SINKHOLE HAZARD AND RISK MANAGEMENT IN POST MINING AREAS

RFCS PROJECT NO 101157400

Co-funded by the European Union

Numerical Model

WP.5. Machine learning modelling

Task 5.3. Modeling the possibility of discontinuous deformation hazards in the form of sinkholes using machine learning techniques and validation process

Responsible Partner: IMG – PAN, Poland

June 2026

1. Introduction

This report presents the achievement of Milestone No 6 (Machine Learning model for sinkholes) within Work Package 5 (WP5) of the SIRIMA project. The main focus of this achievement was the development of a functional forecasting model for sinkhole formation in post-mining areas. This milestone played an important role in establishing the direction of further research. Based on a preliminary analysis of the first sinkhole database provided for the project, this milestone allowed the project team to assess the quality and completeness of available information, identify the most relevant parameters for sinkhole characterization, and evaluate the suitability of selected analytical and machine learning approaches.

2. Acquisition and Initial Assessment of the Sinkhole Database

The first crucial step involved the acquisition and initial organization of the sinkhole database. This step was essential because it provided the first structured dataset describing sinkhole occurrences and their potential controlling factors. The database included a range of parameters related to mining conditions, geological setting, sinkhole geometry, and site characteristics. These parameters created the basis for further analysis and allowed the first attempt to link sinkhole size and occurrence with underground mining and geological conditions.

At the same time, the initial assessment showed that the database was incomplete and heterogeneous. Several parameters contained missing values, while others required additional interpretation or standardization before they could be used in modelling. This was an important finding, as it highlighted that the quality and structure of the database would strongly influence the reliability of any further statistical or machine learning analysis.

The database assessment was therefore a major step forward in the project. It helped identify the strengths and limitations of the available data and showed which parts of the database required further development.

3. Identification of Relevant Parameters

The next step was the preliminary evaluation of parameters included in the database. This analysis helped determine which variables appeared to be more relevant for sinkhole characterization and which variables were less useful at this stage.

The analysis indicated that parameters directly related to mining geometry and geological conditions are particularly important for further modelling. These include, among others, the depth of mining workings, the thickness of the exploited layer or drift and the geological characteristics of the overburden. Such variables are expected to influence the potential development and size of sinkholes, and therefore should be prioritized in the larger database.

In contrast, some parameters were found to be less informative, either because of limited variability, a high number of missing values, or difficulties in interpretation. This does not necessarily mean that these parameters are irrelevant, but rather that they require better documentation or more consistent data collection before they can be used reliably.

This stage was important because it provided a clearer understanding of which information is most valuable for sinkhole modelling. The conclusions from this analysis will support the design and improvement of the larger database and help focus future efforts on the most meaningful variables.

4. Preliminary Modelling and Method Evaluation

The third step was the first application and evaluation of modelling approaches using the initial database. The aim of this step was not to produce a final predictive model, but rather to test which methods are suitable for this type of data and which limitations should be expected.

The preliminary modelling showed that the dataset presents several challenges for machine learning applications. These include a limited number of cases, missing values, heterogeneous site conditions, and possible differences between countries or mining regions. As a result, model performance depended strongly on data preparation and on the ability of the model to handle non-linear relationships and incomplete information.

Despite these limitations, the initial modelling results were encouraging. They showed that machine learning methods are able to capture meaningful relationships between the input parameters and sinkhole characteristics. This confirms that the available data contain useful predictive information and that machine learning can be a valuable tool for further analysis of mining-induced sinkholes. With additional data, improved parameter completeness, and further model refinement, these methods have the potential to provide more reliable predictions of sinkhole geometry and support future risk assessment.

This step was important because it provided practical knowledge about the modelling strategy. It helped define which models should be further developed, which assumptions need to be treated carefully, and what improvements in the dataset are necessary before more advanced modelling can be applied.

5. Guidance for Further Work on the Extended Database

The fourth step is the translation of the preliminary results into guidance for the next stages of the SIRIMA project. The initial database analysis confirmed that the development of a larger and more complete database is a crucial requirement for reliable sinkhole prediction and risk assessment.

The work completed so far provides several important directions for future activities. First, the database should be further standardized, with consistent definitions of parameters and categories. Second, key variables related to mining geometry, overburden conditions, and sinkhole geometry should be prioritized. Third, missing data should be carefully documented and, where possible, addressed using appropriate data preparation or imputation methods. Finally, modelling approaches should be selected not only based on predictive performance, but also on their interpretability and usefulness for understanding sinkhole formation mechanisms. The preliminary analysis therefore represents an important foundation for the next phase of the project.

6. Summary of Main Achievements

In conclusion, the successful completion of this stage marks a significant milestone in the SIRIMA project, centered on the development of a functional, Machine Learning-based predictive model for sinkhole formation. The work carried out has effectively transformed the initial raw dataset into practical, actionable knowledge regarding data requirements, parameter relevance, and future modeling strategies.

Crucially, these efforts led to several key achievements. First, the initial sinkhole database was reviewed, structured, and thoroughly assessed in terms of completeness and usability. Through this process, key parameters influencing sinkhole characterization were identified, and critical limitations related to missing or heterogeneous data were recognized. Furthermore, preliminary modeling tests provided vital insights into which analytical and Machine Learning approaches are promising and which are less suitable for the current dataset. Ultimately, these results have established a solid methodological foundation that will support the next stages of the project, specifically facilitating the implementation of an expanded database and the refinement of more reliable predictive models for mining-induced sinkhole assessment.