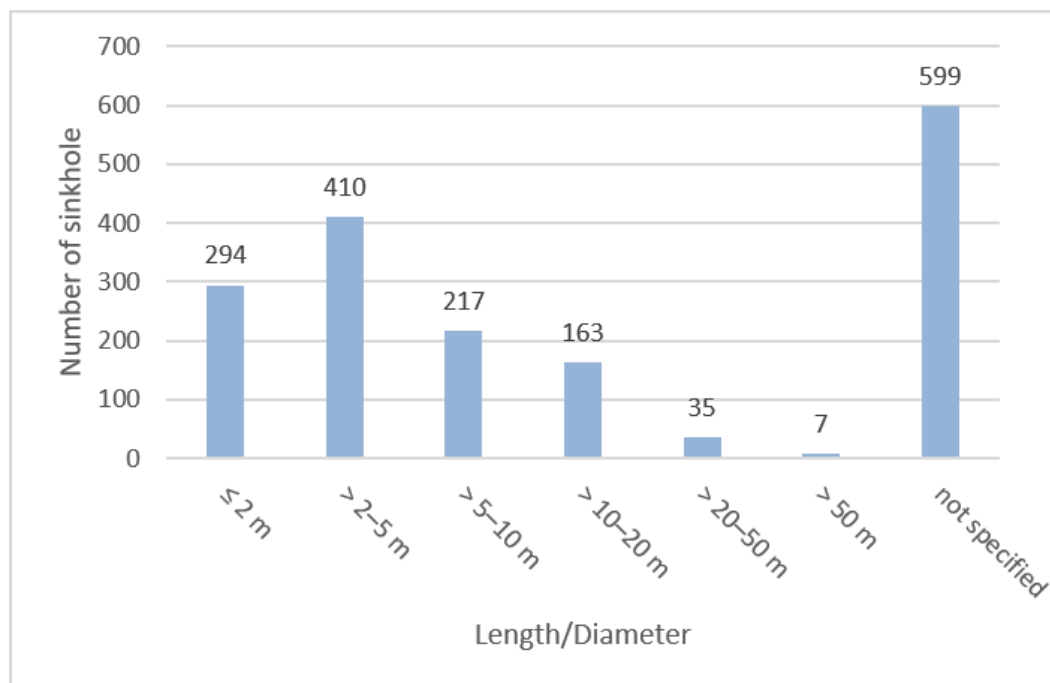


Figure 20 Distribution of length or diameter for sinkholes related to underground workings.

The diameter of the collapse is a key factor controlling the foreseeable consequences for people and property within the area of influence. In the French post-mining hazard mapping approach, it is therefore used as the reference parameter for qualifying sinkhole intensity. The corresponding

intensity classes, based on Franck et al., (2019), are summarised in Table 7. Figure 21 shows the distribution of the SIRIMA sinkholes according to these classes. Most documented sinkholes are classified as low intensity (36,3%), reflecting the predominance of collapse diameters below 5 m in the available dataset.

Table 7 Sinkhole intensity classes based on collapse diameter, following Franck et al. (2019).

Intensity class	Diameter of collapse
Very low	Self-backfilled collapses in immediate proximity to the surface (centimeters in depth)
Low	$\varnothing < 5$ m
Moderate	$5 \text{ m} < \varnothing < 10$ m
High	$\varnothing > 10$ m

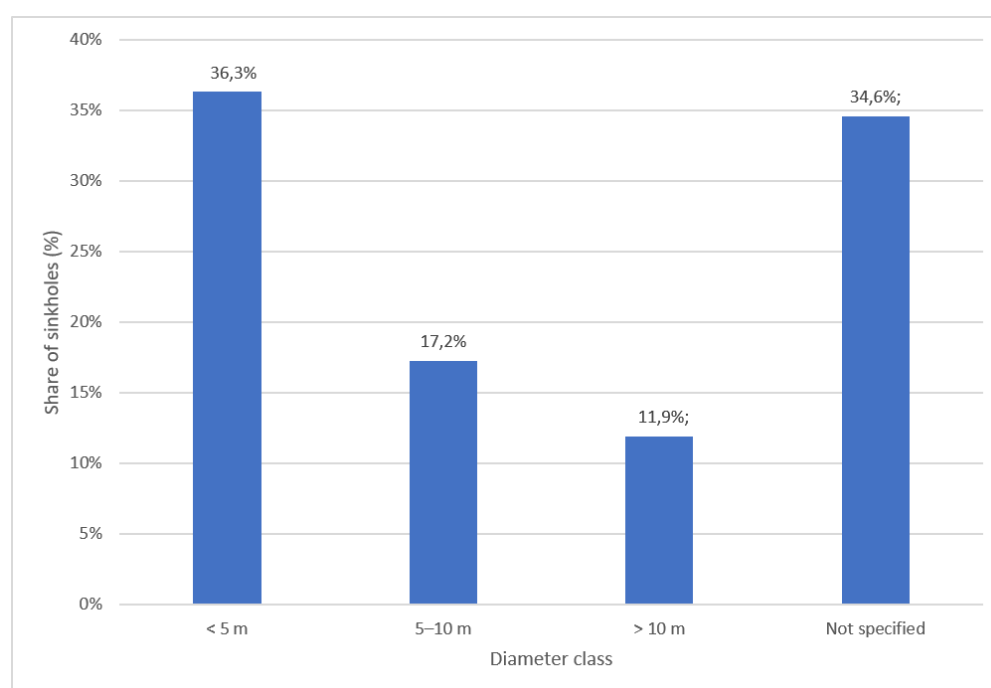


Figure 21 Distribution of sinkholes according to the French intensity classes based on collapse diameter.

Width is the least complete dimensional variable. It is missing for 1,328 sinkholes, corresponding to 77.0% of the dataset. This is probably because most sinkholes are circular or nearly circular (ellipsoidal), so it is generally not necessary to distinguish between length and width. Among the 397 documented cases, 215 have a width of 2 m or less and 140 fall between 2 and 5 m. Together, these two classes account for almost 90% of the available width measurements.

The recorded width values are generally consistent with the corresponding length values. Among the 42 cases with a width greater than 5 m, all have an associated length value, and no width exceeds the corresponding length. Only ten records have a width greater than 10 m. These larger cases have a median length of 50 m and a median width of 25 m, with maximum dimensions reaching 180 × 100 m. They are mainly recorded in France: nine of the ten widths greater than 10 m are located there, several of them in the Commeny area. In the Commeny case, the largest dimensions probably reflect the aggregation of several sinkholes within the same area rather than the size of a single event.

Large-width features therefore exist in the database, but they remain rare within a population otherwise dominated by small to medium-sized sinkholes, as summarised in Figure 16.

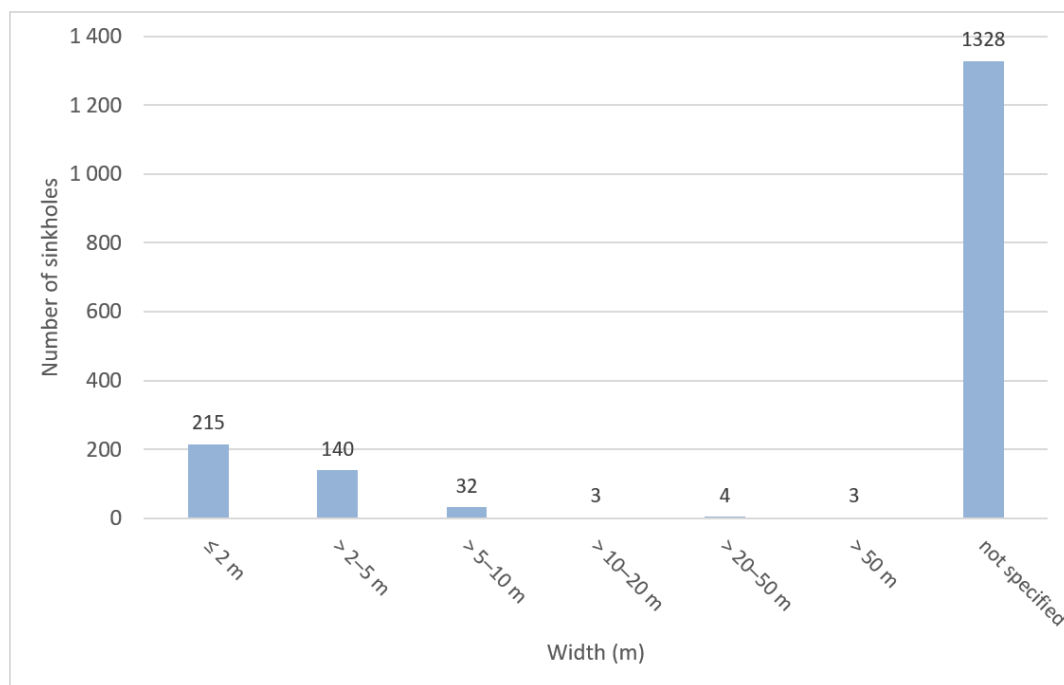


Figure 22 Distribution of width for sinkholes related to underground workings.

Impacted surface area is documented for 1,126 sinkholes and missing for 599 cases (34.7%). This variable was rarely provided directly in the source database. It was therefore recalculated, where possible, from the available dimensions: length multiplied by width for ellipsoidal sinkholes, and a circular surface formula when the recorded value corresponded to a diameter. The largest class is 10-50 m², with 308 records (17.9%). Smaller affected areas are also frequent: 129 cases cover 1 m² or less, 217 falls between 1 and 5 m², and 176 between 5 and 10 m². Overall, 522 sinkholes, or 30.3% of the full dataset, have an estimated or reported impacted area of no more than 10 m².

More extensive impacts are less frequent but not exceptional. The database includes 141 cases between 50 and 100 m², 132 between 100 and 500 m², and 23 above 500 m². These larger values may correspond to grouped events (see above), more rarely to large surface collapses, or to differences in the way the affected area was measured or estimated by each contributor. They should therefore be checked against the original descriptions before being used in comparative or hazard analyses. Where possible, they should also be separated from grouped or poorly located surface disorders. Figure 23 provides the corresponding distribution.

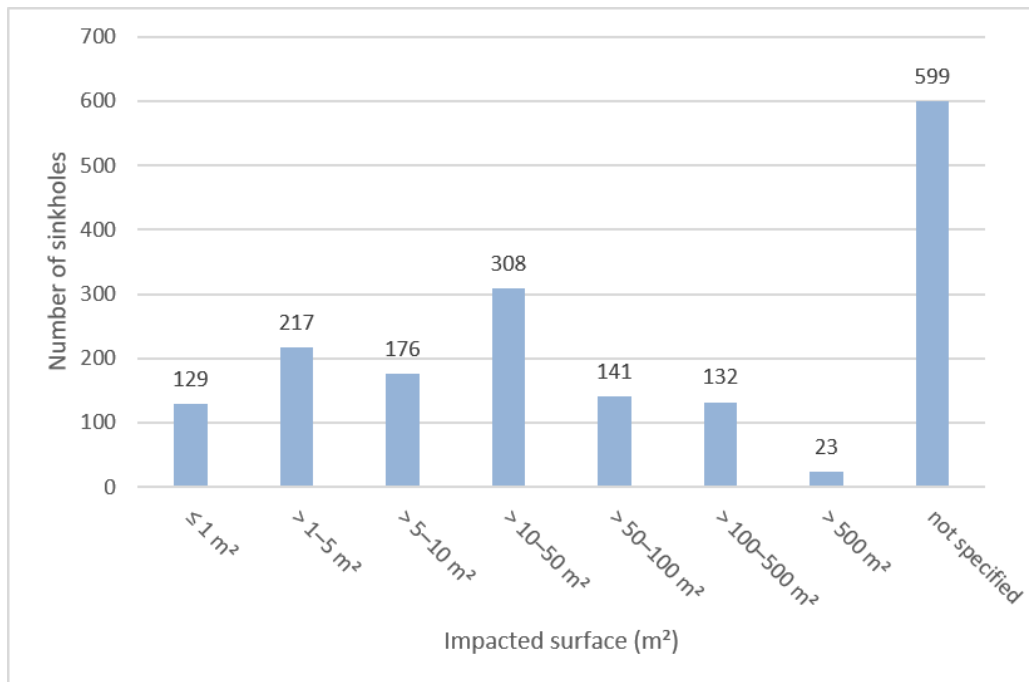


Figure 23 *Distribution of impacted surface area for sinkholes related to underground workings.*

Sinkhole depth is documented for 1257 of the 1725 records, corresponding to 72.9% of the dataset. The distribution is dominated by shallow features. The largest class is 0-1 m, with 472 cases (27.4% of the full dataset), followed by 1-2 m, with 361 cases (20.9%), and 2-5 m, with 328 cases (19.0%). Together, these classes represent 1161 records, meaning that more than 92% of sinkholes with usable depth information are no deeper than 5 m.

Deeper sinkholes are much less frequent. Eighty cases (4.6%) have depths between 5 and 10 m, and only 14 records (0.8%) exceed 10 m. Two additional values are expressed only as lower limits and cannot be assigned reliably to a closed depth class. Depth is considered unspecified for 468 records (27.1%), including 159 entries initially reported as 0 m, since these values may reflect missing information or a reporting convention rather than a measured depth. The dataset also combines field measurements, estimates and threshold values. Figure 24 should therefore be read as a descriptive overview of recorded depths, not as a precise statistical mode¹.

¹ Depth information may also be biased for sinkholes observed long after their formation, as partial natural backfilling can reduce the apparent depth or prevent it from being reported. A similar limitation applies to diameter: the final observed diameter may be larger than the initial opening, due to progressive widening and stabilization of the edges.

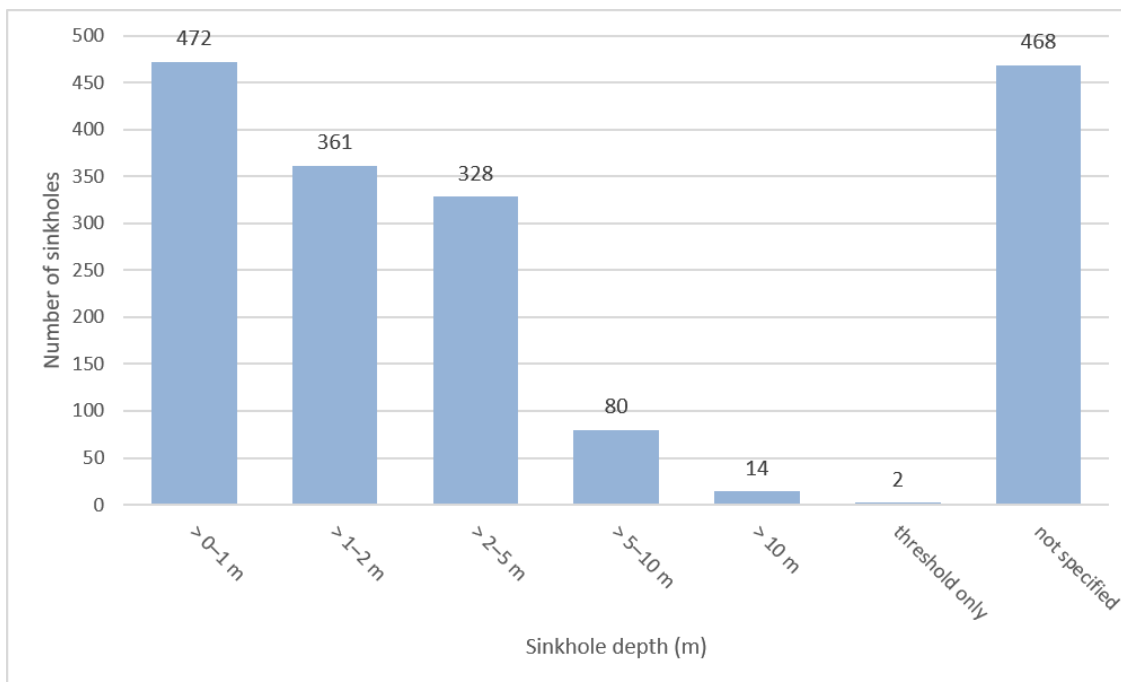


Figure 24 Distribution of sinkhole depth for sinkholes related to underground workings.

The SIRIMA database is dominated by small and shallow sinkholes. Most reported lengths or diameters do not exceed 5 m, available width measurements show the same pattern, and impacted surface areas are commonly below 50 m². Among records with usable depth information, more than 92% of sinkholes are no deeper than 5 m, while depths greater than 10 m remain rare. This overall pattern is consistent with values commonly reported in the literature for sinkholes related to shallow abandoned mine workings (Salmon et al., 2015) .

A smaller group of large or deep features remains important because of its potential consequences and should be examined separately. Some extreme dimensions may correspond to grouped collapses, subsidence depressions or differences in national reporting practices rather than directly comparable individual sinkholes.

The main limitation is data completeness, especially for width and, to a lesser extent, depth. The treatment of values reported as 0 m also requires caution, because they may indicate missing information rather than actual zero-depth features.

6.2 Origin of collapse

The reported origin of collapse is available for 1211 of the 1725 sinkholes. The remaining 514 records, corresponding to 29.8% of the dataset, have no specified origin. The variable “origin of collapse” describes the presumed mechanism responsible for the sinkhole formation, as identified from the available source information. Operators could select among predefined categories corresponding to roof failure above a room, roof failure above a gallery, pillar failure, or another documented origin.

The distribution is clearly dominated by roof-related failure mechanisms. The largest category is “failure of the roof room or roof gallery”, with 859 records, corresponding to 49.8% of the full dataset. A further 297 records are attributed specifically to roof gallery failure, and 44 to roof room failure. Together, these three categories account for almost all records with a documented origin. The database therefore mainly describes sinkholes associated with the upward propagation of collapses from underground rooms and galleries, as shown in Figure 25.

This distribution also reflects the way information was entered in the Excel database. Although only four predefined categories were initially proposed, several contributors added the combined category “failure of the roof room or roof gallery”. This probably reflects the difficulty, in many cases, of

distinguishing between collapse initiated from a room roof and collapse initiated from a gallery roof, especially when underground access or detailed investigation data were limited. For this reason, roof room and roof gallery failures were grouped in a simplified classification, shown in Figure 26. Moreover, in many cases, the information was probably unknown, and the corresponding fields were therefore not filled in (514 records where not specified). This may bias the analysis.

Other collapse mechanisms are rare in the current dataset. Pillar failure is reported for only six sinkholes, and five records fall within the “other” category. This provides useful context, but the origin categories are not evenly represented. Comparisons between mechanisms should therefore be interpreted carefully, especially for pillar failure and other rare origins. The mining method is probably a key factor in this point. Pillar failure generally concerns room-and-pillar workings and requires highly extracted or strongly degraded pillars, especially at shallow depth. It is therefore not necessarily the dominant mechanism in shallow workings, whereas it may occur in deeper mines where vertical stress is higher. Nevertheless, it may also reflect differences in terminology, investigation practices or the level of detail available in the source inventories.

The collapse-origin field should therefore be read as an interpretative descriptor rather than as a confirmed mechanical diagnosis. In many cases, the affected workings are no longer accessible and the available information relies on indirect observations, archival descriptions or post-event interpretation. This limits the ability to distinguish confidently between roof-room failure, roof-gallery failure and other local mechanisms.

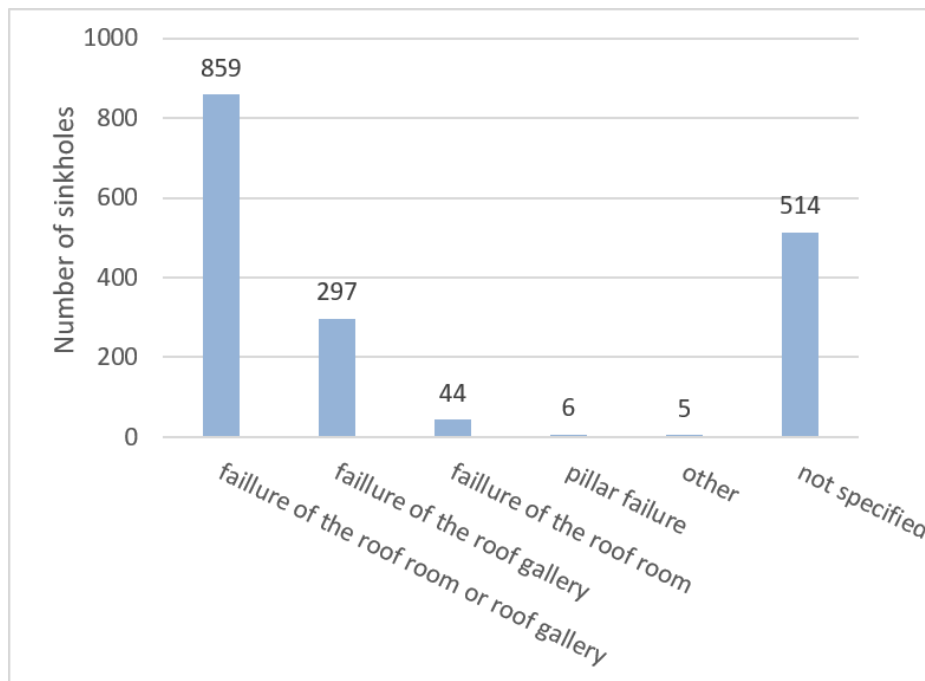


Figure 25 Distribution of reported collapse origins for sinkholes related to underground workings.

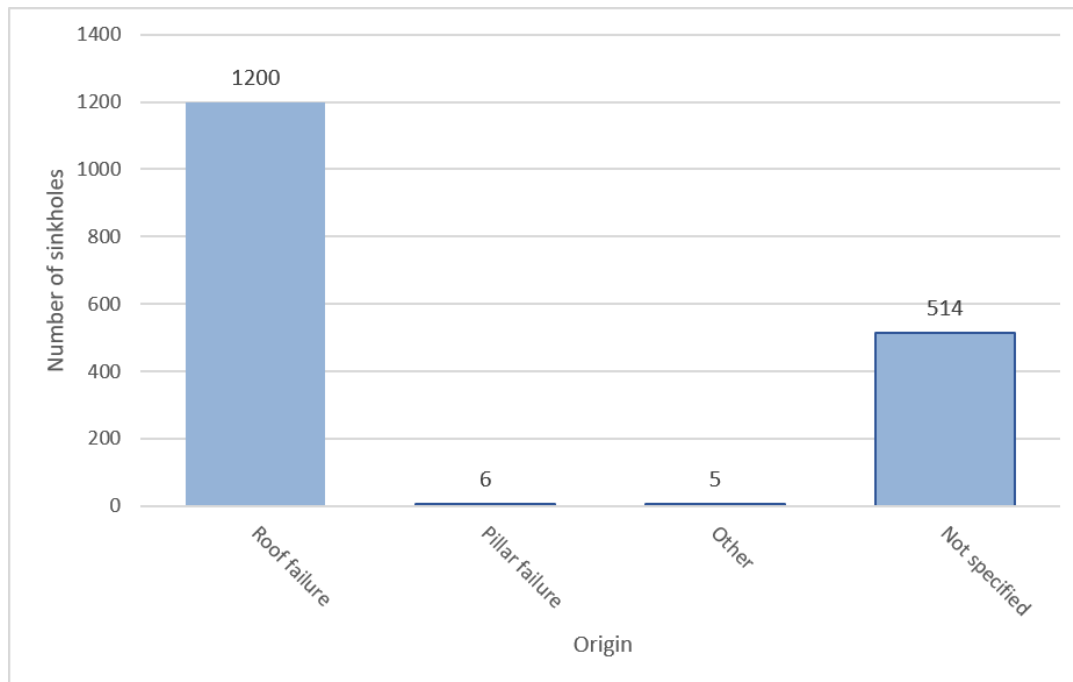


Figure 26 Simplified distribution of reported collapse origins after grouping roof-room and roof-gallery failures.

7. Analysis of mining-related characteristics

7.1 Working depth

The depth of the underground workings was analysed as a key mining-related variable, because it controls the thickness of the overburden separating the void from the ground surface. The distribution, summarised in Figure 25, shows that sinkholes are mainly associated with moderately shallow workings. Within the 1725-record underground-working dataset, the 25-50 m class is the most represented, accounting for 34.0% of records. Workings located at depths of 10-25 m and 50-100 m account for 14.5% and 12.6%, respectively. Shallower workings, at 10 m or less, represent 6.8% of records, while only 4.9% are deeper than 100 m. Depth is unspecified for 27.2% of the dataset, which remains a significant source of uncertainty. Cases involving sinkholes above workings deeper than 50 m should be examined with particular care, since such configurations are uncommon and may involve specific mining, geological or overburden conditions. Thus, Salmon et al. (2015), based on an analysis of more than 1800 localized mine collapses, reported no collapse occurring directly above mine workings located at depths greater than 50 m. Canbulat et al. (2017) made a similar observation based on the analysis of about one hundred sinkholes in New South Wales associated with coal mining. Their study showed that the cover depth generally ranged from 4 m to 50 m, with an average of approximately 15.9 m.

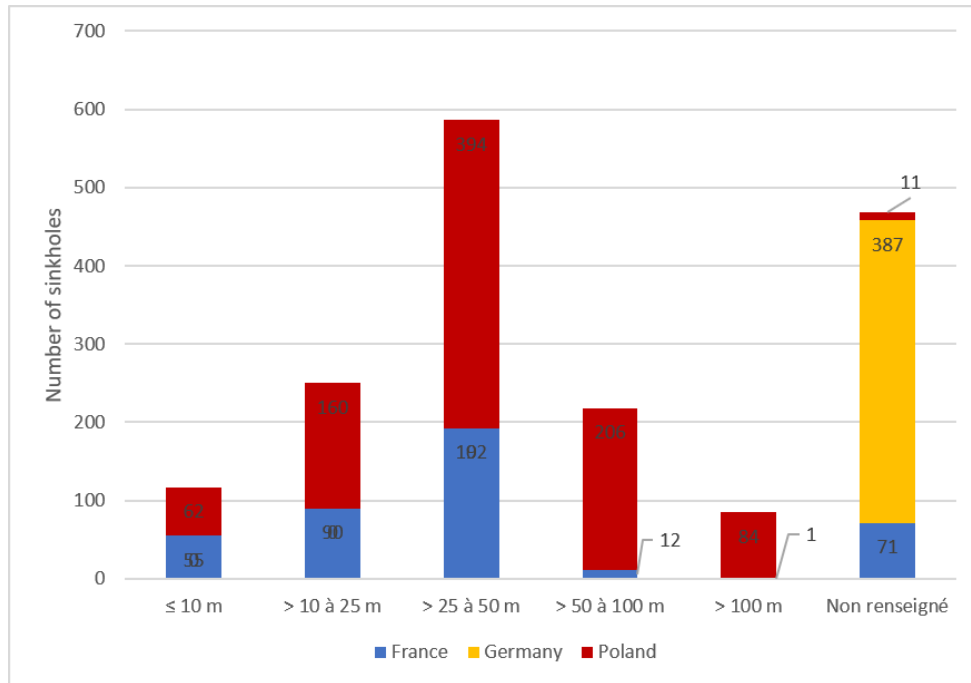


Figure 27 Distribution of underground-working depth below ground level for sinkholes related to underground workings for the different countries

7.2 Working thickness

Figure 28 shows that the 2-10 m class is the dominant category, accounting for 52.3% of the underground-working dataset. Workings thinner than 2 m represent 11.0% of records, while 36.7% of the records have no specified thickness. The distribution therefore remains clearly structured around the 2-10 m class, but the large proportion of missing values limits the interpretation of this parameter and prevents any detailed comparison beyond the classes represented in the figure.

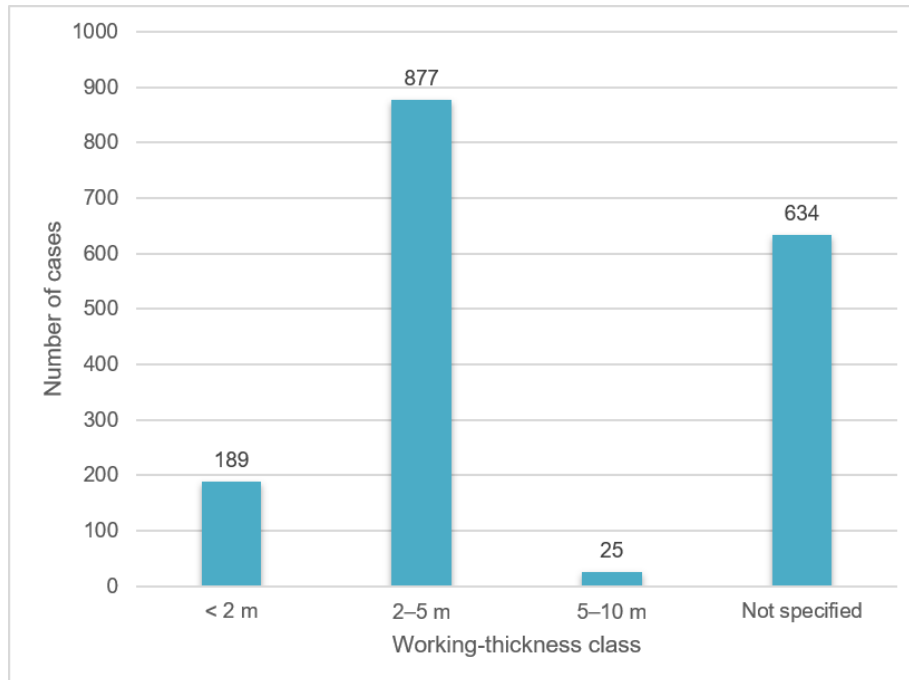


Figure 28 Distribution of working thickness for sinkholes related to underground workings.

7.3 Working-depth to working-thickness ratio

The ratio between working depth and working thickness was examined because it provides a simple indicator of the relative geometry of the mined void with respect to the overburden. In post-mining engineering practice, sinkholes are generally considered much less likely when this ratio is high. For example, in the specific case of quarries in the Paris region Vachat, (1982), showed that localized surface collapses become uncommon when this ratio exceeds about 15, because the overburden is relatively thick compared with the height of the void. The most frequent cases reported by Vachat, (1982), correspond to a ratio of about 6. In Polish conditions, based on recorded sinkhole occurrences, Kotyrba (2005) adopted a three-level classification of sinkhole hazard based on the ratio of mining depth to the extracted coal seam thickness (i.e., the height of the primary void). For values lower than 10, a high sinkhole hazard is assumed, as under such conditions there is a high probability that the caving zone will propagate to the ground surface and cause a loss of its continuity, resulting in the formation of a sinkhole. .

The distribution shown in Figure 29 is consistent with the expected role of the working depth/thickness ratio: sinkholes are mainly recorded where the cover is relatively limited compared with the height of the mined void. Records with ratios greater than 15 are less frequent and should be examined individually, because they may reflect specific local geological or hydrogeological conditions, uncertainty in the reported depth or thickness, or inconsistencies in the source description.

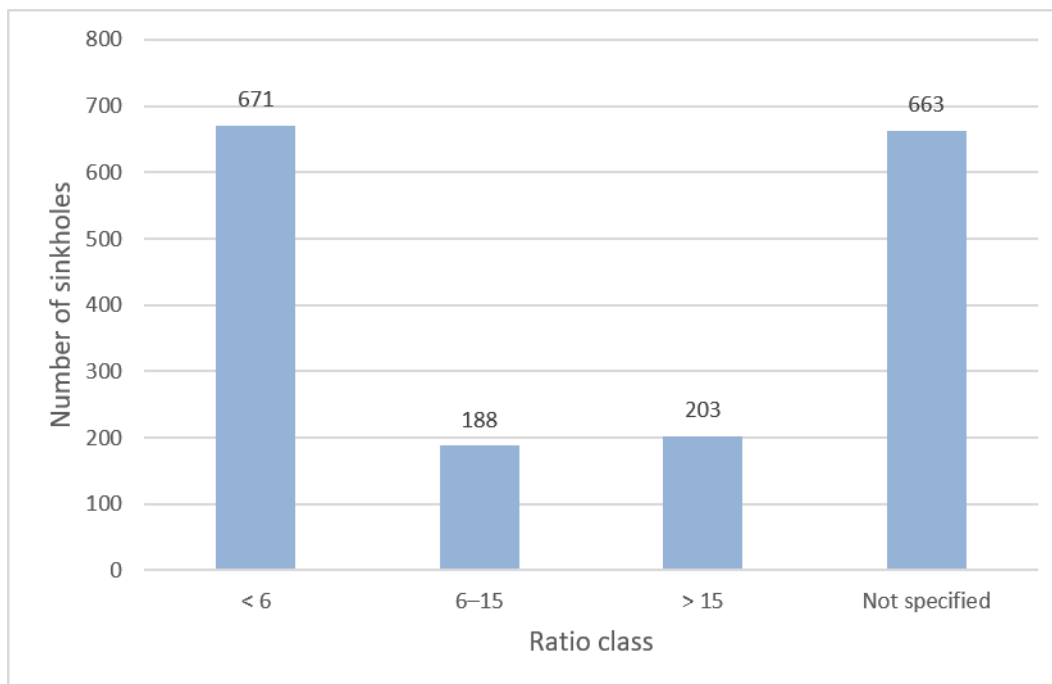


Figure 29 Distribution of the ratio between working depth and working thickness for documented sinkhole records.

7.4 Coal seam dip angle

The dip of the exploited layer is strongly concentrated in the lowest class. Among the 1,725 sinkholes, 999 cases (57.9%) are associated with dips of 10° or less, while only 183 cases (10.6%) correspond to steeper dips. Dip information is missing for 543 records (31.5%). Some documented values also remain uncertain: in four cases, the reported interval ranges from 0° to 70°. When dip values were recorded as intervals, their midpoint was used for classification. These large intervals limit the precision of the comparison, especially for records with poorly constrained dip values. Figure 24 presents the resulting distribution.

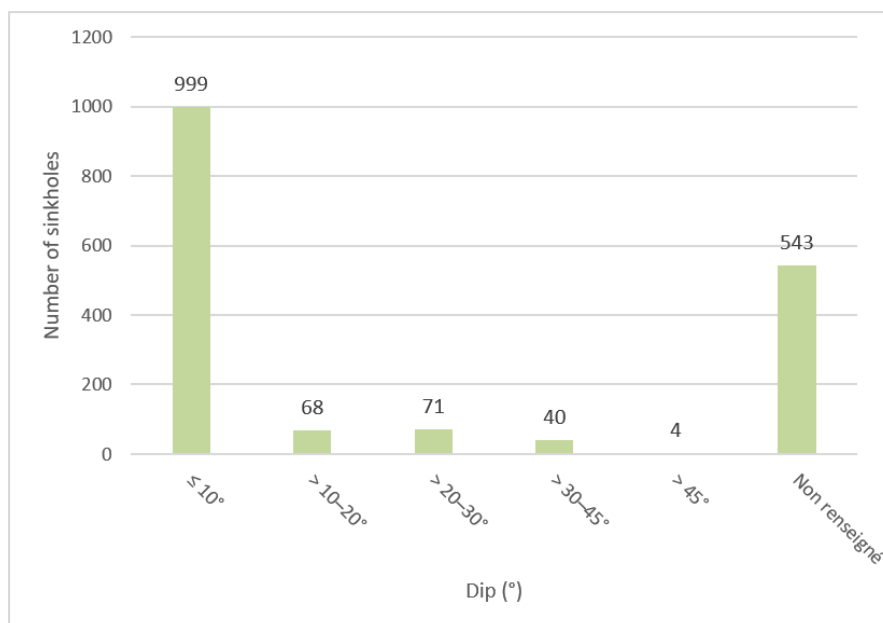


Figure 30 Distribution of the dip of the exploited coal layer.

7.5 Relationships between mining variables and sinkhole characteristics

The relationship between working depth and sinkhole characteristics was examined using median values by depth class. Figure 31 and Figure 32 show that median sinkhole length and impacted surface area increase with working depth up to the 50-100 m class. Median length rises from 3 m for workings shallower than or equal to 10 m to 9 m for workings between 50 and 100 m. Over the same range, the median impacted surface increases from 4 m² to 55 m². These two trends are closely related, since impacted surface area is largely derived from the horizontal dimensions of the sinkhole. Beyond 100 m, both medians decrease, which suggests that the relationship is not linear across the full dataset.

Working depth was divided into five classes: ≤10 m, >10-25 m, >25-50 m, >50-100 m and >100 m (Table 8). When source values were reported as intervals, their midpoint was used for classification. For open-ended values, half the threshold was used for “less than” records, while the stated threshold was retained for “greater than” records. Non-interpretable descriptions were classified as unspecified.

Table 8 Median sinkhole characteristics by underground-working depth class.

Working-depth class	Number of sinkholes	Median length / diameter (m)	Median width (m)	Median sinkhole depth (m)	Median impacted surface (m ²)
≤ 10 m	142	3	2	2	4
> 10-25 m	382	3	3	2	8
> 25-50 m	439	6	2	2	28
> 50-100 m	208	9	1	2	55
> 100 m	84	3	3	1	7
Not specified	470	3	2	1	7

Median values were preferred to averages because sinkhole dimensions are asymmetric and include a few very large observations that could distort the mean. The median gives a clearer description of a typical sinkhole within each working-depth class. The comparison is nevertheless limited by data availability: working depth is documented for 73% of records, whereas width is available for only 23% but as it was said before sinkholes are in general considered circular. However, as noted above, most sinkholes are considered circular or nearly circular, which partly explains the limited availability of width values. The decrease observed beyond 100 m is also based on a smaller group of 84 records. The available data indicate an association between working depth and sinkhole dimensions, but they do not demonstrate direct mechanical control. This result differs from the usual interpretation, reported by some authors, that surface diameter generally decreases as depth increases (Vachat, 1982). However, this relationship also depends on the nature of the overburden materials, especially near the surface. These cases should therefore be examined in more detail.

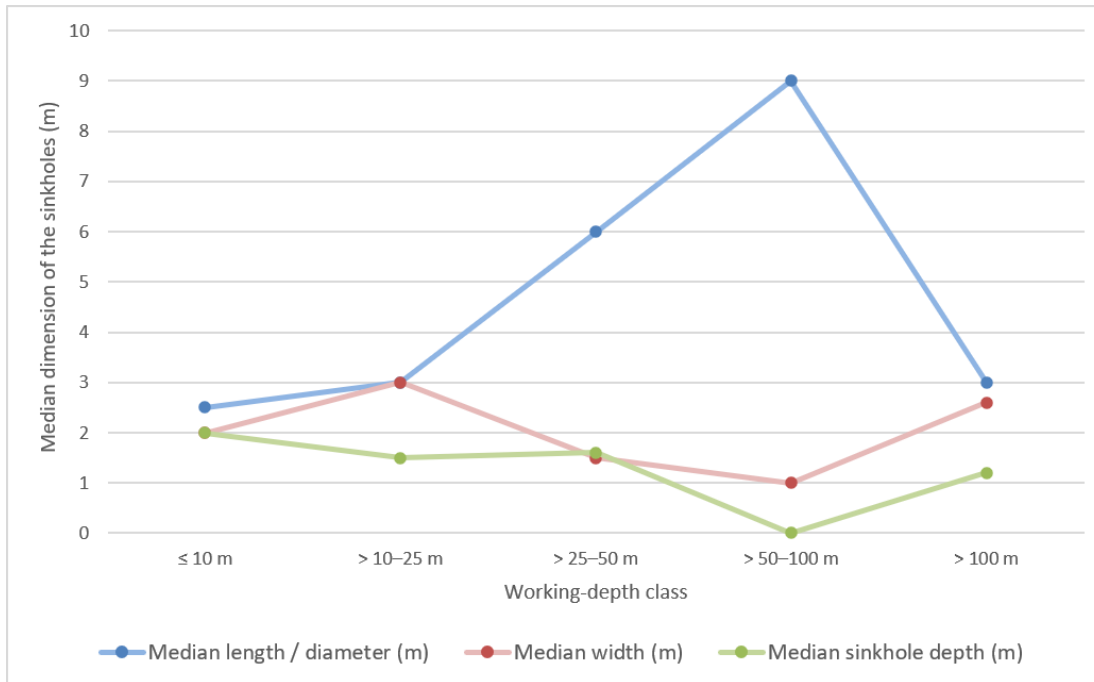


Figure 31 Median sinkhole horizontal dimensions by underground-working depth class.

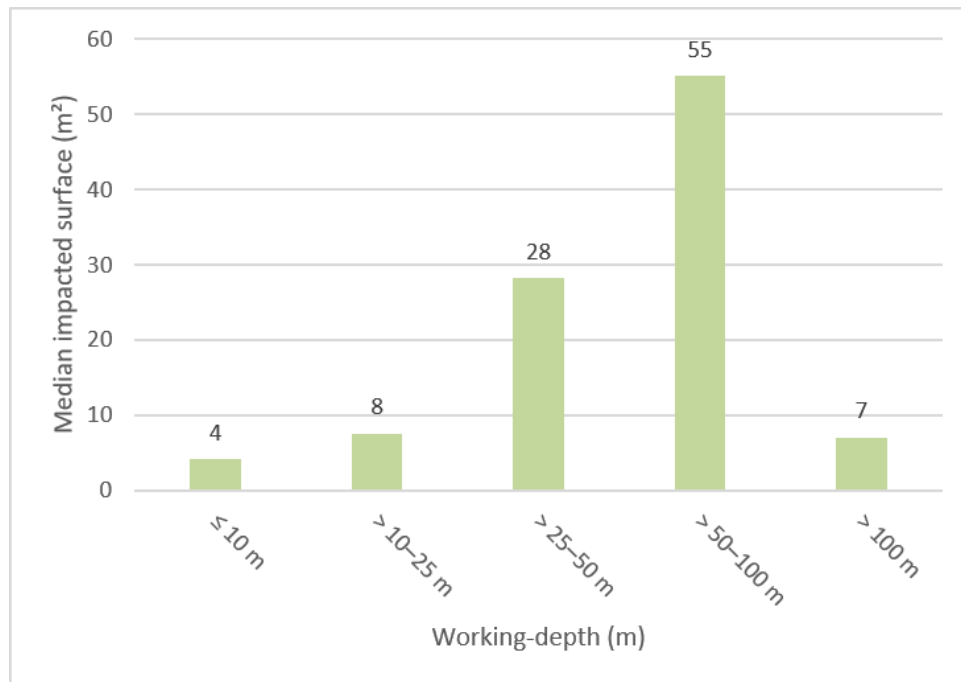


Figure 32 Median impacted surface area by underground-working depth class.

Figure 33 compares median sinkhole depth across the working-depth classes. Unlike length and impacted surface area, sinkhole depth shows no clear relationship with the depth of the underground workings. Median values remain within a narrow range, from 1.2 to 2.0 m, and do not increase progressively as workings become deeper. This suggests that deeper underground workings may be associated, in some classes, with a greater horizontal extent of surface disturbance, but not necessarily with deeper surface depressions. The lower median observed for workings deeper than 100 m should remain cautiously interpreted, as this class contains only 84 records.

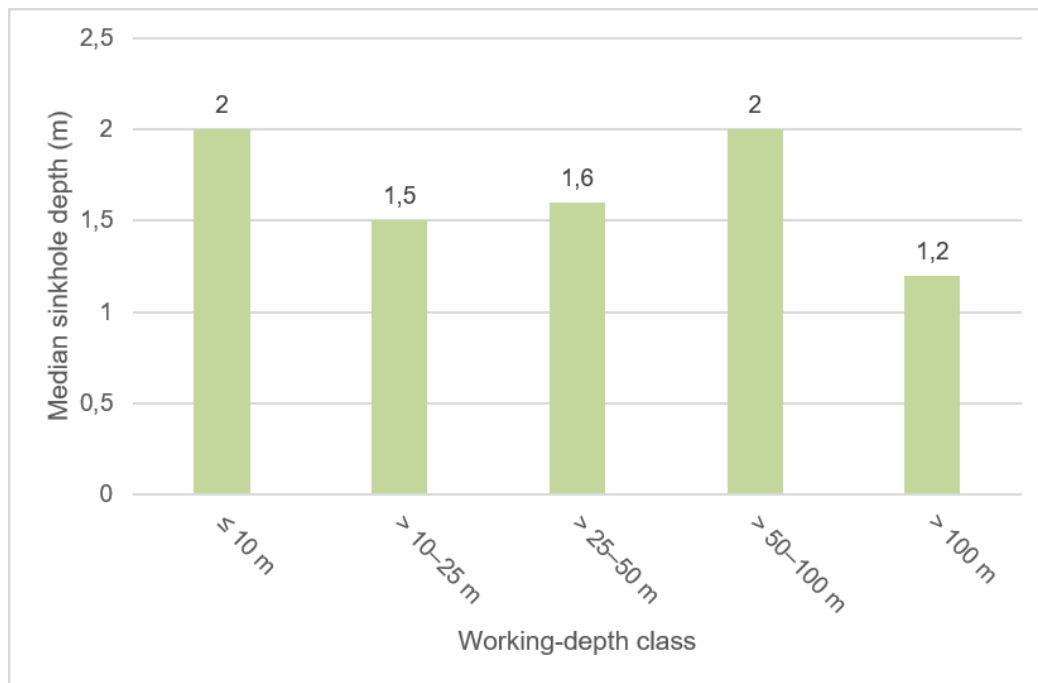


Figure 33 Median sinkhole depth by underground-working depth class.

Figure 34 presents median sinkhole length and depth for each working-thickness class. Records were grouped into three classes: <2 m, 2-10 m and >10-20 m. Medians were calculated separately using the available length and depth measurements.

Median sinkhole length and depth were compared for the working-thickness classes represented in Figure 35. Median length increases from 3 m for workings thinner than 2 m to 5 m for the 2-10 m class. By contrast, median sinkhole depth decreases slightly, from 2.0 m to 1.6 m. This pattern remains descriptive. The medians for length and depth may be based on different subsets, and the interpretation is constrained by missing thickness or sinkhole-dimension data.

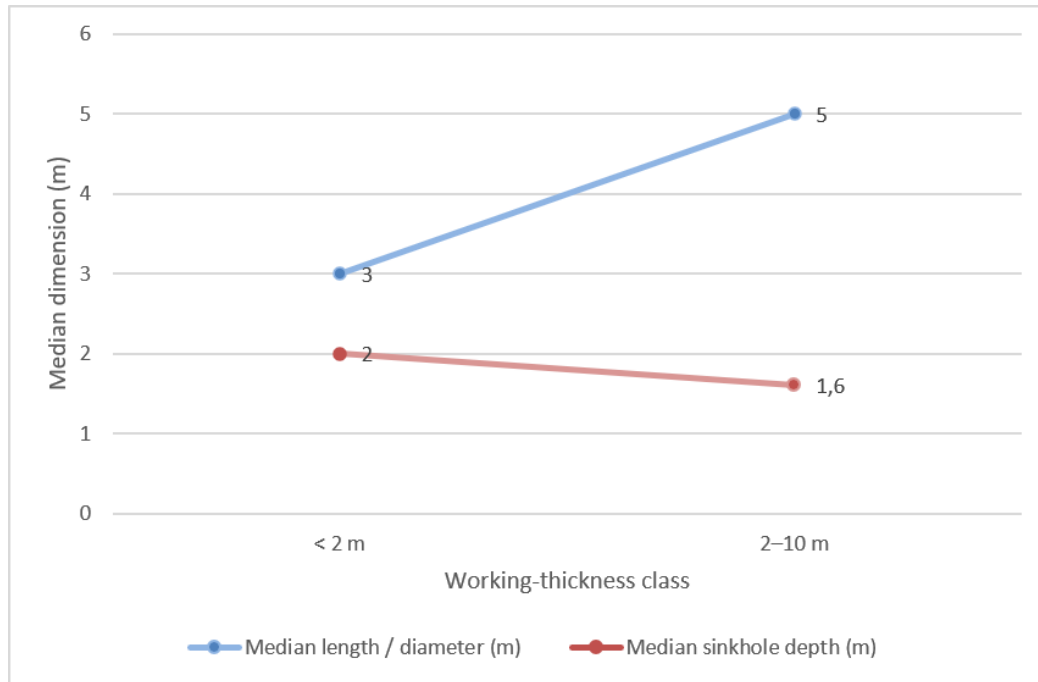


Figure 34 Median sinkhole length and depth by working-thickness class.

7.6 Water conditions and roof conditions

Water conditions in the mining workings are documented for 1186 of the 1725 sinkholes related to underground workings, corresponding to a completion rate of 68.8%. The available records are largely dominated by seepage conditions, which account for 1095 cases, or 63.5% of the full dataset. Flooded workings are reported for 67 cases (3.9%), while runoff and dry conditions remain rare, with 12 and 10 cases respectively. Water conditions are unspecified for 539 records (31.2%).

Based on the documented records, a majority of sinkholes therefore appears to have occurred under seepage conditions.

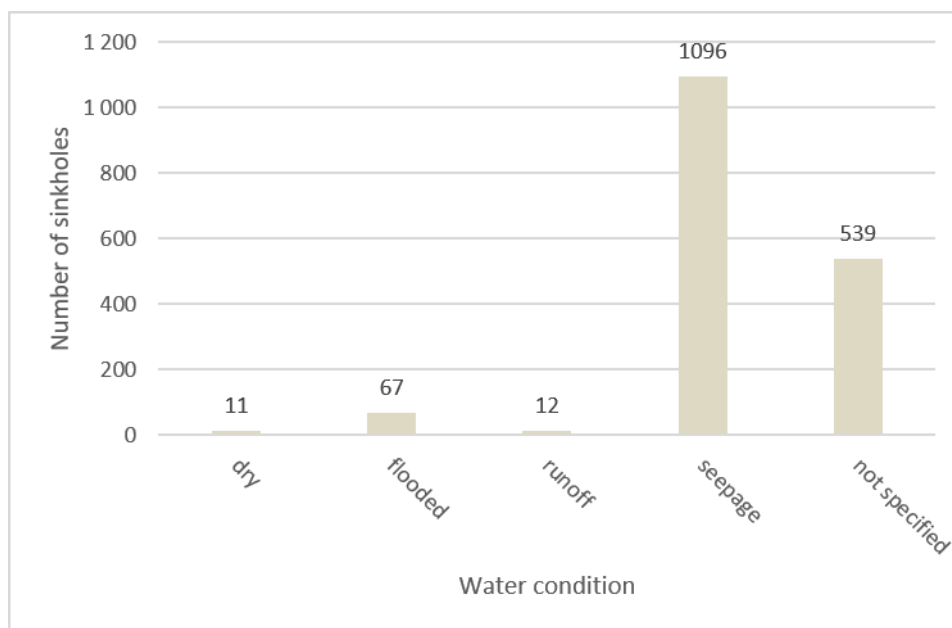


Figure 35 Distribution of sinkholes according to water conditions in the mining workings.

7.7 Type of mining structure: underground workings versus shafts

The distinction between sinkholes related to underground workings and those associated with shafts is essential for the interpretation of the database. These two populations are linked to different underground structures and may involve different collapse mechanisms, geometries and surface expressions. They should therefore not be merged without qualification in a single statistical analysis.

For shaft-related cases, the available information concerns the shaft itself, its depth or diameter when reported, the overburden context, and the observed surface disorder. However, the level of documentation is more heterogeneous than for the underground-working population. Some records are sufficiently detailed for technical interpretation, while others mainly confirm the presence of a shaft-related surface collapse. This heterogeneity limits the possibility of obtaining reliable statistical relationships from the shaft subset alone.

8. Analysis of geological and overburden conditions

8.1 Geological descriptions of the overburden

This section examines the geological and overburden information available for sinkholes related to underground workings. The analysis focuses on the way geological descriptions are documented, harmonised and grouped for interpretation. The objective is not to establish a causal relationship between geology and sinkhole occurrence, but to assess whether the available descriptions can support consistent comparisons between records.

Because the original entries differ in terminology and level of detail, the geological descriptions were first grouped by general geological age, mainly Carboniferous, Permian and Triassic. Records for which the geological age could not be clearly identified were retained as age unspecified or not reported. The resulting distribution is shown in Figure 36.

The descriptions were then grouped into five simplified lithological classes based on dominant composition: A, conglomerates, sandstones and shales; B, conglomerates, sandstones, shales and limestones; C, marls and dolomites; D, limestones and dolomites; and E, superficial deposits. Table 10 summarises the cross-distribution between general geological age and lithological class. This harmonisation highlights the predominance of Carboniferous sandstone-shale formations, but the

interpretation remains limited by the proportion of missing or insufficiently detailed geological information.

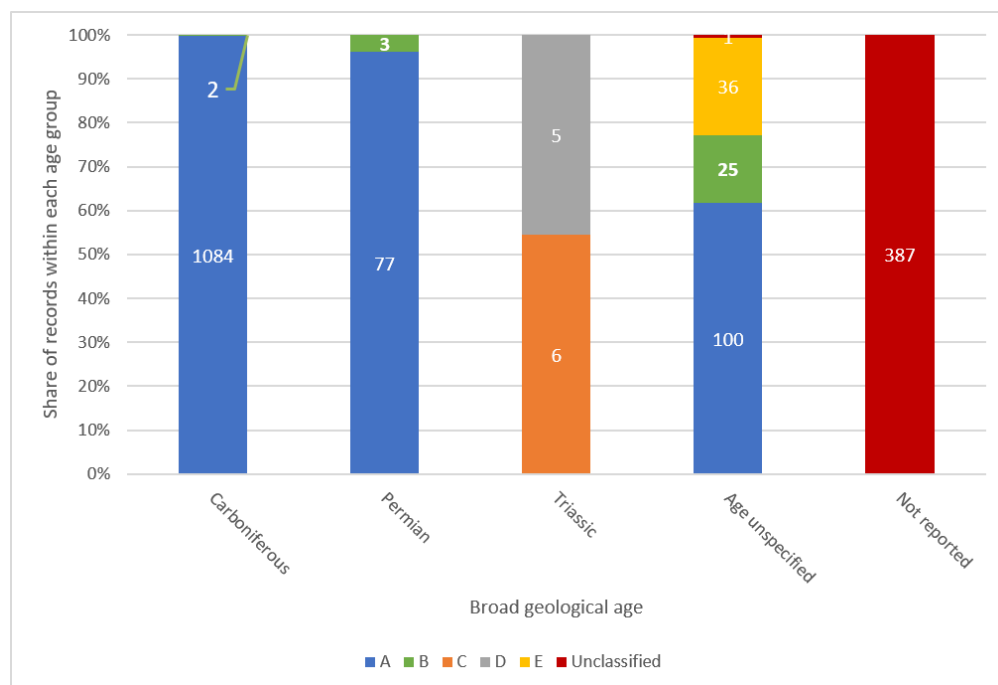


Figure 36 *Distribution of general geological ages assigned to sinkholes related to underground workings.*

Table 9 *Cross-distribution of general geological age and simplified lithological classes for sinkholes related to underground workings.*

General geological age	A	B	C	D	E	Unclassified
Carboniferous	1084	2	0	0	0	0
Permian	77	3	0	0	0	0
Triassic	0	0	6	5	0	0
Age unspecified	100	25	0	0	36	1
Not reported	0	0	0	0	0	387
Total	1261	30	6	5	36	388

The overburden material resistance field was completed by the database operators using five predefined choices: weak, mild, powdery, resistant and highly resistant. As for other categorical variables, some entries did not exactly match these expected categories. They therefore required cleaning and harmonisation, and in a limited number of cases assumptions had to be made to assign the closest consistent resistance class. The cleaned resistance information was then compared with the simplified lithological classes, when both variables were available (Figure 31). Resistance information is available for only 384 of the 1,337 records assigned to lithological classes A-E, corresponding to 28.7% of this subset. Among documented class A records, weakly resistant material predominates (224 cases; 67.1%), whereas class B is mainly highly resistant (6 of 8 documented cases). All documented class C cases are classified as resistant, and all class E cases as powdery.

These patterns remain fragile because classes B and C contain very few documented observations. No resistance information is available for class D.

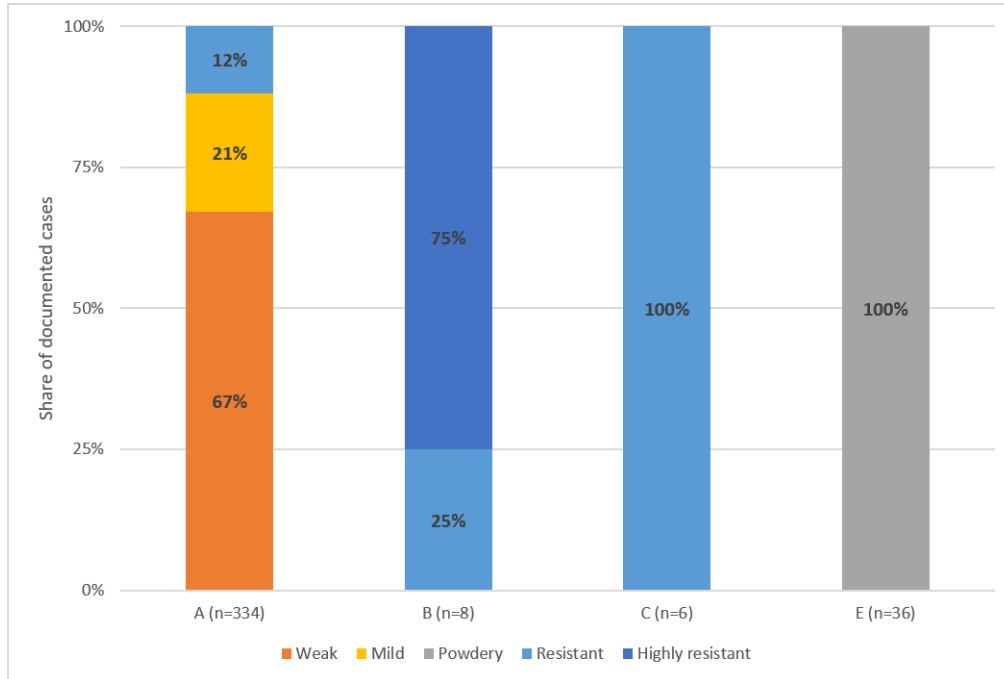


Figure 37 Cross-distribution of cleaned overburden material resistance classes and simplified lithological classes for sinkholes related to underground workings. Only records with both variables documented are included.

The resistance variable must be used carefully. Because it was assigned by individual operators, it may reflect differences in interpretation between contributors rather than a directly measured mechanical property of the overburden. In practice, it remains difficult to infer the actual resistance of the overlying strata from such a global and partly subjective classification.

8.2 Presence of an aquifer in the overburden

The presence of an aquifer in the overburden is documented for 887 of the 1,725 sinkholes related to underground workings, corresponding to a completion rate of 51.4%. Among the full dataset, an aquifer is reported for 39.4% of records, whereas 12.0% indicate no aquifer. The information remains unspecified for 48.6% of cases. Figure 32 summarises this distribution.

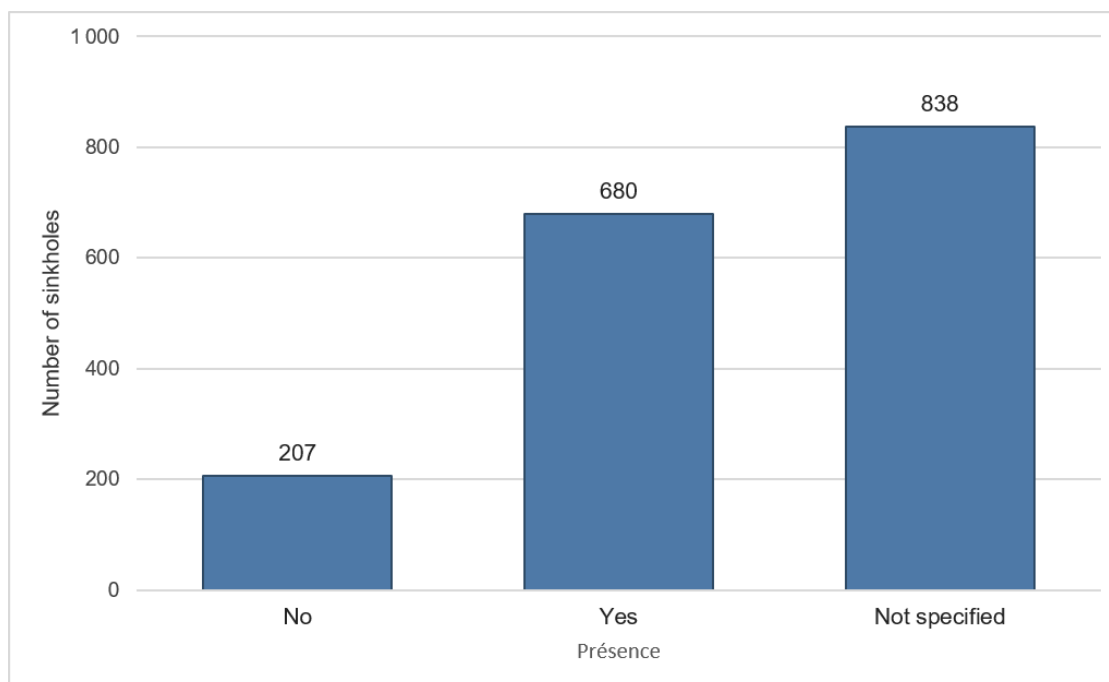


Figure 38 *Distribution of sinkholes according to the reported presence of an aquifer in the overburden.*

The cross-distribution by country shows that the reported presence of an aquifer is highly heterogeneous across the national datasets. Most records with an aquifer are from Poland, with 594 cases, whereas France includes 75 records with an aquifer and 207 records without an aquifer (Table 10). In Germany, only 11 records report the presence of an aquifer and no record indicates its absence. This pattern combines possible geological differences with differences in data availability and reporting practices between countries. In particular, the absence of “No” values for Germany and Poland cannot be read as evidence that aquifers are systematically present; it more likely reflects a limitation of the current inventory.

Table 10 *Distribution of sinkholes according to the reported presence or absence of an aquifer in the overburden, by country. Only records for which the aquifer field is documented as “Yes” or “No” are included.*

Presence of an aquifer by country		
Country	Yes	No
France	75	207
Germany	11	0
Poland	594	0

8.3 Links between overburden characteristics and observed sinkholes

The links between overburden characteristics and observed sinkhole geometry are examined as screening indicators. The database does not provide sufficient control on all relevant geological, mining and hydrogeological factors to infer mechanisms directly from these comparisons.

Working depth was used as a proxy for overburden thickness, as it represents the vertical distance between the underground workings and the ground surface. Median impacted areas were calculated for each thickness class because sinkhole dimensions are highly skewed and include a limited number of very large values. Overburden thickness is available for 1,255 of the 1,725 records, while

impacted area is documented for 1,126 records. The median impacted area increases from 4.2 m² for overburden thicknesses of 10 m or less to 55.1 m² for the 50–100 m class, before decreasing to 7.1 m² for depths greater than 100 m (Figure 39). This pattern suggests that the largest median surface impacts are associated with intermediate overburden thicknesses in the available dataset, but it should not be interpreted as a direct mechanical threshold.

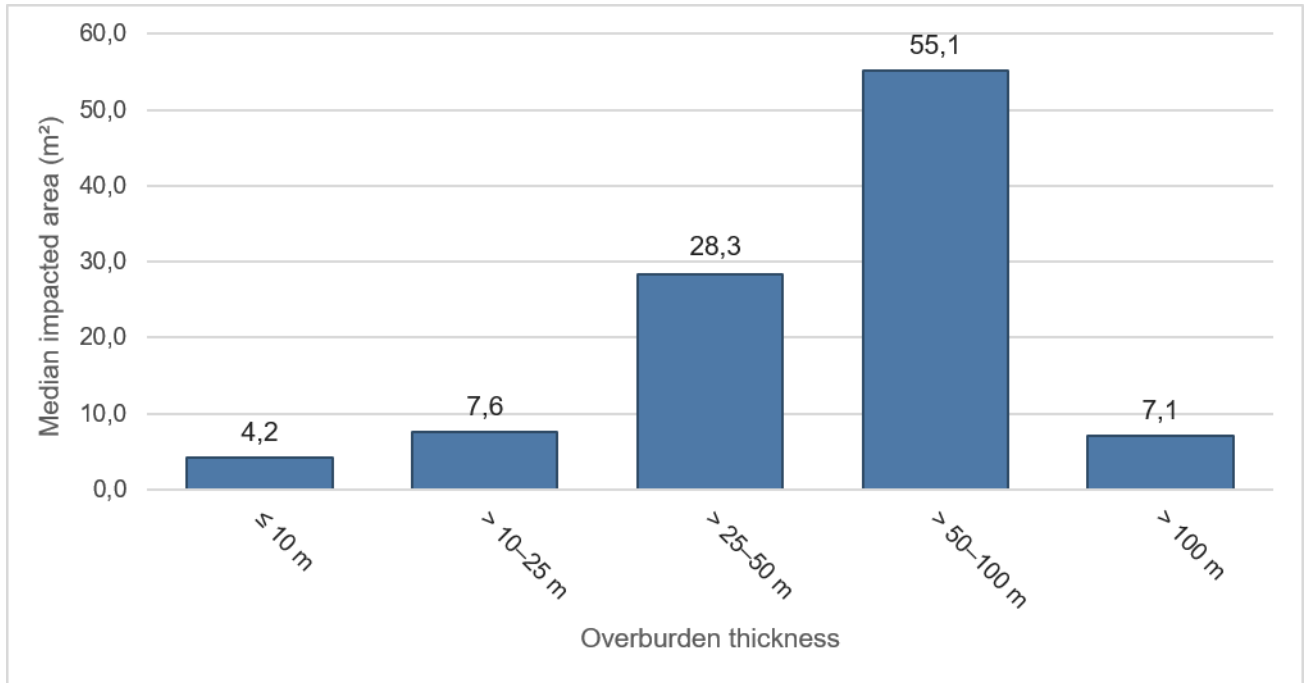


Figure 39 Median impacted surface by overburden-thickness class.

Overburden resistance is documented for 384 of the 1,725 sinkholes related to underground workings, corresponding to a completion rate of 22.3%. Among these records, 301 also include an impacted surface area and can therefore be used to compare median surface impacts by resistance class. The median impacted area decreases from 12.3 m² for weakly resistant overburden to 9.0 m² for mildly resistant material, 4.6 m² for resistant material and 4.0 m² for powdery material. The median value of 9.0 m² obtained for highly resistant overburden is based on only five paired records and should not be treated as a stable class value. The available data suggest larger median impacted areas for weakly resistant overburden, but the low completion rate and unequal sample sizes prevent a firm mechanical interpretation. Figure 40 presents this comparison.

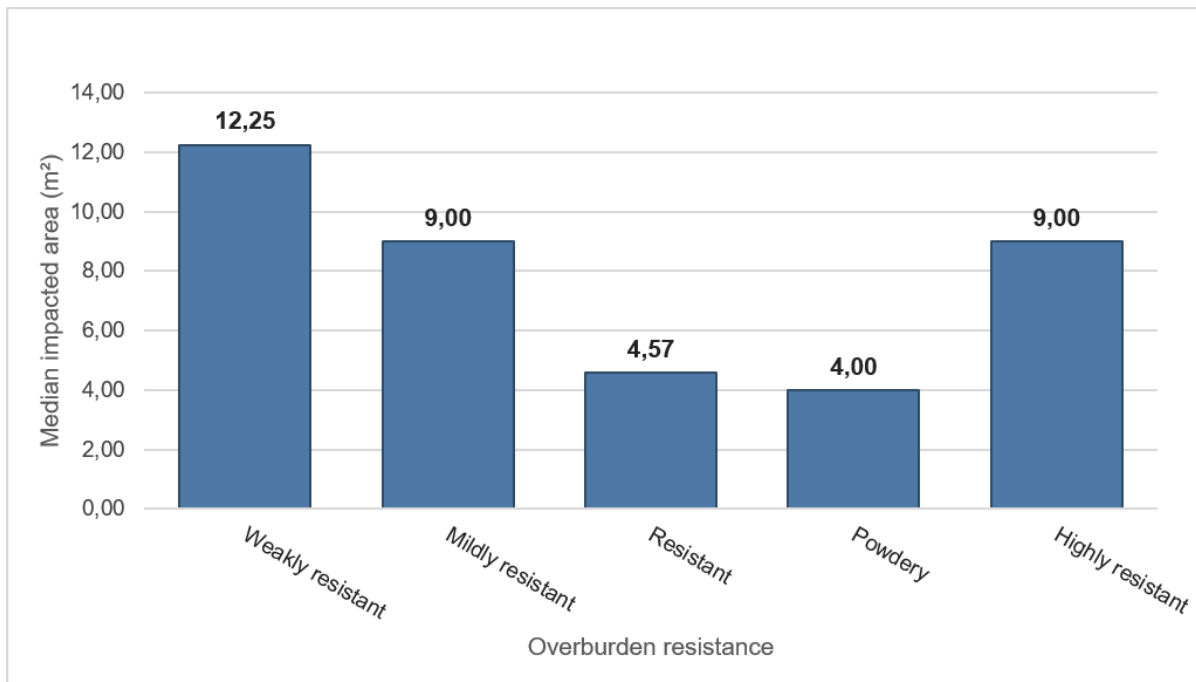


Figure 40 Median impacted surface area by overburden resistance class. Only records with both resistance information and impacted surface area documented are included.

The relationship between overburden resistance and sinkhole depth was examined for records where both variables are documented. This subset includes 227 paired records. Median sinkhole depth ranges from 0.50 m to 2.35 m depending on the resistance class: 1.50 m for weakly resistant overburden, 2.35 m for mildly resistant material, 1.50 m for resistant material, 2.00 m for powdery material and 0.50 m for highly resistant material. No progressive or monotonic trend is observed across the resistance classes. The differences between classes are also weakened by the small number of records in the mildly resistant and highly resistant categories, based on only 16 and 5 paired observations respectively. Within the current dataset, overburden resistance therefore does not appear to control sinkhole depth in a simple way. Figure 35 provides the corresponding comparison.

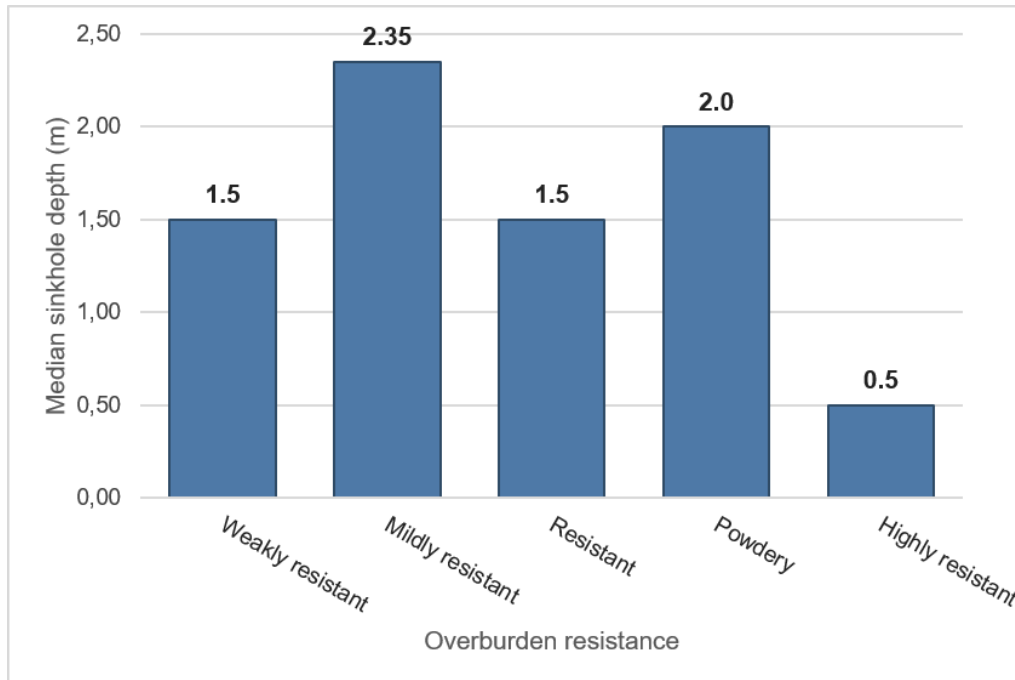


Figure 41 Median sinkhole depth by overburden resistance class. Only records with both resistance information and sinkhole depth documented are included.

9. Statistical relationships between variables

9.1 Scope of the statistical analysis

The statistical analysis follows the selection made in Section 3.3. Only variables with enough documented records were used for the analysis. Variables with very low completion rates, such as void density, ground support, roof fracturing, measurement uncertainty, damage and opening width or diameter of the workings, were not retained as explanatory factors. They are too sparsely documented to support reliable interpretation at the scale of the full database.

Overburden resistance was kept because it is technically relevant, but its low completion rate and partly interpretative nature are kept in mind throughout the analysis. The retained variables were compared with two sinkhole descriptors: impacted surface area and sinkhole depth. Numerical variables were assessed using Spearman's rank correlation coefficient. Spearman's rank correlation coefficient was used because the analysed numerical variables do not follow a normal distribution and include a limited number of large values. As a non-parametric measure, it is more appropriate than Pearson's correlation for assessing monotonic relationships in this exploratory analysis.

Categorical variables were assessed using the eta coefficient after logarithmic transformation of sinkhole dimensions to reduce the influence of very large observations. Because each coefficient is calculated on the subset for which both variables are documented, the reported values do not necessarily refer to the same underlying population. Direct comparisons between coefficients should therefore be made cautiously.

9.2 Relationships with sinkhole dimensions

A link with the impacted surface area is clearly observed for working thickness ($\rho = 0.467$; $n = 920$). The dip of the mined layer ($\rho = 0.330$; $n = 971$), the presence of an aquifer ($\eta = 0.312$; $n = 688$) and working depth ($\rho = 0.313$; $n = 1,063$) also show moderate associations. These variables therefore appear useful for identifying possible relationships with the surface extent of recorded sinkholes. They should not, however, be read as direct evidence of cause and effect.

By contrast, the link with sinkhole depth is weaker. Overburden geology shows the strongest association ($\eta = 0.310$; $n = 1\,027$), followed by overburden resistance ($\eta = 0.231$; $n = 227$). The dip of the mined layer, working thickness and working depth have coefficients close to zero. Within the current dataset, sinkhole depth therefore appears less directly related to the retained mining variables than impacted surface area.

These statistical links remain sensitive to the structure of the database. Missing data, unequal national contributions and the concentration of many records in a small number of mines or basins may all influence the results. Some associations may therefore reflect reporting practices, data access conditions or local inventory strategies rather than general physical relationships.

9.3 Relationships with post-closure delay

A separate analysis was carried out for the delay between mine closure and sinkhole occurrence. It was restricted to the 281 sinkholes recorded after closure. The 385 cases that occurred before closure were excluded because they correspond to active-mining or pre-closure conditions. The delay was calculated as the difference between the normalised year of sinkhole occurrence and the mining end year.

Numerical factors were assessed using Spearman's rank correlation coefficient. Categorical factors were assessed using the eta coefficient after logarithmic transformation of the delay, in order to reduce the influence of exceptionally long intervals. Results based on fewer than 100 paired observations were treated as exploratory.

Overburden geology shows the strongest association with post-closure delay ($\eta = 0.864$; $n = 176$), followed by aquifer presence ($\eta = 0.707$; $n = 165$), and water conditions in the workings ($\eta = 0.380$; $n = 171$). Among numerical mining characteristics, greater working thickness and greater working depth are moderately associated with shorter delays, with Spearman coefficients of -0.310 ($n = 151$) and -0.315 ($n = 168$). The dip of the mined layer shows only a weak positive relationship with delay ($\rho = 0.240$; $n = 156$). These results suggest that post-closure delay may be related to geological, hydrogeological and mining conditions, but the interpretation remains strongly dependent on the composition of the available records.

9.4 Limits

The statistical analysis provides a useful first screening of the variables that may be informative for hazard modelling. Its main limits are missing or uncertain values, differences between national inventories, and the variable number of paired observations available for each comparison. These limits explain why the results should be used to guide further modelling rather than to define causal rules directly.

Improving the analytical value of the database will require controlled and traceable imputation methods, applied only where the missing information can be estimated in a technically justified way. This is the purpose of the complementary work carried out by SYSTRA Subterra in Part 2 of the deliverable, which focuses on more advanced statistical analysis and on imputation approaches intended to improve the completeness, consistency and modelling potential of the SIRIMA database.

10. Main findings from the database analysis

The analysis of the SIRIMA database gives a structured view of the information currently available on sinkholes associated with shallow abandoned coal mine workings. The database is sufficiently organised to describe the main spatial, temporal, geometrical, mining-related and geological characteristics of the recorded events. At the same time, it shows clearly where the evidence remains too sparse or too uneven to support direct interpretation.

The spatial analysis shows a strong concentration of records in a limited number of mining areas. Poland contributes the largest number of underground-working sinkholes, mainly in a few municipalities of the Upper Silesian Coal Basin; the German dataset is largely concentrated in the

Essen area, whereas the French records are more dispersed across a larger number of mines. At the 1 km² grid scale, this concentration becomes even more apparent: a small number of grid squares contains a large share of the recorded sinkholes. This pattern is informative, but it is not a direct map of hazard intensity. It mainly reflects the structure of the available inventories, the location of well-documented basins, and differences in data access and reporting practices between countries.

The inventory also confirms the need to distinguish between sinkholes related to underground workings and those associated with shafts. The first group forms the main analytical population and is large enough to support descriptive statistics and selected cross-analyses. Shaft-related events are fewer and more heterogeneous. They correspond to a different type of mining structure and probably involve different failure mechanisms. For this reason, the two populations should not be merged without qualification in future modelling work.

The best-documented variables are those related to location, mine identification and general spatial context. Several variables describing the sinkholes themselves are also usable, including length or diameter, impacted surface area, depth, occurrence date and reported origin of collapse. Mining end dates, working depth, working thickness, dip of the mined layer, water conditions and broad geological descriptions are documented often enough to support cautious comparison. By contrast, variables such as roof fracturing, ground support, void density, and observed damage remain too incomplete to be used as robust explanatory factors at the scale of the full database.

The dimensional analysis indicates that most recorded sinkholes are small to medium-sized and shallow. This pattern is consistent with the French post-mining hazard mapping approach, where the diameter of localized collapse is a key parameter because it largely controls the expected consequences for people and property within the affected area (Franck et al., 2019). Previous French analyses of more than 1,800 localized mine collapses show that about 90% are smaller than 10 m in diameter, with very large collapses being exceptional and linked to specific mining contexts (Salmon et al., 2015). The SIRIMA database shows a similar predominance of small to medium-sized sinkholes. A limited number of larger events is nevertheless present and should be examined separately because of their potential consequences. The clearest descriptive associations with impacted surface area involve working thickness, working depth, the dip of the mined layer and the reported presence of an aquifer. Associations with sinkhole depth are weaker. Overburden geology and overburden resistance show some statistical relationship with depth, but the available data do not support a simple mechanical interpretation.

The temporal analysis adds an important perspective. The available records show that sinkholes may occur not only during or shortly after mining activity, but also several decades after mine closure, and in some cases more than a century later. However, the temporal structure of the database is strongly affected by the quality and availability of archives, by relative or incomplete dates, and by national reporting practices. The apparent differences between countries in post-closure delay must therefore be interpreted carefully, as they may reflect the composition of the inventory as much as physical differences between mining basins.

Overall, the results provide a solid descriptive basis for selecting variables that may be useful in post-mining sinkhole hazard modelling. Some of the variables identified by the analysis stand out as expected, since they have long been recognised as predisposition factors for sinkhole occurrence in post-mining contexts. This is notably the case for mining-related parameters such as working depth, working thickness or the geometry of the exploited layer (Tritsch, 2005). However, the results also show that other variables should be considered more explicitly in future hazard modelling. In particular, the geology of the overburden, and possibly the reported presence of an aquifer within it, appear to provide relevant information. These variables suggest that the nature, structure and hydrogeological context of the cover above the workings may influence the characteristics of the observed sinkholes but also on the mechanism controlling upward propagation and should not be treated as secondary descriptors only. The main constraints are the heterogeneous completeness of the variables, the imbalance between countries, the concentration of many records in a small number of mines or basins, and the changing number of paired observations available for each comparison.

Part 2

Operational Tool for Predicting of Sinkhole Hazard

11. General overview

The phenomenon of sinkhole formation in post-mining areas represents one of the most profound and unpredictable geological hazards. As the legacy of centuries of intensive subterranean extraction, specifically within shallow coal mining environments, the earth's subsurface is frequently left with abandoned voids, collapsed galleries, and deteriorating infrastructural supports. Consequently, the proper management of the coal mine abandoning process is essential for ensuring the safety of subsequent land use (Misa et al., 2025).

Surface deformations in mining regions manifest in two primary forms: continuous subsidence, which presents as broad, trough-like depressions, and discontinuous subsidence, which presents as localized, abrupt sinkholes or chimney caving. Continuous subsidence is the typical result of deep, modern longwall mining with caving, where the extraction of a wide coal seam causes the immediate overlying strata to collapse, fracture, and ultimately bend in a continuous profile reaching the surface (Yan et al., 2023). Conversely, discontinuous deformations emerge almost exclusively from shallow mining operations, particularly those employing historical room-and-pillar techniques or rudimentary drift mining. Over decades or centuries, these pillars and timber support undergo progressive geomechanical deterioration due to weathering, stress redistribution, and biodegradation (Zhuang et al., 2025).

When sinkholes materialize in densely urbanized regions, their impacts are multi-dimensional and devastating. They compromise the structural foundations of commercial and residential buildings (Marino, 2020), rupture sensitive subterranean utility networks (including water, gas, and electrical lines), and sever critical transportation arteries (Clarke et al., 2006). Beyond the immense economic ramifications, which can run into hundreds of millions of euros in highly susceptible regions, these events inflict profound psychological distress, anxiety, and a chronic sense of insecurity on local populations, thereby severely undermining public confidence in regional land-management and municipal authorities.

To effectively combat this escalating geohazard, the paradigm of risk management must transition from reactive emergency response to proactive, data-driven predictive methodologies. To address this, an Operational Tool for predicting Sinkhole Hazard is proposed. This tool is a machine learning, and statistical model that has been developed to help various stakeholders better predict the sinkhole hazard, a capability that is particularly vital when dealing with historical sites characterized by limited, fragmented data.

This comprehensive research report proposes and details the solution developed to meet an automated, predictive tool encapsulated within a web-based application. The developed framework systematically integrates data processing techniques, statistical evaluations of critical variables and algorithmic threshold optimizations (Mazumder et al., 2025). To accomplish this, a dual-modelling approach to define the Hazard Index is proposed.

The susceptibility model is expected to determine the probability of occurrence of an event and the magnitude model to predict the potential severity and the probability of size of the sinkhole. The study is based on an inventory of 1,868 observations across France, Poland, and Germany. The pipeline is divided into four core analytical modules: susceptibility assessment, magnitude prediction, and hazard Integration, culminating in a deployment-ready algorithmic engine. The final Hazard Index is expected to be a 3X3 matrix that crosses susceptibility model with magnitude model:

Magnitude / Susceptibility	Low	Medium	High
High Magnitude (M3)	Moderate	High	Critical
Medium Magnitude (M2)	Low	Moderate	High
Low Magnitude (M1)	Very Low	Low	Moderate

Figure 42. Hazard matrix derived from the integration of susceptibility and magnitude models

12. The Susceptibility Model

The susceptibility model explains the potential of a sinkhole occurrence. This was accomplished through a weighted sum approach (Michael & Samanta, 2016) that integrates objective variables with expert knowledge to define predictive model behaviour. The model does not evaluate the potential physical size or destructive capacity of the event; rather, it focuses exclusively on mapping the preconditions that facilitate instability.

12.1 Selection of variables of interest

The selection of conditioning variables for the susceptibility model is based on a convergence of empirical principles of mining geomechanics and the statistical findings detailed in Part 1 of this document. As demonstrated in the exploratory phase, variables relating to local geology and rock mechanic show the clearest statistical relationship with sinkhole occurrence. Building on these relationships, three critical variables govern the physical behaviour of sinkhole formation: the depth of the drift relative to the surface (overburden thickness), the working thickness, and the overburden resistance material (Kotyrba 2005, Kotyrba & Kortas, 2016; Karagüzel et al., 2020; Orhan et al., 2020).

The selection of these three parameters responds directly to rock mechanics: depth dictates the proximity of the drift/gallery to the surface and defines whether the caving zone has the potential to collapse; working thickness represents the volume of the induced void and the potential energy of the failure; and overburden resistance material determines the ground's capacity to sustain the load-bearing bridge or yield to applied stresses.

12.2 Material overburden resistance

This variable represents the bridging capacity of the rock mass above the mining void. As previously detailed in Part1, section 8.1, the raw dataset originally contained textual descriptions (e.g., “weakly resistant”, “powdery”, “highly resistant”) that required significant harmonisation to establish a valid numerical scale mapped into discrete levels: 1 (Weakly Resistant), 2 (Mildly Resistant), and 3 (Resistant) being powdery/weakly resistant assigned to level 1 and resistant/highly resistant to level 3.

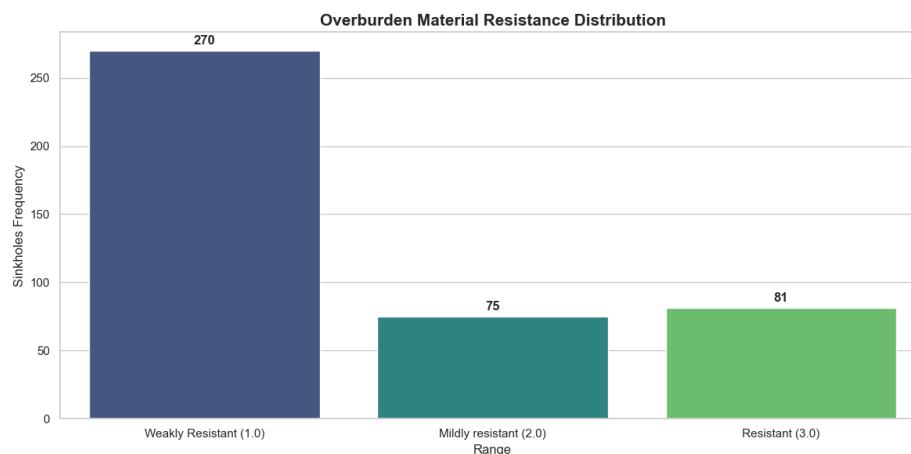


Figure 43. Bar plot of discrete overburden material resistance variable

As shown in Figure 43, only 426 of the initial 823 records remained non-null after data cleaning and transformation. This reduction occurred because entries originally labelled as 'Unknown' were converted to not a number value to ensure technical accuracy. Due to lithology description and overburden resistance represent critical variables for predicting sinkhole occurrence (as discussed in sections 9 and 10), an imputation strategy was developed that leverages the parallel *geology of the overburden* variable.

12.3 Imputation of material resistance variable

The *geology of the overburden* variable contained 1475 non-null valid data and was identified as a critical feature for imputing missing values in the *overburden resistance* variable. The initial strategy for addressing missing data relied on a lexical analysis aimed at identifying specific keywords associated with rock competence. By isolating key geological terms such as "sandstone," "shale," or "limestone" within the lithological descriptions to correlate their frequencies with known resistance classes. However, this approach was insufficient. A frequency analysis across the three designated resistance levels showed that individual geological terms were distributed uniformly, appearing with comparable frequency across all entries, regardless of whether they were classified as low or high resistance.

Table 11 Keywords count for every overburden material resistance level.

Keyword	Weakly Resistant (1)	Mildly Resistant (2)	Resistant (3)
sandstone	234	76	63
shale	122	76	27
clay	122	1	25
carboniferous	77	0	31
limestone	0	0	17
sand	270	76	63
coal	23	59	11
conglomerates	23	75	11

This lack of clear terminological segregation demonstrated that the resistance of the overburden is not governed by single keywords, but rather by the complex, contextual lithological sequence described in the geological records.

To overcome the limitations of keyword-based mapping, a more sophisticated, integrated approach was adopted through an unsupervised Machine Learning pipeline. Instead of evaluating isolated words, a methodology combining Natural Language Processing (NLP) with a K-Means clustering algorithm was implemented. First, a vectorizer was utilized to transform the raw, unstructured geological text into a mathematical feature matrix. Subsequently, an unsupervised K-Means clustering algorithm was deployed to group similar geological profiles into distinct clusters (Al-Obaydy et al., 2022; Wang et al., 2022). By calculating the average resistance class for the known records within each of the four generated clusters, a statistically informed rules to fill the gaps in the dataset was established. This data-driven approach populated the missing values, rescuing hundreds of records for the susceptibility model without introducing arbitrary human bias.

Table 12. Serial clustering keywords and average resistance calculations

Average Resistance	Known Data	Missing Data	Keywords
1.00	53	887	carboniferous, quaternary, shale, sands, clays, sandstone, triassic, topped, westphalian, shales, siltstones, upper, silty, sandstones, schistose, sand, quartz, puddingstone, predominantly, stephanian, whesphalien, pebbles, permian, molasse, middle, locally, limestones, mezozic, marls, labradorite, keuper, iridescent, glacial, geological, fragments, limestone, levels, formation, formed, followed, facies, dolomite, deposits, eocene, dolomites, conglomerates, conglomeratic, clayey, coalfields, clay, bituminous, alluvium, age
1.67	187	45	molasse, middle, westphalian, silty, puddingstone, stephanian, sandstone, clayey, limestone, clay, schistose, facies, levels, sands, quartz, geological, followed, formation, whesphalien, labradorite, predominantly, conglomeratic, dolomite, shale, marls, keuper, conglomerates, glacial, deposits, carboniferous, pebbles, fragments, bituminous, formed, upper, iridescent, limestones, dolomites, triassic, shales, quaternary, clays, topped, siltstones, sandstones, permian, sand, mezozic, locally, eocene, coalfields, alluvium, age
1.87	92	40	shales, sandstones, coalfields, conglomerates, topped, locally, age, alluvium, eocene, sand, limestone, clayey, mezozic, siltstones, clays, carboniferous, triassic, stephanian, upper, westphalian, whesphalien, pebbles, permian, predominantly, puddingstone, quartz, quaternary, sands, sandstone, silty, schistose, shale, molasse, marls, middle, limestones, levels, formation, formed, fragments, geological, glacial, iridescent, keuper, labradorite, conglomeratic, deposits, dolomite, dolomites, followed, facies, clay, bituminous
1.34	94	77	permian, sandstone, shale, triassic, clays, carboniferous, quaternary, sands, limestone, topped, upper, westphalian, whesphalien, sandstones, schistose, siltstones, sand, quartz, shales, stephanian, silty, pebbles, predominantly, puddingstone, molasse, marls, locally, mezozic, middle, labradorite, keuper, iridescent, glacial, geological, fragments, limestones, levels, formation, formed, followed, facies, dolomite, deposits, eocene, dolomites, conglomerates, conglomeratic, clayey, coalfields, clay, bituminous, alluvium, age

Missing values in the *overburden material resistance* variable were then imputed based on the assigned cluster of their corresponding average resistance and geological description (Table 12). This methodological shift ensured that the imputed values were coherent with the broader lithological context of each record. As a final refinement, these relationships were encapsulated into a supervised Random Forest classifier (Trott et al., 2022), ensuring that the model could generalize this contextual learning to predict resistance levels for future geological descriptions. This encapsulation will be explained in detail in section 15.

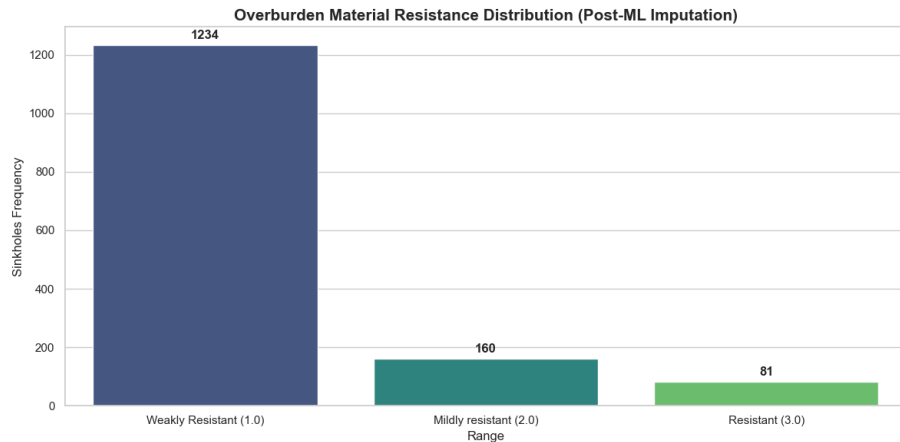


Figure 44. Overburden material resistance distribution post-ML imputation.

Clearly, all the observations are concentrated on the weakly resistant level. This imbalance likely stems from the lack of records for historical events or from an observer bias during sinkhole recording. Consequently, relying on these raw statistical frequencies would introduce a massive bias into the predictive algorithm, artificially underestimating the risk in more competent rock masses due to historical underreporting. As a direct response to this limitation, assigning expert weights acts as a corrective measure within the weighted sum susceptibility model.

12.4 Depth of the drift relative to the surface

The depth of the drift defines the proximity of the underground void to the surface and directly influences the likelihood of a collapse propagating upward. As detailed in Part 1, Sections 2.2 and 7.1, the raw data for this variable underwent extensive parsing to resolve formatting heterogeneities and convert qualitative descriptors, ranges, or inequalities into a standardized continuous numerical format. The exploratory analysis in Part 1 also established a heavily right-skewed distribution, reflecting both observational biases (shallow sinkholes are more easily detected and recorded) and geomechanically realities, such as the bulking factor of caved rock preventing deeper collapses from reaching the surface.

To integrate this parameter into the operational predictive framework, the corrected continuous depth values were first discretized into four geomechanically meaningful intervals based on standard engineering criteria: Very Shallow (0-15m), Shallow (15-30m), Intermediate (30-60m), and Deep (>60m).

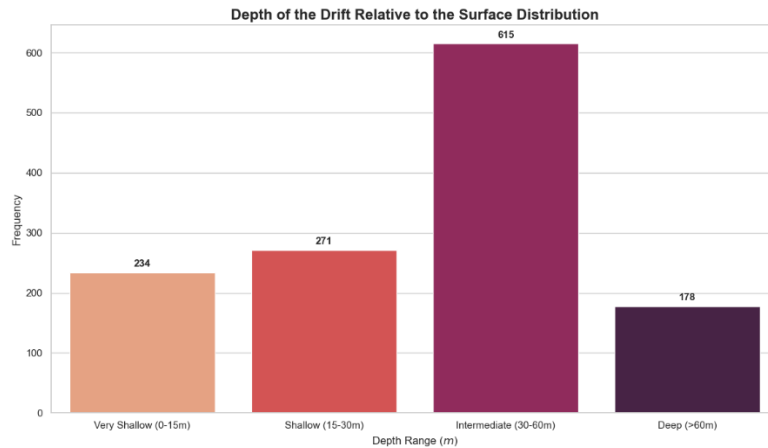


Figure 45. Depth of the drift relative to the surface frequency distribution

It is imperative to note that the frequency distribution is strictly statistical and does not account for implicit biases arising from the incompleteness of the historical catalogue. Several factors may contribute to this observational bias. First, a shallow sinkhole (e.g., at 30 meters) is inherently more likely to be detected and officially recorded than a deep subterranean anomaly (e.g., at 200 meters). Second, collapses occurring over ancient, very shallow mining operations often lack formal documentation, leading to a historical underreporting in the shallowest strata. Finally, from a geomechanical perspective, collapses initiated in deeper mines are less likely to propagate to the surface due to the bulking factor of the caved rock (Palchik, 2015), which naturally expands and fills the void before reaching the topsoil. To analyse the statistical distribution of the working depth, a box plot and kernel density estimation (KDE) visualizations were generated.

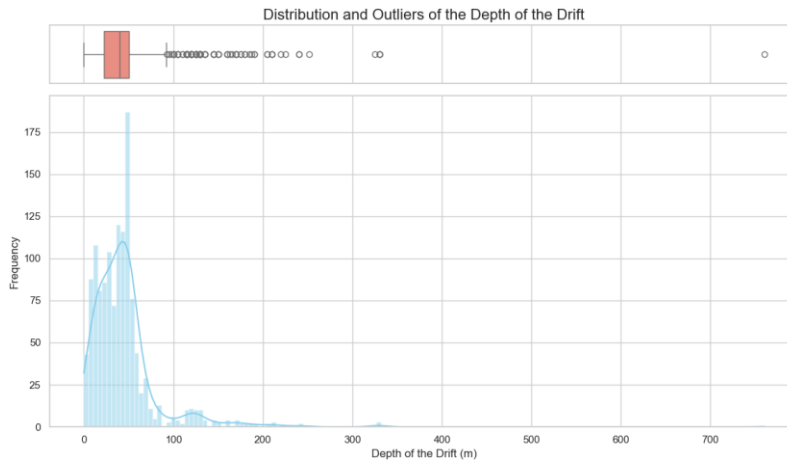


Figure 46. Distribution and outliers of the working depth variable

These analyses revealed a heavily right-skewed distribution, with most of the values highly concentrated between 0 and approximately 70 meters. However, the distribution also exhibits a pronounced, elongated tail extending to extreme values exceeding 700 meters. Although these observations were retained within the dataset as they represent valid, real-world deep mining operations rather than measurement errors. To effectively integrate this highly skewed continuous variable into the weighted sum susceptibility model without disproportionately biasing the algorithm, an inverse logarithmic normalization technique was applied and explained in more detail in section 12.8.

12.5 Coal working thickness

This variable refers to the thickness of the void space created by coal extraction activities. As established during the baseline descriptive analysis in Part 1, Sections 2.2 and 7.2, the raw data for this parameter required extensive parsing to resolve formatting inconsistencies and exhibited a distribution heavily influenced by specific Polish mines (Siersza and Jan Kanty). While the fundamental harmonisation was covered in Part 1, integrating this variable into the operational predictive framework required a targeted cross-variable reconciliation process.

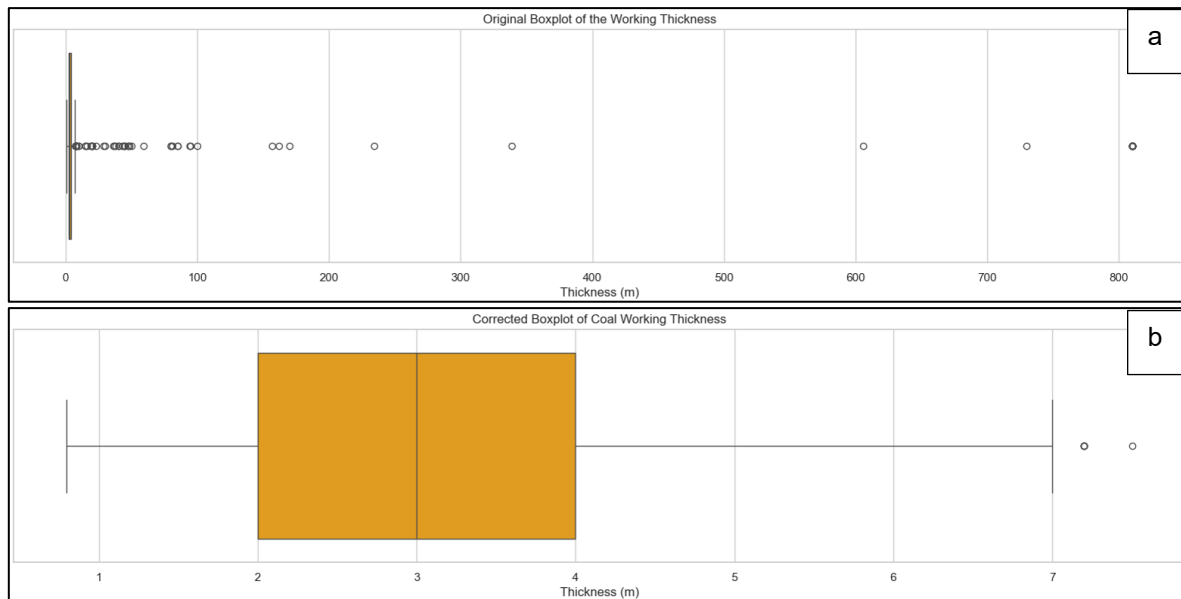


Figure 47. Comparison of boxplots for the coal working thickness variable before and after corrections. (a) Original boxplot, and (b) Corrected boxplot.

Following the initial data extraction, a descriptive statistical analysis revealed critical anomalies, with certain working thickness values reaching physically impossible dimensions for European coal seams (e.g., up to 810 meters – Figure 47/a). A cross-variable evaluation determined that these extreme outliers closely matched the expected working depths, exposing a systematic data transposition error that occurred during the historical digitization of the mining catalogue. To rectify this structural flaw without discarding valuable data, unrealistic thickness values (≥ 8.0 meters) were nullified within the thickness array and simultaneously transposed to impute missing records in the depth of the drift variable. This cross-variable reconciliation successfully purged the thickness dataset of invalid outliers while recovering essential spatial information.

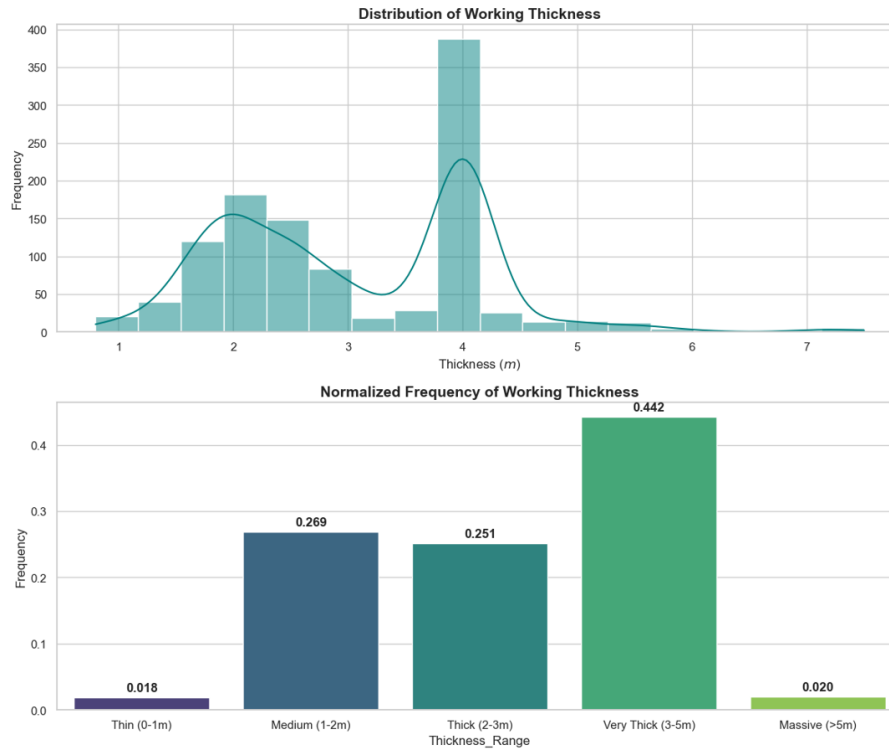


Figure 48. Distribution of coal working thickness after correction

With the variable structurally corrected, the continuous thickness values were discretized into five representative intervals: Thin (0-1m), Medium (1-2m), Thick (2-3m), Very Thick (3-5m), and Massive (>5m). A normalized frequency plot was subsequently conducted to calculate the proportional occurrence of historical sinkholes within each specific interval (Figure 48). The distribution of the coal working thickness reveals a bimodal behaviour, and the most prominent frequency peak occurs at approximately 4 meters; a verification of the underlying data confirms that the records originate from two specific locations in Poland: The Siersza and Jan Kanty hard coal mines. Consequently, it becomes evident that the statistical representation of the working thickness variable is inherently biased by this database imbalance.

12.6 Multicollinearity analysis

To guarantee the robustness of the susceptibility model based on the weighted sum method, it was imperative to ensure the statistical independence of the predictor variables. The inclusion of highly correlated parameters could distort the mathematical weighting of the algorithm, leading to an artificial overestimation of certain variable effects in the final risk calculation. Therefore, before consolidating the final parameters, a comprehensive multicollinearity analysis (Hair et al., 2010) was conducted on the set of structural variables, which included the depth of the drift, the coal working thickness, and the overburden resistance.

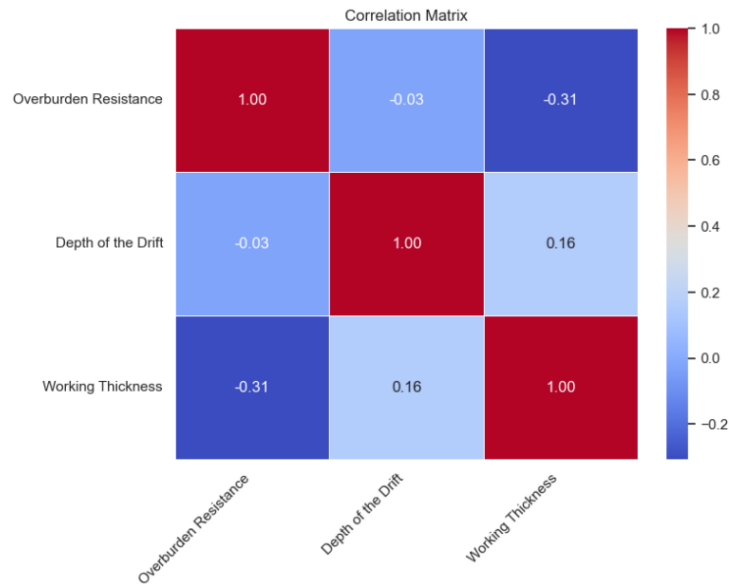


Figure 49. Correlation matrix of conditioning variables of susceptibility model

The first phase of this analysis involved the development and evaluation of a Pearson correlation matrix. This statistical tool allowed for the measurement and quantification of the linear relationships between all possible pairs of variables within the historical database. By establishing a strict correlation threshold to identify redundancies, the matrix results revealed that the selected predictor variables exhibited low linear correlation with one another reaching values between 3% and 31%.

Table 13. Criteria for the Variance Inflation Factor

VIF = 1	No correlation.
1 < VIF < 5	Moderate correlation. Generally acceptable and does not require variable elimination.
VIF > 5	Critical multicollinearity. The model coefficients may be unreliable.
VIF > 10	Must eliminate one of the variables involved because they are redundant.

To definitively validate the two-dimensional results of the Pearson matrix and account for simultaneous interactions among all parameters, the Variance Inflation Factor (VIF) was calculated (Hair et al., 2010; Joshi et al., 2026). The VIF is an advanced metric that quantifies the extent to which the variance of a model is inflated due to linear correlation among multiple predictors. Adhering to statistical modelling standards, a critical acceptance threshold was established according to the table 14.

Table 14. VIF results for the susceptibility variables

Variable	VIF
Working Thickness	1.13
Overburden Resistance	1.11
Depth of the Drift	1.03

The calculations demonstrated that all variables yielded VIF values substantially below the threshold. This dual statistical validation ruled out any risk of severe multicollinearity, mathematically justifying

the safe integration of these three independent variables into the operational algorithm. This empirical finding validated the premise that both the geometry of the void (depth of the drift and thickness of the coal workings) and the competence of the ground (resistance) capture distinct and independent geomechanical phenomena during the formation of a sinkhole.

12.7 Balance of susceptibility variables

Despite the rigorous data cleaning procedures applied to the 1,868 total recorded events, an evaluation of the dataset revealed significant intersectional gaps among the three core variables required for the susceptibility model.

Table 15. Completeness of conditioning variables of the susceptibility model

Variable	Non-Null Count	Data Type
Id_bd_sinkhole	1868	Object
Longitude (wgs84)	1868	Float
Latitude (wgs84)	1868	Float
Overburden Resistance	1475	Float
Depth of the Drift	1339	Float
Working Thickness	1102	Float

While the geographical coordinates (latitude, longitude) are completely populated for all entries, the predictive variables present varying degrees of incompleteness: The overburden resistance contains 1,475 non-null values, depth of the drift has 1,339, and working thickness is limited to 1,102. Because the weighted sum algorithm strictly requires concurrent values across all three parameters to calculate a valid spatial risk score, simply discarding incomplete rows would result in an unacceptable loss of critical historical data. Consequently, a spatial imputation strategy was adopted to bridge this lack of information; missing variables were systematically filled by their spatial geological knowledge and geomechanical trends.

To address the incompleteness of the data within the fundamental variables of the susceptibility model, a spatial imputation strategy based on the principle of geological autocorrelation was designed and implemented. The foundational premise of this approach is that rock mass characteristics, stratigraphy, and historical mining methodologies tend to maintain strong consistency within a specific local basin or district. Therefore, the geomechanical conditions of an undocumented event can be deduced with high precision from well-documented collapses. The first step in this workflow was the transformation of the events' original geographic coordinates (latitude and longitude in WGS84) into a Universal Transverse Mercator (UTM) ETRS89-LAEA. This conversion was an indispensable mathematical requirement to translate the data into a Euclidean space, enabling the algorithm to calculate true physical distances in meters rather than angular distortions, thereby guaranteeing the geometric accuracy of the spatial model.

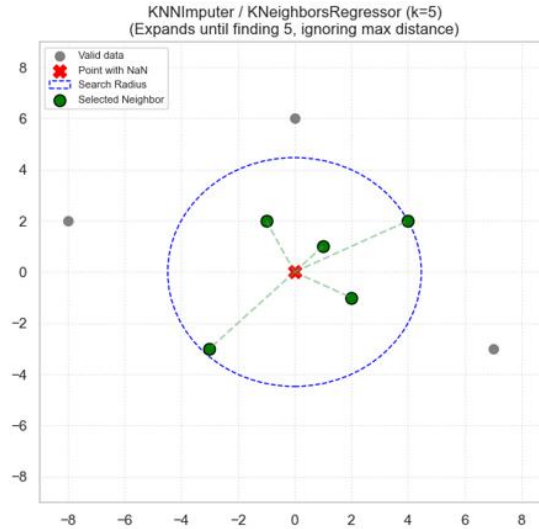


Figure 50. Representation of applied imputation strategy

Once the dataset was projected, a spatial machine learning imputation approach was applied utilizing the integrated Nearest Neighbors and K-Nearest Neighbors algorithms (Ahmed et al., 2023). To prevent the model from averaging incompatible data originating from distant geological regions, the algorithm was calibrated to operate exclusively within a strictly defined radius of influence, establishing a maximum search distance of 5 kilometers ($k = 5$ km). This spatial boundary was not arbitrary; it represents a coherent geological domain scale, ensuring that the nearest neighbor algorithm only utilized information from adjacent mines and rock strata that share an identical structural, lithological, and operational context to estimate the missing values of the depth of the drift, coal working thickness, and overburden material resistance.

Through this configuration, the K-Nearest Neighbors algorithm iteratively evaluated each incomplete record, identifying valid sinkholes within the 5 km radius and calculating an estimated value based on the spatial proximity of those neighbors. This data-driven methodology successfully filled the information gaps using exclusively local spatial geological knowledge. As a direct result of this spatial interpolation technique, the structural integrity of the database was recovered, providing the weighted sum susceptibility model with a complete set of continuous parameters without sacrificing fidelity of the historical records.

Table 16. Completeness of conditioning variables after imputation

Variable	Non-Null Count	Data Type
Id_bd_sinkhole	1501	Object
X (LAEA)	1501	Float
Y (LAEA)	1501	Float
Overburden Resistance	1501	Float
Depth of the Drift	1501	Float
Working Thickness	1501	Float

Finally, a quality control (QC) phase was implemented to validate the integrity of the algorithmic outputs through a comprehensive visual verification utilizing a point-by-point geographic mapping approach.

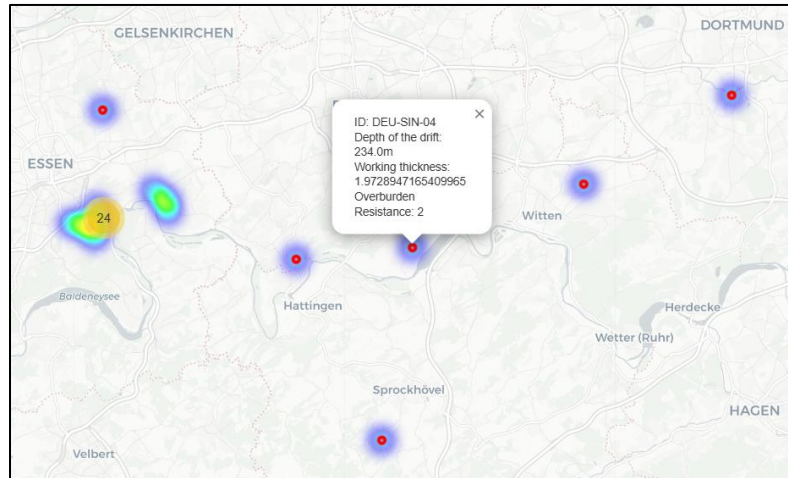


Figure 51. Point-to-point verification of imputed values

By plotting the imputed parameters alongside the original historical data on a localized spatial map, it was possible to manually inspect the geomechanical coherence of the results. This mapped verification process served to confirm that the estimated values for the depth of the drift, the coal working thickness, and the overburden resistance seamlessly aligned with the known geological and operational trends of their immediate surroundings.

12.8 Normalization process

To integrate the conditioning variables into a unified mathematical framework for the weighted sum model, it was necessary to transform all variables onto a standardized 0-to-1 scale. However, applying a single mathematical transformation across the board would have ignored the unique physical and statistical behaviours of each parameter. For the depth of the drift variable, an inverse logarithmic min-max normalization was implemented. As established in the distribution analysis (Figure 46), the depth data exhibited a severe right-skew with a prolonged tail of deep-mining outliers. A standard linear scaler would have disproportionately compressed the shallow data points into an indistinguishable narrow band. The logarithmic transformation effectively compressed this extreme variance without discarding the outliers.

In contrast, the coal working thickness variable required a different mathematical treatment and was transformed using a standard linear min-max normalization. From a phenomenological perspective, the thickness of the extracted seam exhibits a direct, proportional relationship with the potential energy of the failure due to the total volumetric of the subterranean void. Simply put, larger extraction volumes are expected to generate greater instability and a higher likelihood of upward caving propagation. This linear approach ensures that the dimensional reality of the mining voids was proportionally represented in the final algorithmic weights.

As previously identified, the overburden material resistance variable was severely imbalanced, heavily concentrated in the "weakly resistant" category (Figure 44). Applying a standard mathematical normalization based on data distribution would have intrinsically transferred the catalogue's observational bias into the predictive model, artificially skewing the risk assessment. To smooth this critical flaw, an expert-based normalization criterion was adopted. Instead of relying on raw statistical frequencies, normalization values were assigned to the discrete resistance classes based on their proportional expert criteria weights. However, it should be noted that these proportions require further refinement through a more robust study regarding the resistance of the overburden material, as these discrete classes are currently supported solely by observational criteria.

12.9 Susceptibility Score

To synthesize the individual parameters into a unified framework, the model employs an empirical data-driven approach combined with expert weighting (w) to calculate the Susceptibility Score (S). The calculation was performed as a weighted sum of each conditioning variable (F). The formula used is:

$$S = \sum F * w$$
$$S = F_{depth} * \alpha + F_{thickness} * \beta + F_{resistance} * \gamma$$

Weights were assigned based on technical expert judgement, ensuring the sum equals 1:

- α (Depth of the drift): 0.4
- β (Working thickness): 0.3
- γ (Overburden resistance): 0.3

These weights (α, β, γ) were assigned based on expert-driven proportional principles rather than statistical frequencies, an approach chosen to mitigate the observational biases and incompleteness of the historical catalog.

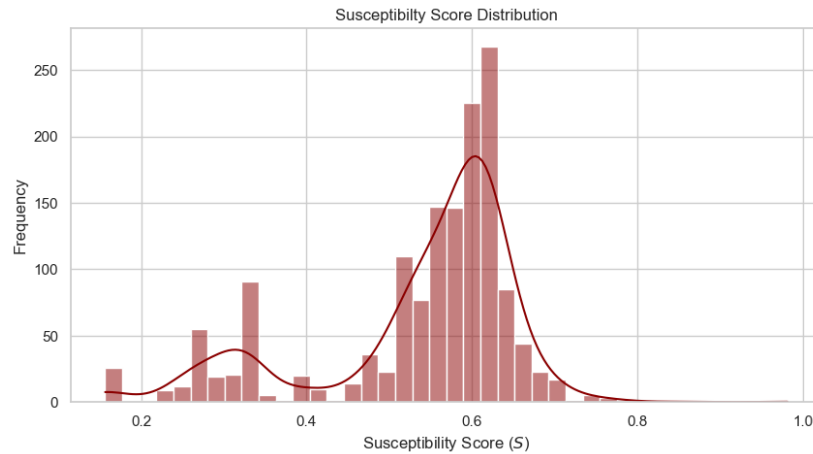


Figure 52. Distribution of susceptibility score

The distribution curve was generated to analyse the statistical behaviour of the final weighted results. The Figure 52 provide a continuous probability density profile of the calculated susceptibility scores across the entire historical dataset, revealing the central tendencies and the natural clustering of the spatial risk metrics.

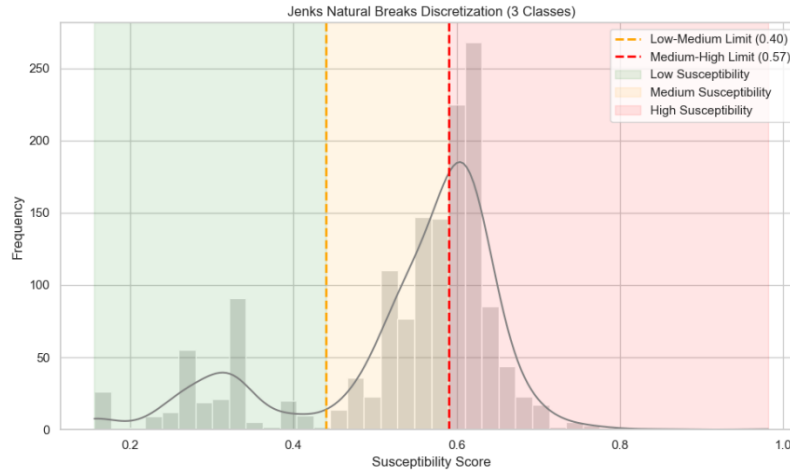


Figure 53. Thresholds definition for susceptibility score

While a continuous probabilistic value between 0 and 1 offers high mathematical precision, it is inherently difficult to interpret within the context of a practical Decision Support System (DSS). Consequently, it became necessary to translate these continuous mathematical probabilities into discrete values specifically, Low, Medium, and High susceptibility levels (Figure 53). To avoid the introduction of arbitrary cut-off points that could distort the risk assessment, the discretization process was governed by the Jenks Natural Breaks optimization algorithm (Ke et al., 2023). This geostatistical method mathematically identifies optimal classification thresholds by minimizing the variance within each class while maximizing the variance between different classes, ensuring that the defined categories accurately reflect the intrinsic groupings of the empirical data. As a result, the following thresholds were selected: values < 0.40 define low susceptibility, values between 0.40 and 0.57 define medium susceptibility, and values > 0.57 define high susceptibility.

Furthermore, a model stability analysis was conducted to validate these Jenks-derived boundaries. The analysis was executed utilizing a repeated random subsampling methodology. Specifically, a simulation comprising 100 iterations was performed, wherein the dataset was randomly partitioned during each cycle, allocating 70% of the data to compute the Jenks thresholds while withholding a 30% subset to deliberately introduce structural variability. Across these 100 independent realizations, the transitional boundary values separating the Low-Medium (L1) and Medium-High (L2) susceptibility classes were systematically recorded.

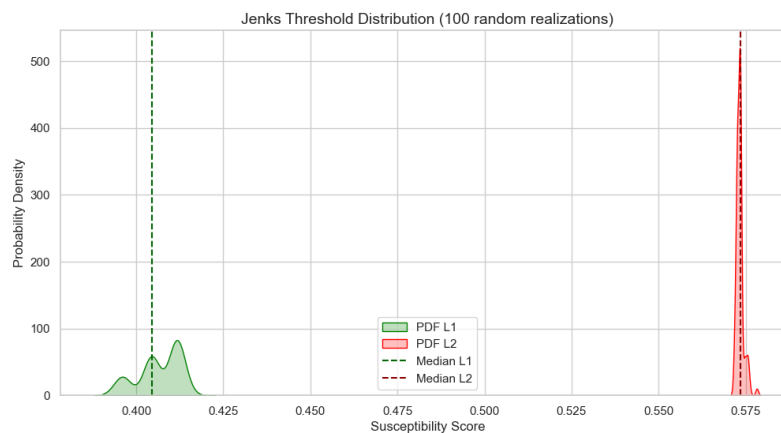


Figure 54. Model stability analysis. Thresholds variation across 100 random realizations

By evaluating the descriptive statistics of these iterative outputs (including the mean, median, and standard deviation) and mapping their Probability Density Functions (PDF), the analysis quantified the sensitivity of the thresholds to underlying data fluctuations. The resulting probability distributions exhibited minimal variance and a tight concentration around their respective medians, empirically demonstrating that the established Jenks thresholds are intrinsically stable, resilient to localized data anomalies, and mathematically reliable for definitive integration into the predictive model.

Table 17. Descriptive statistics for model stability analysis

Statistics	L1 Low-Medium	L2 Medium-High
mean	0.407	0.573
std	0.006	0.001
median	0.404	0.573
min	0.395	0.572
max	0.416	0.578

12.10 Susceptibility Model Results

To visualize the mathematical behaviour of the weighted sum model, scatter plots were generated mapping the final susceptibility score against the primary geometric variables, with data points coloured by the discrete overburden resistance classes. The relationship between the depth of the drift variable and the calculated susceptibility score illustrates that as the depth of the mining void increases, the susceptibility score exhibits a sharp, non-linear decline, confirming that shallow legacy workings dominate the highest hazard tiers. Furthermore, the colour overburden resistance variable reveals a stratification; events characterized by weakly resistant overburden (Class 1) consistently form the uppermost boundary of the point cloud, mathematically pushing the susceptibility scores higher compared to highly resistant rock masses (Class 3) situated at identical depths.

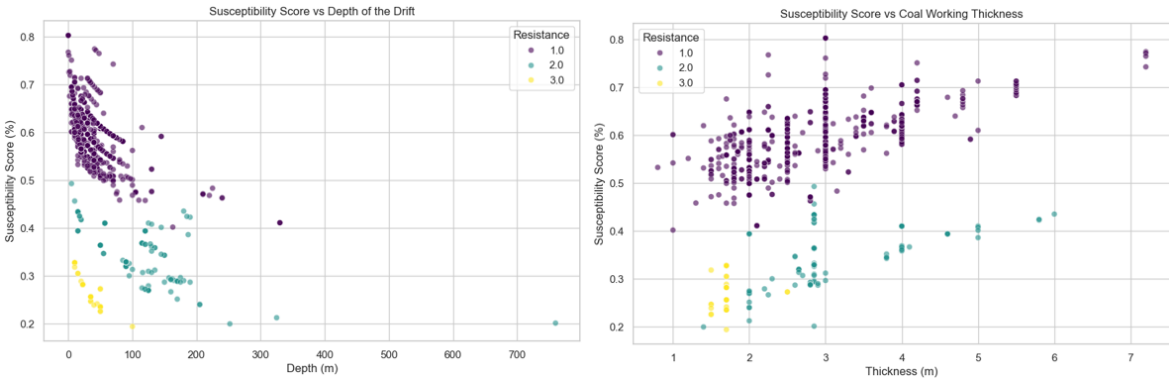


Figure 55. Susceptibility model results compared to conditioning variables

Conversely, the relationship between the susceptibility score and the coal working thickness demonstrates a clear positive linear trend. As the thickness of the extracted seam increases, the corresponding susceptibility scores rise proportionally, reflecting the increased potential energy associated with large underground voids. Within this upward linear trend, the stratification remains highly visible; the point cloud is distinctly segmented by the overburden resistance classes across all thickness intervals. This visual evidence confirms that while the physical geometry of the void establishes the baseline probabilistic trajectory, the structural competence of the rock mass (overburden resistance) acts as a decisive mathematical modifier, successfully validating the multi-variable integration framework of the predictive engine.

13. The Magnitude Model

While the susceptibility model quantifies the probability of a sinkhole occurrence, a magnitude module was developed to infer the severity of that prospective sinkhole. This module evaluates the expected magnitude by employing an approach based on conditional probabilities derived from historical occurrences. By analysing the intersection of specific operational and structural parameters, the model calculates the likelihood of an event manifesting as a Low, Medium, or High magnitude collapse. Consequently, this predictive framework allows the system to establish a data-driven worst-case scenario, providing stakeholders with critical context regarding the potential physical impact of the failure.

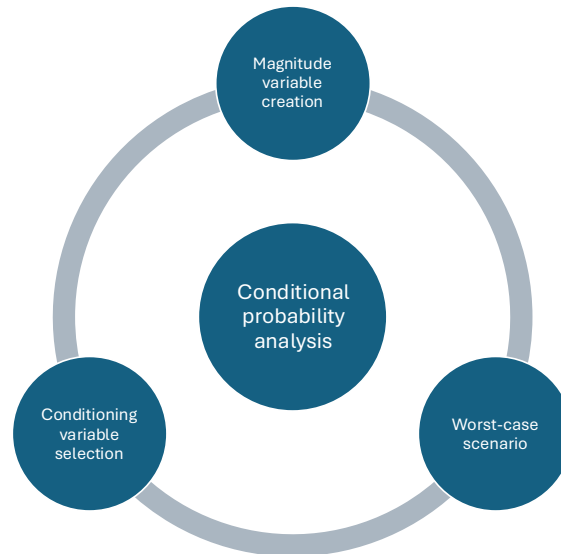


Figure 56. Workflow for magnitude model creation

13.1 Sinkhole characteristics variables

To construct the discrete target variable for sinkhole magnitude (categorized into Low, Medium, and High severity), the model relies on the sinkhole characteristics previously detailed in Part 1, Section 6. As established during the exploratory data analysis, variables such as impacted surface area and sinkhole width suffer from severe incompleteness (less than 50%) and were consequently discarded from this predictive phase. Because these parameters directly represent the high-impact physical scale of the event, employing mathematical data imputation to fill their missing values was strictly avoided to prevent the introduction of artificial bias into the worst-case scenario assessment.

Consequently, the severity derives from the analysis on four key parameters: the number of sinkholes, the length or diameter of the sinkholes, the depth of the sinkhole, and the origin of the collapse.

The subsequent sections detail how these specific parameters are synthesized and mapped to operational thresholds to classify the expected physical severity of the collapse, building directly upon the statistical distributions already defined in the baseline database analysis.

13.2 Depth of the sinkhole

The depth of the sinkhole represents one of the two core geometric parameters directly utilized to calculate the expected physical severity of the collapse event. As comprehensively described in Part 1, Sections 2.2 and 6.1, the raw data array for this variable underwent extensive parsing to remove symbols, standardize measurement units, and filter out non-numeric textual noise.

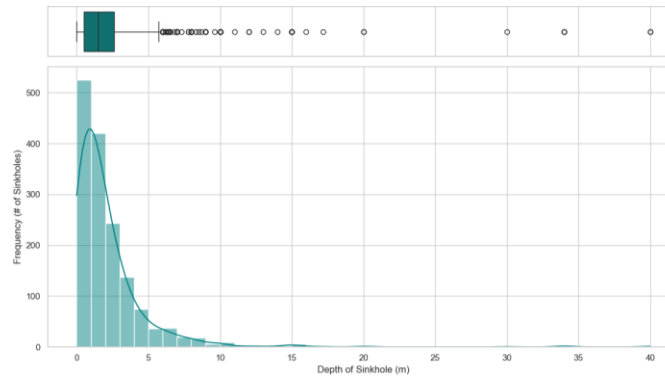


Figure 57. Distribution and outliers of the depth of the sinkhole variable

This foundational harmonisation successfully recovered 1,547 valid continuous entries, preserving 82.8% of the total dataset. As established during the baseline statistical evaluation in Part 1, the dataset yields an average sinkhole depth of 2.13 meters and exhibits a strongly logarithmic distribution. Observations are densely concentrated between 0 and 10 meters, though the data features an extended right tail containing extreme outlier events that reach maximum depths of 40 meters.

Table 18. Descriptive statistics of the depth of the sinkhole variable

Statistics	Value
count	1547
mean	2.13
std	3.08
min	0
25%	0.5
50%	1.5
75%	2.6
max	40

13.3 Diameter of the sinkhole

The diameter (or length) of the sinkhole serves as the primary geometric indicator of the physical extent of the surface collapse. The structural harmonisation, measurement validation, and statistical distribution of this variable are comprehensively detailed in Part 1, Sections 2.2 and 6.1.

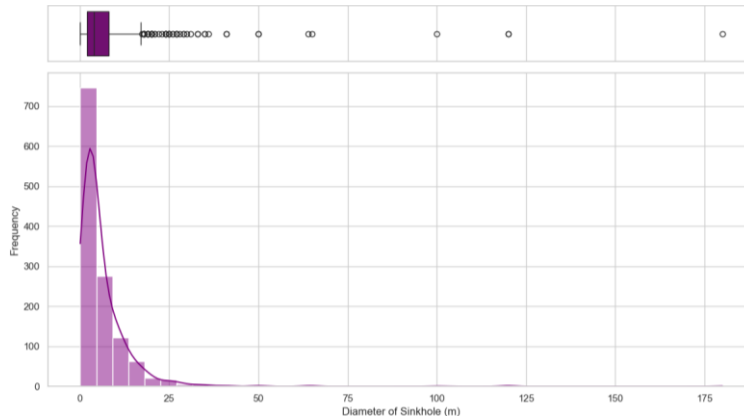


Figure 58. Distribution and outliers of the diameter of sinkhole variable

As established during the baseline analysis, the dataset retains 1,263 valid entries (a 67.6% completeness rate) and is predominantly characterized by small to medium-sized features, with most recorded diameters falling below 5 meters. However, the data also captures highly infrequent, extreme outlier events reaching maximum horizontal dimensions of up to 180 meters.

Table 19. Descriptive statistics of the diameter of the sinkhole variable

Statistics	Value
count	1263
mean	6.29
std	9.57
min	0.1
25%	2
50%	4
75%	8
max	180

By relying on the structural cleaning and spatial distribution already validated in the preliminary analysis, the operational magnitude model directly utilizes continuous measurements such as depth and diameter of the sinkhole, mapping them to discrete severity thresholds to classify the physical scale of the hazard without requiring redundant exploratory processing.

13.4 Origin of Collapse

The structural failure mechanism that initiates sinkhole formation could serve as a geometric indicator for assessing the physical severity of the resulting event. The foundational data standardisation process and the baseline statistical distribution of this categorical variable, specifically highlighting the overwhelming predominance of roof room and roof gallery failures, are comprehensively detailed in Part 1, Section 6.2.

The relationship between the origin of the collapse and the physical magnitude of the resulting sinkhole, specifically defined by its median values of diameter and depth of the sinkhole, was analysed. The results indicate that the underlying failure mechanism directly influences the spatial dimensions of the sinkhole event. The most frequent origin, "failure of the roof room or roof gallery" (875 recorded cases), is associated with the largest typical diameters (5 m) and depths of 2 m. Interestingly, while "failure of the shaft head" is much less common (15 cases), it produces some of the most severe dimensions, presenting diameters of 4.5 m and the greatest comparative depths of 3 m. Other recurring mechanisms, such as "failure of the roof gallery" (292 cases) and "collapse of the shaft filling material" (117 cases), tend to generate moderate-sized sinkholes, with median

diameters of 3 m and depths of 1.5 m. Conversely, "pillar failure" is a rare event (6 cases) and typically results in the smallest diameters (1.75 m) alongside depths of 2 m.

Table 20. Median value of diameter and depth of the sinkhole related to the origin of collapse

Origin of collapse	Diameter of sinkhole	Depth of sinkhole	Count
Failure of the roof room or roof gallery	5	2	875
Failure of the shaft head	4.5	3	15
Failure of the roof room	3	2	43
Failure of the roof gallery	3	1.5	292
Collapse of the shaft filling material	3	1.5	117
Pillar failure	1.75	2	6

13.5 Number of sinkholes

The number of sinkholes variable quantifies the clustering of events, indicating how many distinct surface collapses occur within a specific localized swarm. Consistent with the treatment of the other parameters, this variable was subjected to a preliminary data cleaning process to ensure analytical reliability (Section 2.2). Subsequently, an analysis was conducted to determine how the size of the sinkhole swarm correlates with the individual diameter and depth of the sinkholes.

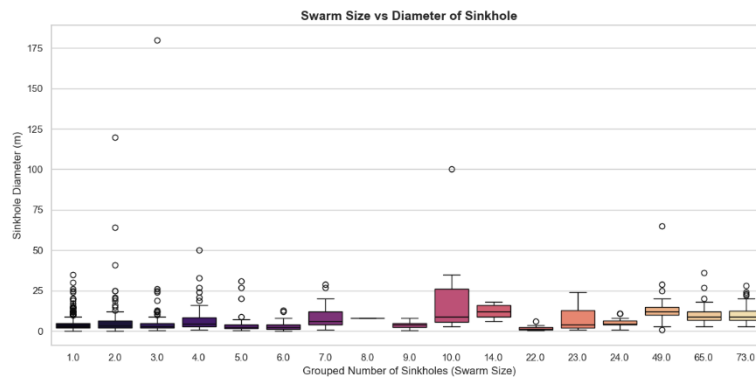


Figure 59. Swarm size compared to their diameter of sinkholes

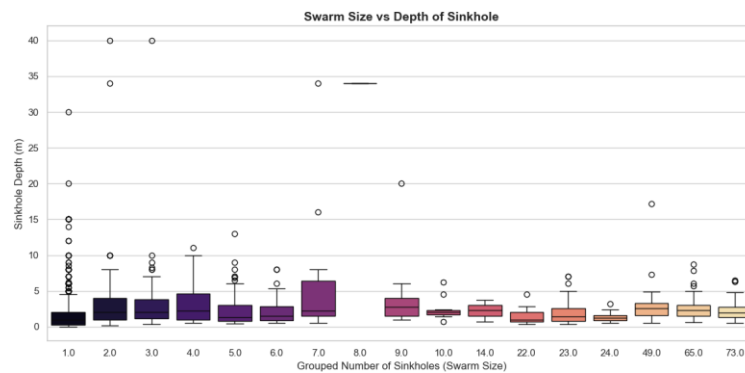


Figure 60. Swarm size compared to their depth of sinkholes

The results indicate that isolated events or very small swarms (e.g., 1 to 4 sinkholes) are associated with the most severe geometric outliers, reaching maximum depths of up to 40 m and exceptional diameters of up to 180 m. Conversely, as the swarm size significantly increases, the individual sinkholes comprising the cluster tend to be much more uniform and comparatively smaller in both depth and diameter, lacking the massive, extreme dimensions.

Table 21. Comparison of swarm of sinkholes with the origin of collapse

Origin of collapse	Number of sinkholes		
	Median	Max	Count
Failure of the roof room or roof gallery	5	73	874
Collapse of the shaft filling material	1	23	116
Failure of the roof gallery	1	23	288
Failure of the roof room	1	5	43
Pillar failure	3	3	6

A comparative analysis was conducted to evaluate the relationship between the origin of the collapse and the resulting size of the sinkhole swarms. The results demonstrate that the underlying failure mechanism significantly dictates the clustering potential of the collapse. The most frequent cause, "failure of the roof room or roof gallery" (874 recorded cases), generates the most extensive swarms, presenting a typical swarm size of 5 sinkholes and extreme occurrences reaching up to 73 individual collapses. In contrast, mechanisms such as the "collapse of the shaft filling material" (116 cases) and "failure of the roof gallery" (288 cases) predominantly manifest as isolated events with a median of 1, although they can occasionally form moderate clusters reaching up to 23 sinkholes. Similarly, a "failure of the roof room" (43 cases) typically produces single occurrences (median of 1) with a maximum observed cluster of 5. Finally, "pillar failure," representing a less common event (6 cases), consistently results in small, uniform swarms exhibiting both a median and maximum size of 3.

13.6 Magnitude variable derivation

To construct the categorical target variable representing the magnitude of the sinkhole, two parameters with a direct physical impact on the severity of the event were selected: the diameter and the depth of the sinkhole. The classification thresholds for these dimensions were initially established using practical expert criteria and common sense; however, within the system's architecture, these values work as dynamic control parameters that can be adjusted and updated in the future without disrupting the pipeline. Table 22 shows the baseline thresholds for the model, translating physical measurements into actionable magnitude levels.

Table 22. Defined thresholds for the Magnitude of sinkhole variable

Magnitude	Diameter [m]	Depth [m]
Low	<=3	<=1
Medium	<=8	<=3
High	>8	>3

Finally, the worst-case scenario was selected to determine the categorical magnitude level of the sinkhole for each observation. These quantitative boundaries define the baseline criteria for the model, translating physical measurements into severity levels.

As detailed previously, while other factors such as the number of sinkholes and the origin of the collapse variables were thoroughly analysed and initially considered for this categorization, multiple

preliminary tests concluded that relying exclusively on the most direct metrics yields a simpler, more robust, and highly interpretable model.

13.7 Selection of conditioning variables

Conditioning variables represent the geological and operational parameters that physically influence and govern the resulting severity of a sinkhole event, acting as the independent predictors for the sinkhole magnitude. To identify which factors, exert the greatest control over the collapse dimensions, a statistical evaluation was performed across the dataset. This analysis aimed to measure the strength of association between the available conditioning features and the discrete magnitude tiers (Low, Medium, and High).

A multivariate evaluation was performed to determine the predictive strength of five geological and mining-related factors in relation to sinkhole magnitude. Four variables are derived from the mining characteristics group, and one from the overburden characteristics group. It begins by selecting only those observations for which the actual worst-case magnitude is available, thereby ensuring that the analysis is based on valid and complete target data. This filtering step is essential because any record without a defined outcome would weaken the reliability of the statistical assessment and distort the interpretation of the predictors.

A variable discretization procedure was then applied to transform the original predictors into operational categorical classes consistent with engineering interpretation. Mining type was recoded into *shaft* and *drift* categories based on the cleaned textual field. Dip was grouped into three classes (*flat*, *gentle*, and *inclined*) using fixed angular thresholds. Working depth and working thickness were similarly discretized into ordered classes reflecting increasing structural or operational intensity. Overburden resistance was treated as a categorical factor. This harmonization step enabled the construction of contingency tables and ensured methodological consistency across all predictors.

Table 23. Thresholds used to convert predictor variables into categorical classes

Predictive variable	Thresholds / rule applied	Resulting categories
Mining type (Drift/Shaft)	Categorical	1_Shaft, 2_Drift
Dip of the layer	0–5°, >5–15°, >15°	1_Flat (0-5°), 2_Gentle (5-15°), 3_Inclined (>15°)
Depth of the Drift	0–30 m, 30–60 m, >60 m	1_Shallow (0-30m), 3_Deep (30-60m), 4_Very_Deep (>60m)
Working Thickness	0–2 m, 2–3.25 m, >3.25 m	1_Thin (0-2m), 2_Medium (2-3.25m), 3_Thick (>3.25m)
Material Overburden Resistance	Categorical	1_Weakly, 2_Mildly, 3_Resistant

The statistical evaluation was performed independently for each predictor through cross-tabulation against the observed sinkhole magnitude. For every predictor–response pair, a chi-square test of independence was computed to assess whether the distribution of magnitude classes varied significantly across predictor categories. In parallel, the strength of association was quantified using bias-corrected Cramér’s V and p-value (Bergsma, 2013). The use of the corrected formulation is relevant in this context because it reduces the effect that may arise in contingency tables with limited sample size or asymmetric dimensions, thereby providing a more robust estimate of explanatory strength.

Finally, the results are represented in a ranking of predictor variables ordered by their Cramér's V values, from the strongest to the weakest association. This ranking allows the analyst to identify which of the five factors has the greatest predictive influence on sinkhole magnitude.

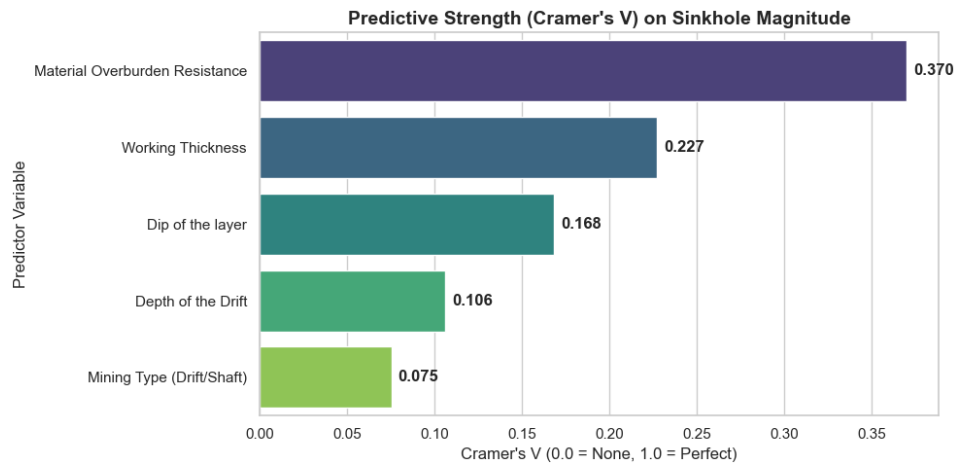


Figure 61. Cramer's V analysis for predictive variables

Two conditioning variables were selected for the magnitude model: Working Thickness and Dip of the layer. These variables were retained because they showed relevant predictive value and offered a suitable basis for structuring the conditional probability model to be applied in the magnitude assessment. Material Overburden Resistance was not included, as this variable resulted from imputations performed in previous stages of the analysis; its use as a predictor could therefore introduce additional bias into the model and affect the robustness of the results. The selection was intentionally limited to two predictors to simplify the conditional probability framework and ensure a more transparent and stable implementation of the magnitude model.

13.8 Conditional probability model

To predict the sinkhole magnitude and its probability of occurrence based on historical data, a conditional probability model was developed. The process began by transforming the two primary conditioning variables, working thickness and coal seam dip angle from continuous numerical values into discrete categorical classes (Table 23). This discretization step was crucial for grouping the historical records into defined structural scenarios, allowing the system to calculate empirical probabilities based on actual observed cases rather than relying on theoretical interpolations.

The selection of the specific thresholds used for this discretization was performed heuristically. The primary criterion was to carefully observe how the working thickness and dip of the layer were distributed in relation to the predefined low, medium, and high magnitude levels. The cutoff boundaries were strategically defined to capture and box the highest possible number of combinations of thickness and dip relative to the historical cases. For example, the working thickness was categorized into thin (<2.0 m), medium (2.0-3.25 m), and thick (>3.25 m) intervals, while the dip angle was classified as flat (0-5°), gentle (5-15°), and steep (>15°), ensuring that the bins effectively represented distinct behaviours.

Table 24. Conditional probability matrix for coal working thickness and coal seam dip angle relative to magnitude classes

Working Thickness	Dip of coal layer	Magnitude Level			Historical Cases
		Low (%)	Medium (%)	High (%)	
Thin (<2.0m)	Shallow dip (0-5°)	24.4	48.9	26.7	45
	Moderate dip (5-15°)	17.1	46.4	36.4	140
	Steep dip (>15°)	14.9	59.6	25.5	47
Medium (2.0-3.25m)	Shallow dip (0-5°)	47.6	31.0	21.4	42
	Moderate dip (5-15°)	24.1	51.9	24.1	133
	Steep dip (>15°)	33.3	16.7	50.0	6
Thick (>3.25m)	Shallow dip (0-5°)	30.8	61.5	7.7	13
	Moderate dip (5-15°)	9.5	36.4	54.0	420

Finally, a conditional probability matrix was constructed by calculating the frequency of each magnitude level within every unique combination of these newly created thickness and dip categories. This resultant table translates raw historical counts into standardized predictive percentages. It is this final, three-dimensional matrix that is directly integrated into the operational model to predict the expected magnitude level. Consequently, when the system evaluates a new scenario, it uses this matrix to return the magnitude level that exhibits the highest probability of occurrence. By isolating this maximum percentage, the model identifies the statistically dominant outcome for those exact conditions. Furthermore, alongside this predicted magnitude and its associated confidence percentage, the system simultaneously retrieves and reports the total number of historical cases that constitute those specific conditions. This ensures that the final prediction is not only the predicted severity but also a probability of confidence and the exact number of historical cases that validate that specific structural combination.

14. Hazard Integration Framework

The final stage of the predictive operational tool involves integrating the two analytical modules, Susceptibility and Magnitude models, to construct a comprehensive Hazard Matrix (Figure 42; Degas & Al Heib, 2024; Nalmpant-Sarikaki et al., 2025). This matrix works as an assessment tool, seamlessly translating the independent evaluations from each model into a unified, actionable hazard classification. To determine the final hazard level, the system maps the outputs derived from the specific conditioning variables processed by each respective model:

- **Susceptibility Level (probability of occurrence):** The first axis of the matrix evaluates the likelihood that a sinkhole event will occur in each area. This susceptibility is calculated by the weighted sum approach (section 12) based on the structural and geological parameters of mine, specifically relying on the depth of the gallery, the working thickness, and the level of resistance of the overburden material.
- **Magnitude Level (expected severity):** The second axis of the matrix defines the potential physical scale and destructive capacity of the expected collapse. This is determined by the conditional magnitude model, which categorizes magnitude levels based on the working thickness and the coal seam dip angle.

By cross-referencing these two distinct dimensions, the Hazard Matrix (Figure 42) provides a holistic evaluation of the terrain. It ensures that the definitive hazard rating accounts simultaneously for both the probability of an event happening (Susceptibility) and the physical impact it would have if it were to materialize (Severity). Furthermore, to ensure the predictive model functions as a truly operational tool, it is essential that each resulting hazard level should be paired with a specific mitigation plan or actionable strategy. This matrix classifies the final terrain hazard into five distinct operational categories: Very Low, Low, Moderate, High, and Critical. Consequently, end-users can seamlessly associate these predefined hazard classifications with tailored engineering interventions, zoning regulations, or monitoring protocols, effectively bridging the gap between theoretical hazard prediction and practical on-site safety management.

The methodological framework of the operational tool presents several inherent limitations must be acknowledged regarding the underlying dataset. The input data utilized to train and validate the models is exclusively derived from historical coal mining regions in France, Poland, and Germany. Consequently, the completeness of the sinkhole catalogue across these specific countries remains a significant challenge, as historical mining records are often fragmented, lost over time, or inconsistently documented. This inherent incompleteness, coupled with the profound spatial and geological variability between the different mining basins, introduces a degree of data imbalance within the dataset. Because the predictive engine relies on historical frequencies and observed patterns, this uneven distribution of recorded incidents may constrain the model's ability to generalize its predictions across entirely unmapped territories or significantly different geological settings.

Furthermore, specific methodological limitations exist concerning the assumptions made to define the mechanical resistance of the overburden material. While the statistical analysis confirmed that the overburden resistance is a highly critical conditioning variable for determining sinkhole susceptibility, its current categorization relies on simplified processing of available geological descriptions. The precision of this specific parameter could be significantly enhanced by integrating advanced geomechanical models or more detailed field tests to robustly support the resistance values that feed into this system. Therefore, this current tool represents a functional prototype; it establishes a fully defined, scalable system architecture and operational workflow, but its predictive behaviour and overall reliability are designed to be continuously improved by incorporating higher-resolution data, more exhaustive historical catalogues, and rigorous geomechanical assessments.

15. Operational Deployment

To bring the predictive framework from theoretical development to practical application, an automated system was orchestrated to deploy the operational hazard tool via an interactive web application. This tool was orchestrated through several technologies: Python served as the core programming language to build the architecture, coordinating how individual files and scripts communicate. Specifically, the orchestration script handles the packaging of data and models into standardized *pickle* (.pkl) files, which are then consumed by the inference engine (Engine) to process live user requests. This backend logic communicates directly with the user-facing Streamlit application, ensuring that physical and geological inputs are instantly evaluated through both the susceptibility and magnitude modules to render the hazard assessment.

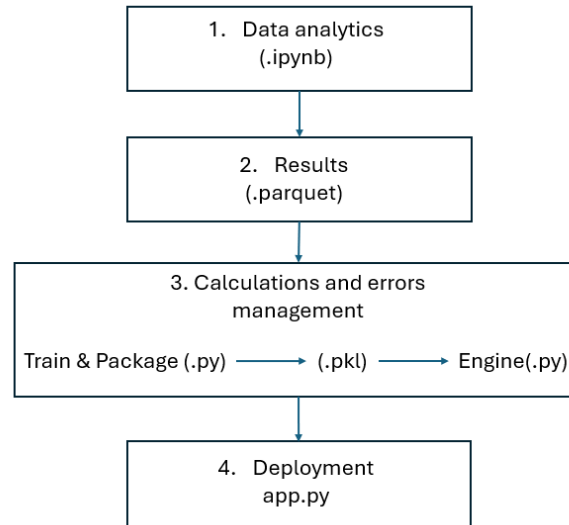


Figure 62. General workflow for the hazard operational tool deployment

(1) The workflow initiates within a dedicated Jupyter notebook (.ipynb), which serves as the foundational environment for comprehensive data treatment and experimentation. Within this exploration, extensive data analysis was performed, encompassing rigorous data cleaning, the removal of formatting anomalies, and the standardization of both column headers and textual contents across the database. Furthermore, this notebook environment facilitated the statistical correlation analysis such as Cramer's V evaluations and the initial training, tuning, and validation of the Machine Learning pipeline components prior to their export and integration into the production-ready operational scripts.

(2) Then, the final processed datasets and structural results were exported into (.parquet) files. This specific file format was chosen for the production environment due to its highly optimized, columnar storage structure, which makes it significantly lighter and more memory-efficient than traditional data formats like CSV. By utilizing *parquet* files, the system ensures that the underlying data is much easier and faster for the web application to ingest, parse, and process. This data handling optimizes the operational tool's overall performance, minimizing latency and ensuring smooth, responsive user experience during real-time hazard assessments.

(3) To ensure the continuous relevance and accuracy of the operational tool, a circular updating subprocess was established to seamlessly retrain the models, execute necessary calculations, and implement error handling. Once the data is processed, the system automatically packages the updated model parameters and their corresponding results into standardized *pickle* files. This serialization process is highly comprehensive; every resulting parameter, evaluated threshold for each variable, and even the specific Machine Learning model trained to impute the missing overburden material resistance values are securely stored within a *pickle* file, making them readily accessible for the live inference engine.

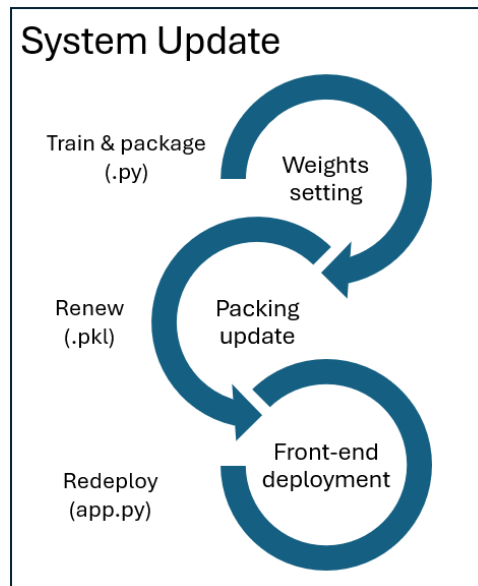


Figure 63. Workflow for parameter settings update

This automated pipeline operates through two distinct but interconnected procedures: the systematic packing of the model parameters and a dedicated module for operational execution and error management. The packing procedure acts as the system's core memory, saving all the critical components required to rebuild the predictive models during runtime. For the susceptibility model, this involves storing the calculated feature weights, the internal Machine Learning parameters, the min/max normalization boundaries, and the specific Jenks natural breaks thresholds used to classify the level of susceptibility. Simultaneously, this packing procedure records all the rules defining the magnitude model. It stores the numerical thresholds for the physical variables that dictate the severity of the event as well as the specific categorization boundaries established for the magnitude conditioning variables. This operational script features built-in error handling protocols designed to validate user inputs, ensuring that out-of-bound values or unexpected data entries are managed without crashing the web application, thereby guaranteeing stable and reliable user experience.

(4) Finally, leveraging the *pickle* file containing all the validated model parameters alongside the dedicated operational module for calculations and error management, the interactive web application is deployed using Streamlit. This final integration step bridges the backend architecture with a user-friendly frontend interface. By utilizing Streamlit, the system is capable of securely receiving live inputs, validating them through the established error-handling protocols, and instantly querying the pre-packaged predictive models to dynamically render the final hazard matrix assessments in real-time.

To foster open collaboration and continuous improvement, the complete source code for this project has been made publicly available in a GitHub repository at:

<https://github.com/gerobena/sinkholes-prediction-model>

This repository can be readily cloned, allowing researchers and developers to directly address the limitations, refine the methodology, and further enhance the predictive capabilities of the tool. Alternatively, for users seeking a more straightforward and immediate application without the need to interact with the underlying codebase, the fully operational model has been successfully deployed and is ready for practical use via the live web application at:

<https://sinkholes-prediction-model-sirima.streamlit.app/>

16. Conclusions and perspectives

The research document presents a methodology for assessing and predicting sinkhole hazards in European post-mining areas, based on a database of 1,868 historical records from shallow abandoned coal mines in France, Germany, and Poland. This deliverable was prepared within Task 4.1 of the SIRIMA project. It uses the common database built during WP2, which compiles sinkhole records related to shallow abandoned coal mine workings and shafts. The database is incomplete and heterogeneous, but it remains a rare European inventory of sinkholes related to shallow abandoned coal mine workings.

Because sinkholes related to galleries and those associated with shafts correspond to different mechanisms, shaft-related cases were not included in the analyses presented in this report. Nevertheless, the SIRIMA inventory does not describe a single homogeneous phenomenon. Underground-working sinkholes are mainly associated with the upward propagation of collapses from rooms (for room and pillar mining method) or galleries (access, infrastructures, etc.), whereas shaft-related sinkholes are controlled by the presence, condition and geometry of the shaft itself. This distinction was applied in the hazard model, introducing the mining features as explicit input variables.

A descriptive analysis of the SIRIMA sinkhole database was provided and used for the hazard modelling and prediction objective. The database confirms that most of documented sinkholes have a diameter smaller than 5 m (Salmon et al., 2015). This result is important for hazard assessment, since collapse diameter is commonly used to qualify sinkhole intensity, as presented in the intensity framework proposed by Franck et al., (2019). However, a notable result of the analysis is the relatively large number of sinkholes associated with workings deeper than 50 m, whereas several authors have shown that sinkholes generally develop above shallow mine workings (Canbulat et al., 2017; Salmon et al., 2015). These cases should therefore be looked at more closely, either to better understand their specific context or to check uncertainties in the source data.

The statistical analysis was deliberately restricted to variables with sufficient documentation, in line with the selection made in Section 3.3. Variables with very low completion rates were not retained as explanatory factors. This choice avoids overinterpreting very small or unrepresentative subsets. Among the retained variables, working thickness, working depth, dip of the mined layer and the reported presence of an aquifer show the clearest links with impacted surface area variable. Links with sinkhole depth are weaker. Overburden geology and overburden resistance show some relationship with depth of sinkhole, but the available data do not support a simple mechanical interpretation. These results are useful for identifying candidate variables for modelling, not for proving causal mechanisms.

To transition from reactive observation to risk management, the project developed a statistical predictive model to capture the susceptibility and the magnitude of the predictive tool to perform a hazard assessment. The Susceptibility model measures search to measure the level of a sinkhole occurrence using a weighted sum algorithm governed by the depth of the drift, the working thickness, and the overburden material resistance. To overcome critical gaps in the historical dataset, this module successfully integrates Machine Learning and statistical techniques. In parallel to predicting the likelihood of an event, a conditional probability-based Magnitude Model was constructed to forecast the potential severity of a collapse. By correlating the dimensions of historical geometric extremes with the extracted working thickness and the coal seam dip angle, this module categorizes the expected physical impact into discrete Low, Medium, or High severity levels based on a worst-case scenario approach.

The outputs from both the Susceptibility and Magnitude models were combined into a unified Hazard Matrix, which cross-references the level of occurrence with the expected severity to classify the terrain into distinct operational risk categories. This comprehensive analytical framework has been successfully deployed as an interactive, Python-based web application. The automated tool provides land-management authorities with a practical decision-support system designed to guide measure of actions such as engineering interventions or establish zoning regulations and facilitate proactive safety management.

17. References

- Ahmed, M.-S., N'diaye, M., Attouch, M. K., & Dabo-Niang, S. (2023). K-nearest neighbors prediction and classification for spatial data. *Journal of Spatial Econometrics*, 4(1), Article 12. <https://doi.org/10.1007/s43071-023-00041-2>
- Al-Obaydy, W. N. I., Hashim, H. A., Najm, Y. A., & Jalal, A. A. (2022). Document classification using term frequency-inverse document frequency and K-means clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, 27(3), 1517–1524. <https://doi.org/10.11591/ijeecs.v27.i3.pp1517-1524>
- Al Heib, M. M., Franck, C., Djizanne, H., & Degas, M. (2023). Post-mining multi-hazard assessment for sustainable development. *Sustainability*, 15(10), Article 8139. <https://doi.org/10.3390/su15108139>
- Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*, 42(3), 323–328. <https://doi.org/10.1016/j.jkss.2012.10.002>
- Canbulat, I., Zhang, C., Black, K., Johnston, J., McDonald, S., Kay, D., Tong, S., 2017. Assessment of Sinkhole Risk in Shallow Coal Mining. Presented at the Mine Subsidence: Adaptive Innovation for Managing Challenges. <https://doi.org/10.26190/unsworks/26513>
- Clarke, B. G., Welford, M., & Hughes, D. B. (2006). The threat of abandoned mines on the stability of urban areas (Paper No. 379). In *Proceedings of the 10th IAEG International Congress*. The Geological Society of London.
- Degas, M., & Al Heib, M. (2024). Methodology of interaction between post-mining hazards (Deliverable D2.3, PoMHaz Project, Grant No. 101057326). Ineris / Research Fund for Coal and Steel (RFCS).
- Franck, C., Salmon, R., Didier, C., Paquette, Y., Pokryszka, Z., 2019. Post-Mining Hazard Evaluation and Mapping in France | Ineris (Ineris).
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall.
- Joshi, B. R., Bhandary, N. P., Acharya, I. P., & K.C., N. (2026). Comparative assessment of quantitative landslide susceptibility mapping using feature selection techniques. *ISPRS International Journal of Geo-Information*, 15(1), Article 20. <https://doi.org/10.3390/ijgi15010020>
- Karagüzel, R., Mahmutoğlu, Y., Erdoğan Topçuoğlu, M., Şans, G., & Dikbaş, A. (2020). Susceptibility mapping for sinkhole occurrence by GIS and SSI methods: A case study in Afsin-Elbistan coal basin. *Pamukkale University Journal of Engineering Sciences*, 26(8), 1353–1359.
- Ke, C., He, S., & Qin, Y. (2023). Comparison of natural breaks method and frequency ratio dividing attribute intervals for landslide susceptibility mapping. *Bulletin of Engineering Geology and the Environment*, 82(10), Article 384. <https://doi.org/10.1007/s10064-023-03392-0>
- Kotyrbá (2005). Zagrożenie i ryzyko zapadliskowe terenów Górnośląskiego Zagłębia Węglowego. *Wiadomości górnicze* 7-8/2005, 348-358.
- Kotyrbá, A., & Kortas, Ł. (2016). Sinkhole hazard assessment in the area of abandoned mining shaft basing on microgravity survey and modelling — Case study from the Upper Silesia Coal Basin in Poland. *Journal of Applied Geophysics*, 130, 62–70. <https://doi.org/10.1016/j.jappgeo.2016.04.007>
- Marino, G. G. (2020, April). Building above underground coal mines. *STRUCTURE Magazine*. <https://www.structuremag.org/article/building-above-underground-coal-mines/>
- Mazumder, D., Das, R., & Das, S. (2025). Multi-criteria decision analysis framework for landslide susceptibility mapping with analytical hierarchy process in parts of Assam–Arakan Fold Belt, India. *Geological Journal*, 61(2), 307–322. <https://doi.org/10.1002/gj.5229>
- Michael, E. A., & Samanta, S. (2016). Landslide vulnerability mapping (LVM) using weighted linear combination (WLC) model through remote sensing and GIS techniques. *Modeling Earth Systems and Environment*, 2(2), Article 88. <https://doi.org/10.1007/s40808-016-0141-7>
- Misa, R., Sroka, A., & Mrochęń, D. (2025). Evaluating surface stability for sustainable development following cessation of mining exploitation. *Sustainability*, 17(3), Article 878. <https://doi.org/10.3390/su17030878>

- Nalmpant-Sarikaki, D. M., Theocharis, A. I., Koukoulzas, N. C., & Zevgolts, I. E. (2025). A framework for effective multi-hazard risk assessment in post-mining areas. *Safety*, 11(1), Article 18. <https://doi.org/10.3390/safety11010018>
- Orhan, O., Yakar, M., & Ekercin, S. (2020). An application on sinkhole susceptibility mapping by integrating remote sensing and geographic information systems. *Arabian Journal of Geosciences*, 13(17), Article 886. <https://doi.org/10.1007/s12517-020-05841-6>
- Palchik, V. (2015). Bulking factors and extents of caved zones in weathered overburden of shallow abandoned underground workings. *International Journal of Rock Mechanics and Mining Sciences*, 79, 227–240. <https://doi.org/10.1016/j.ijrmms.2015.07.005>
- Salmon, R., Daupley, X., Charmoille, A., Ghoreychi, MehdiMe, 2015. Retour d'expérience sur les effondrements localisés miniers | Ineris (Ineris).
- Tritsch, J.-J., 2005. Les mécanismes de ruptures des carrières souterraines abandonnées, in: *Evaluation et Gestion Des Risques Liés Aux Carrières Souterraines Abandonnées : Séminaire de Restitution et de Valorisation Des Travaux INERIS -Réseau Des LPC*. LCPC. Paris, Paris, France, pp. 74–92.
- Trott, M., Leybourne, M., Hall, L., & Layton-Matthews, D. (2022). Random forest rock type classification with integration of geochemical and photographic data. *Applied Computing and Geosciences*, 15, Article 100090. <https://doi.org/10.1016/j.acags.2022.100090>
- Vachat, J.C., 1982. Les désordres survenant dans les carrières de la région parisienne. Étude théorique et pratique de l'évolution des fontis. CNAM.
- Vuidart, I., Conil, N., Al Heib, M., 2026. Post-mining Sinkhole Hazard and Risk Management in France: Feedback and Outlook. Presented at the IPMS 2026 and ECOP 2026, Turkey.
- Wang, B., Ma, K., Wu, L., Qiu, Q., Xie, Z., & Tao, L. (2022). Visual analytics and information extraction of geological content for text-based mineral exploration reports. *Ore Geology Reviews*, 144, Article 104818. <https://doi.org/10.1016/j.oregeorev.2022.104818>
- WP2 Sirima partners, 2026. Work Package 2 Progress Data bases creation and advanced analysis.
- Yan, W., Guo, J., & Yan, S. (2023). Difference in surface damage between deep and shallow mining of underground coal resources in China. *Sustainability*, 15(9), Article 7296. <https://doi.org/10.3390/su15097296>
- Zhuang, X., Yu, X., Mao, C., Yi, S., & Cao, A. (2025). Study on deformation mechanism and stability analysis of mined-out areas in small coal mines. In *Advances in Engineering Research: Vol. 267. Proceedings of the 2025 International Conference on Engineering Management and Safety Engineering (EMSE 2025)* (pp. 359–366). Atlantis Press. https://doi.org/10.2991/978-94-6463-780-9_33

Appendix – How to Use the Operational Tool

You can access the Operational Tool through:

<https://sinkholes-prediction-model-sirima.streamlit.app/>.

This tool evaluates sinkhole hazard related to coal mining using four variables: working depth, working thickness, dip of the layer and overburden rock resistance.

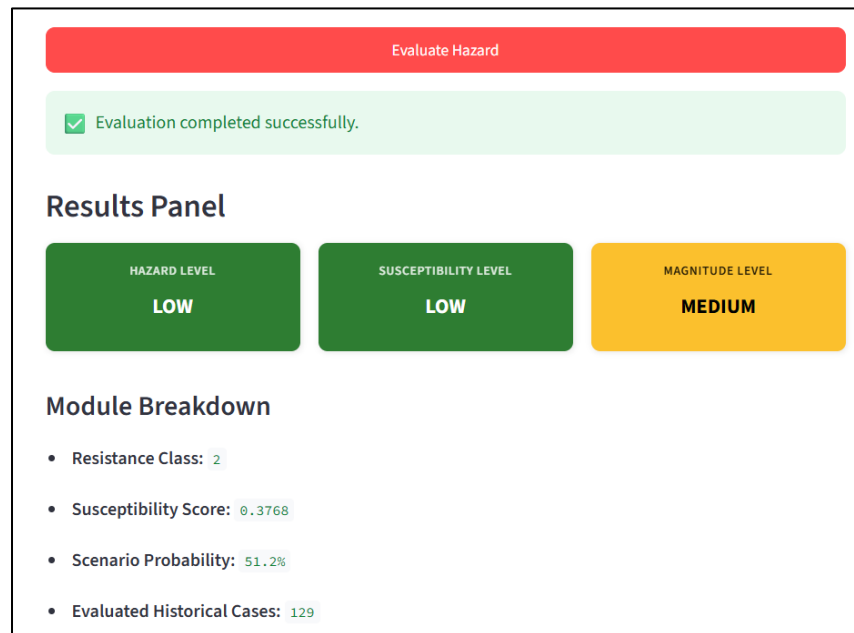
- Depth of the Drift (m): The vertical distance from the surface to the mining cavity. Must be a positive numerical value.
- Coal Working Thickness (m): The thickness of the extracted coal seam. Must be a positive numerical value.
- Dip of coal layer (°): The coal seam dip angle.
- Geological Description: The raw text describing the overburden lithology in English (e.g., "sands and shale"). The built-in NLP model will process this text to automatically assign the correct resistance class.
- Manual Resistance: An alternative to the text description. Leave the text field empty and directly input 1 (Weak), 2 (Medium), or 3 (Strong) based on your own geotechnical assessment.

1. Input parameters

The screenshot displays the SIRIMA System interface for Comprehensive Hazard Assessment. It is divided into two main sections: Physical Parameters and Geological Parameters. Under Physical Parameters, there are three input fields: Drift Depth (m) with a value of 10.00, Dip (°) with a value of 2.00, and Working Thickness (m) with a value of 4.18. Each field has minus and plus buttons for adjustment. Under Geological Parameters, there is a radio button selection for the resistance evaluation method, with 'Geological Description (NLP)' selected. Below this is a text input field for the rock description, which contains the text 'clay and sand'.

Users input positive numerical values in meters for working depth and working thickness. For rock resistance, the system accepts either a manual categorical value (1 to 3) or a raw geological text description. When text is provided, an integrated NLP engine analyzes lithology description and automatically assigns the correct resistance class based on historical patterns. Once all parameters are set, the algorithm applies a normalization process with calibrated expert weights for the susceptibility model and applies the conditional probability matrix for the magnitude model.

2. Output results



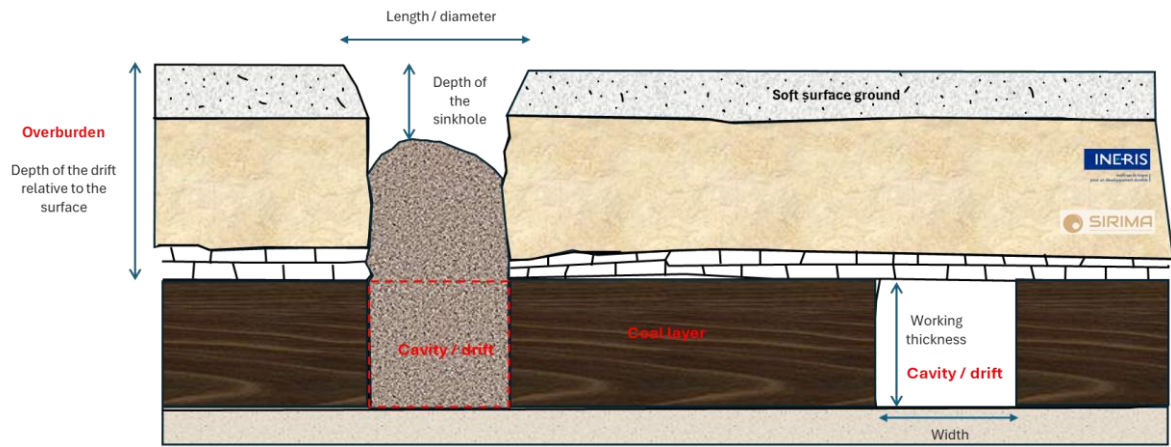
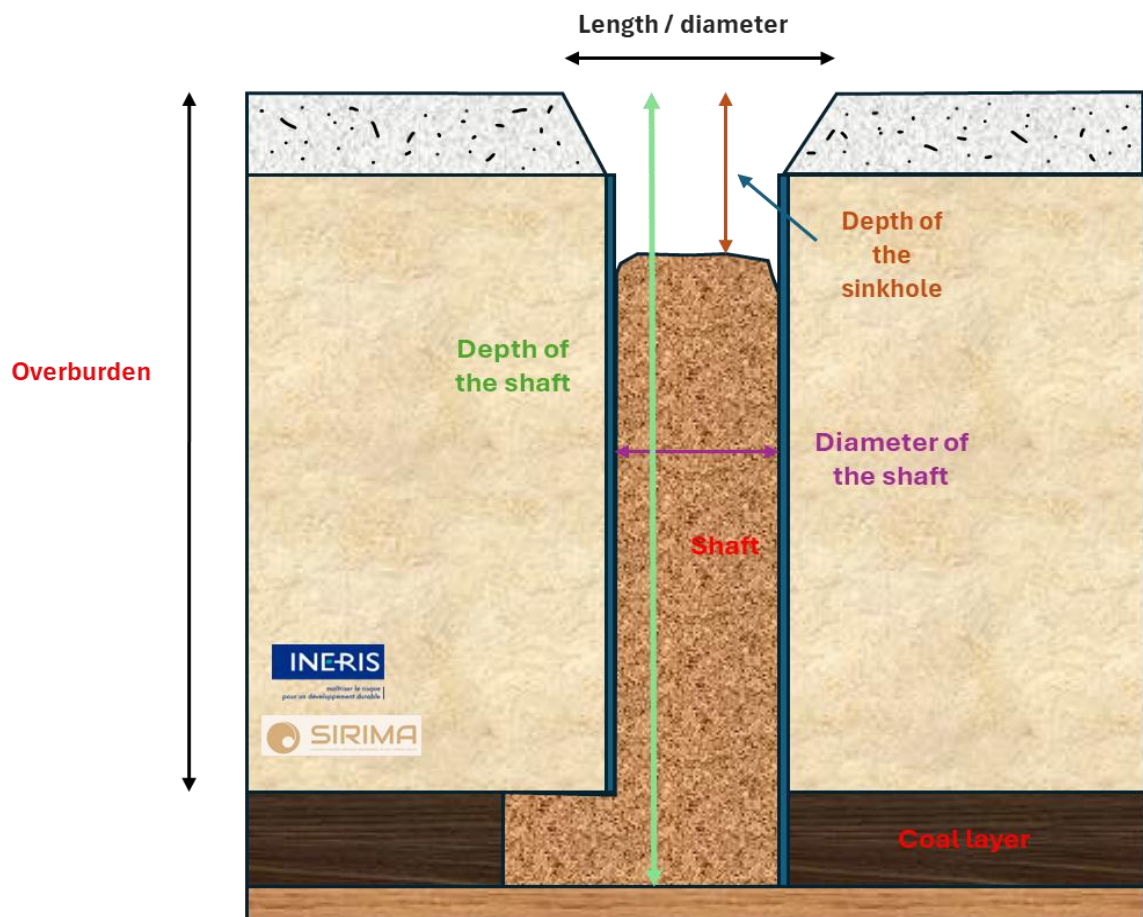
After clicking the “Evaluate Hazard” button, the output dashboard displays the predicted magnitude, susceptibility and integrated hazard levels.

For greater detail, a breakdown panel is displayed:

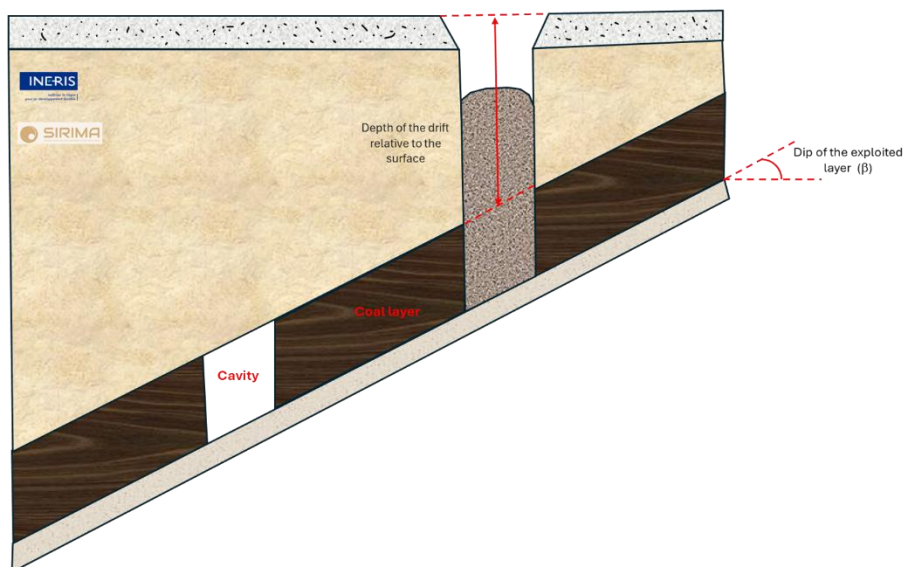
- Resistance Class: Indicates the corresponding overburden material resistance level.
- Susceptibility Score: Represents the calculated susceptibility level.
- Scenario Probability: Shows the percentage of events that match the input parameters for the magnitude model.
- Evaluated Historical Cases: Indicates the number of past events that fit the given parameters for the magnitude model.

Appendix – Description of database variables (Vuidart et al., 2026; WP2 Sirima partners, 2026)

	Sinkhole characteristics
Object	Comments
Number of sinkholes	In some cases, there may be many sinkholes on a small surface area. Each sinkhole will not be described, but the number of sinkholes observed will be indicated. The characteristics retained will be the most representative of the sinkholes in the area.
Surface impacted (m ²)	Area of the zone or surface of the sinkhole
Length (m)	If the sinkhole is circular, enter the diameter only in the length field. If the information is unknown, enter “unknown”
Width (m)	Width of the impacted surface
Depth of the sinkhole (m)	cf. explanatory diagram
In-situ measurement or estimation	Specify whether the dimensions were measured on site or whether they are an estimate (e.g. archive data)
Date of appearance of the sinkhole	Specify the date of appearance of the sinkhole. If it is not known, give an approximate date or period and specify the uncertainty in the following field
Uncertainty of date (years)	
Origin of collapse	Failure of the shaft head Collapse of the shaft filling material Faillure of the roof room Faillure of the roof gallery Pillar failure Unknow
Similar phenomena known nearby	Yes or no



	Characteristics of mining work (shaft/cavity-drift)
Object	Comments
Start date of the mining	Specify the start date of mining operations in the sinkhole sector
End date of the mining	Specify the end date of mining operations in the sinkhole sector
Accuracy of date	
Dip of the exploited layer (°)	cf. explanatory diagram
Type of mining work (shaft/cavity-drift)	shaft/cavity-drift
Width / diameter	Width of the cavity / drift, Diameter of the shaft (cf. explanatory diagram)
Working thickness / Deep of the shaft	In the case of a gallery or a stope, specify the height of the galleries (cf. explanatory diagram)
Void density	Specify whether it is an isolated galley (low density) or room and pilars or slotes or block caving
Depth of the drift relative to the surface (m)	Taken from the roof (cf. explanatory diagram)
Uncertainty (m)	Uncertainty of the depth of the drift relative to the surface
Ground support	Concrate, lining segments, Set, rock bolting, masonry, bricks, none
Accessibility	Is the cavity or shaft accessible for inspections or monitoring ?



	Overburden characteristics*
Object	Comments
Water conditions in the minig work	Specify if there are water inlets in the mine and their nature : dry, seepage, runoff, flooded
Roof fracture condition	No fracturation, fractured, highly fractured
Geology of the overburden	Short description of the nature of rocks (limestone, sandstone, marl,,,))
Overburden resistance	highly resistant, resistant, weakly resistant, low strength powdery
Presence of an aquifer in the hanging wall	yes or no
*	The high of the overburden is given in the characteristics of mining work (Depth of the drift relative to the surface)