

Automatic Classification of Borehole Lithological Descriptions for 3D Geological Modeling: Application to Aix-en-Provence (France)

¹Julie Regnier, ²Mickael Binois, ³Ilias Djermoun

¹Géoazur/CEREMA CNRS, ²INRIA Acumes Team, ³Université Côte d'Azur, Sophia Antipolis, France

Context

Three-dimensional geological models rely on borehole lithological descriptions to represent the subsurface. However, these descriptions are recorded by different geologists using heterogeneous terminology, abbreviations and varying levels of detail. This heterogeneity limits their direct integration into automated geological modeling workflows.

Problem

How can heterogeneous borehole lithological descriptions be automatically transformed into lithological families to support the construction of a 3D subsurface model of Aix-en-Provence ?

The study focuses on boreholes located in the Lower Var Valley (Alpes-Maritimes, France) an area with a dense network of boreholes used for hydrogeological and geological investigations and in Aix-en-Provence, where the objective is to build a 3D geological model of the subsurface.

Objectives

- Compare supervised machine learning algorithms for lithology classification
- Automatically cluster lithological descriptions using BERTopic
- Prepare structured geological information for 3D geological modeling

Preliminary Results

Four supervised learning models were evaluated. Random Forest achieved the highest Top-1 classification accuracy (48.7%), TabPFN (42.9%), XGBoost (44.5%) and Logistic Regression (27.7%). Despite these moderate Top-1 performances, Top-2 and Top-3 accuracies increased substantially for all models (~70%), indicating that the correct lithological family is frequently ranked among the first candidate predictions.

The results highlight the potential of combining supervised classification and topic modeling for the automatic processing of borehole lithological descriptions. While supervised learning provides direct prediction of lithological families, BERTopic reveals the semantic organization of textual descriptions and helps identify consistent geological groups. The large difference between Top-1 and Top-2 accuracies suggests that many prediction errors correspond to ambiguous lithological descriptions rather than completely incorrect predictions.

Future work will focus on improving classification performance by integrating additional spatial and geological features, refining the lithological families with expert interpretation, and incorporating the lithological families obtained by clustering into the construction of a three-dimensional geological model of the Aix-en-Provence subsurface.