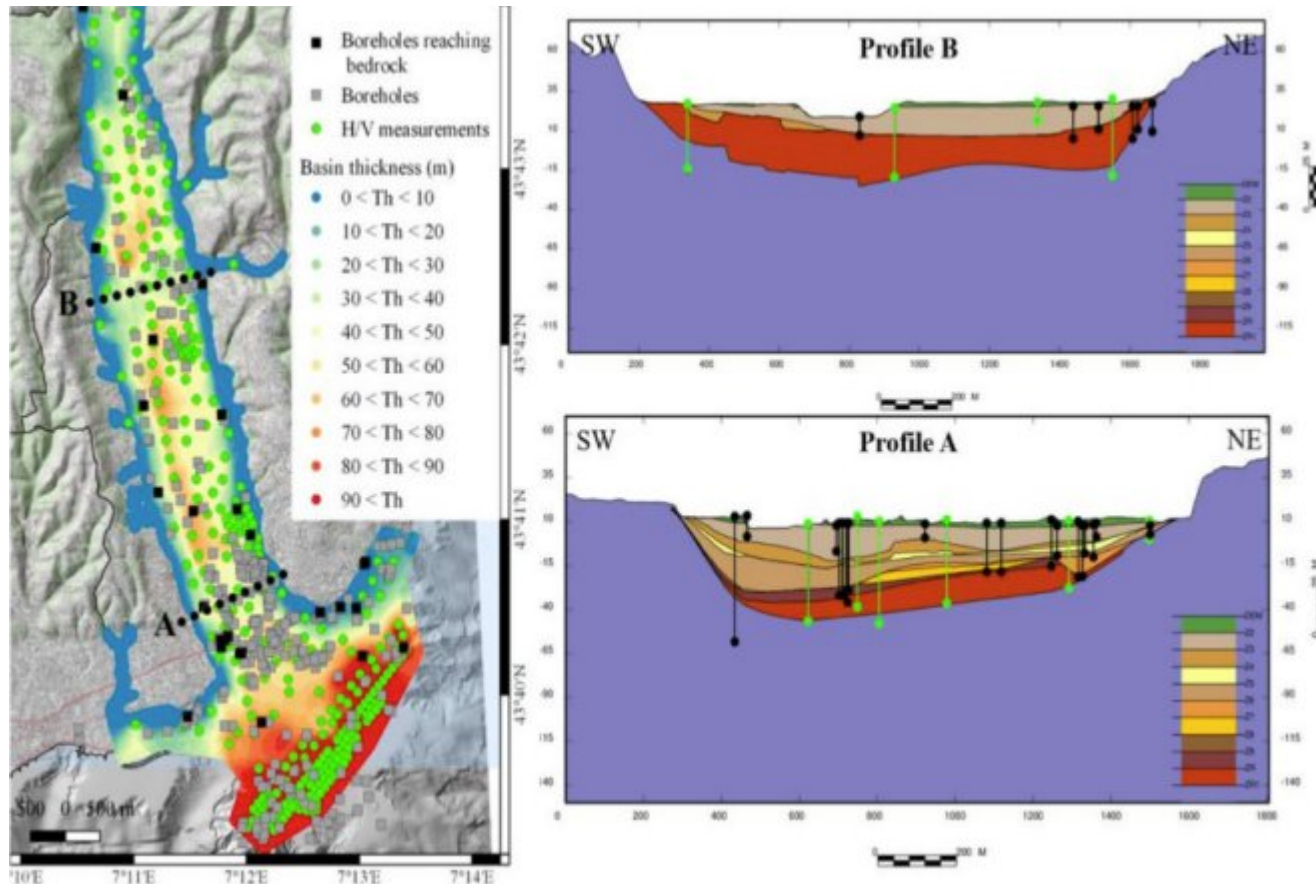


Machine learning Applied to Borehole data for the definition of typical lithostratigraphic soil columns ; towards a 3D soil model of the city of Aix-en-Provence

Mickaël BINOIS (Inria/ACUMES), Julie REGNIER (CEREMA/GEOAZUR), Ilias DJERMOUN
(Polytech Nice Université Côte d'Azur)

Motivation



Layer description	Code	Geology (1/50000 ^e geological map)
Filling, concrete. Topsoil	1	Anthropogenic deposits
Sand + pebbles + gravels	2	Recent alluvium
Fine sand. Silty sand or clay sand.	3	Deltaic deposits
Silt, clay (sometimes muddy)	4	Muddy deposits and pro-delta
Sand + pebbles + gravels	5	Recent alluvium. Fluvial deposits
Fine sand. Silty sand or clay sand	6	Deltaic deposits
Silt, clay (sometimes muddy)	7	Muddy deposits and pro-delta
Coarse sand	8	Ancient alluvium. Fluvial deposits
Sand + pebbles + gravels	9	Ancient alluvium. Fluvial deposits
Bedrock	R	Pliocene, Cretaceous, Jurassic, Trias

Left: Lower Var Valley map with boreholes' locations.

Right: reconstructed geological sections from the corresponding model

Source: Rohmer et al. (2020).

Motivation

Left: BRGM
borehole
database

Right: example
of borehole data

Profondeur	Formation	Lithologie	Lithologie	Stratigraphie	Altitude
14.00	Formations d'épandages, de piedmont ou de glacis et dépôts résiduels associés		Limon argileux	Quaternaire	174.00
30.00			Calcaire beige oxydé		158.00
60.00			Calcaire beige altérée et calcite		128.00
65.00			Calcaire fissuré		123.00
90.00			Calcaire beige à passage fissuré à 80m		98.00
175.00			Calcaire gris fin, recimenté de calcite	Jurassique	13.00
235.00			Calcaire fin, beige légèrement oxydé		-47.00
265.00			Calcaire fin, passage oxydé et argile		-77.00
340.00			Calcaire fin, beige, massif	Jurassique inférieur	-152.00
355.00			Calcaire gris, fin, argileux		-167.00
370.00			Marne noire		-182.00
402.00			Alternance de calcaire gris et de marne noire		-214.00

1. Introduction
1. Supervised Classification of Lithological Descriptions
1. Clustering of Lithological Descriptions
1. Conclusion

1. Introduction – Definitions and Background

- **Borehole** : a drilled hole providing information on the sequence of subsurface materials.
- **Lithological layer** : a depth interval characterized by a relatively homogeneous material type, such as clay, sand, limestone or marl.
- **Lithology** : description of the nature and characteristics of a geological material.
- **Stratigraphy** : the study of the succession and arrangement of geological layers.
- **Lithofacies** : a type of rock or sediment defined by its lithological properties (mineralogy, colour, grain size, etc.).

Key references

- **Ghorbanfekr, Kerstens and Dirix (2025b)** → Automated classification of borehole lithological descriptions / **GEOBERTje**.
- **Grootendorst (2022)** → BERTopic method used for clustering.
- **Kim (2022)** → Lithofacies classification from well data using machine learning.

Research question: How can heterogeneous lithological descriptions be automatically processed to structure and characterize borehole data?

1. Introduction – Scientific Challenges

- **Highly heterogeneous lithological descriptions**
→ Difficulty converting free text into comparable categories.
- **Stratigraphic continuity is difficult to represent**
→ Need to combine textual information with vertical context.
- **No universal reference classification**
→ Difficulty validating automatically identified groups and transferring the method across regions.
- **Sparse and irregularly distributed spatial data**
→ 3D interpolation may create artificial geological continuities.

1. Introduction – Objectives

Objectives :

1. Define a representative lithostratigraphic soil column for Aix-en-Provence.
1. Assign the soil layers defined in the representative lithostratigraphic column to each borehole.
1. Interpolate between boreholes to map the spatial variation in depth to the seismic bedrock.

- **Figure 1** : Reference lithological classes – Lower Var Valley

Layer description	Code	Geology (1/50000° geological map)
Filling, concrete, Topsoil	1	Anthropogenic deposits
Sand + pebbles + gravels	2	Recent alluvium
Fine sand, Silty sand or clay sand.	3	Deltaic deposits
Silt, clay (sometimes muddy)	4	Muddy deposits and pro-delta
Sand + pebbles + gravels	5	Recent alluvium. Fluvial deposits
Fine sand, Silty sand or clay sand	6	Deltaic deposits
Silt, clay (sometimes muddy)	7	Muddy deposits and pro-delta
Coarse sand	8	Ancient alluvium. Fluvial deposits
Sand + pebbles + gravels	9	Ancient alluvium. Fluvial deposits
Bedrock	R	Pliocene, Cretaceous, Jurassic, Trias

Source: Rohmer et al. (2020).

Data :

- Boreholes from the BRGM French Subsurface Database (BSS).
- **Two study areas** : the **Lower Var Valley** (190 boreholes, 1,214 layers) and the **Aix-en-Provence** area (154 boreholes, 1,141 layers).
- **Available variables** : identifier, Lambert-93 coordinates, elevation, PRD/TDEP (depth) and lithology.

- **Figure 2** : Example of a digitized borehole log – Lower Var Valley

Profondeur	Formation	Lithologie	Lithologie	Stratigraphie	Altitude
0.70	Fz	~ ~ ~ ~ ~	Terre végétale.	Holocène	9.90
0.90		~ ~ ~ ~ ~	Argile et sable fin. Alluvions du Var.		9.70
1.95		~ ~ ~ ~ ~	Graviers et galets.		8.65
2.85		~ ~ ~ ~ ~	Galets		7.75
3.00		~ ~ ~ ~ ~	Sable fin.		7.60
6.50		~ ~ ~ ~ ~	Galets et sable fin.		4.10
8.45		~ ~ ~ ~ ~	Galets et graviers.		2.15
9.50		~ ~ ~ ~ ~	Graviers et peu de galets.		1.10
10.50		~ ~ ~ ~ ~	Graviers et peu de sable.		0.10
13.80		~ ~ ~ ~ ~	Sable fin et peu de graviers.		-3.20
14.75	Fy-z	~ ~ ~ ~ ~	Galets et peu de sable	Würm à Holocène	-4.15
16.10		~ ~ ~ ~ ~	Galets et sable fin.		-5.50
18.60		~ ~ ~ ~ ~	Galets, graviers et sable.		-8.00
18.90		~ ~ ~ ~ ~	Argile.		-8.30
20.35		~ ~ ~ ~ ~	Galets, graviers et sable.		-9.75
22.10		~ ~ ~ ~ ~	Graviers et sable agglomérés.		-11.50
24.60		~ ~ ~ ~ ~	Graviers et galets.		-14.00
24.80		~ ~ ~ ~ ~	Graviers et sable.		-14.20
25.80		~ ~ ~ ~ ~	Argile, sable et galets.		-15.20
27.50		~ ~ ~ ~ ~	Galets, sable et graviers.		-16.90
28.00		~ ~ ~ ~ ~	Sable très comprimé.		-17.40
28.40		~ ~ ~ ~ ~	Sable fin.		-17.80
32.45		~ ~ ~ ~ ~	Graviers, galets et sable aggloméré.		-21.85
		~ ~ ~ ~ ~	Graviers et galets.		

Source: BRGM website.

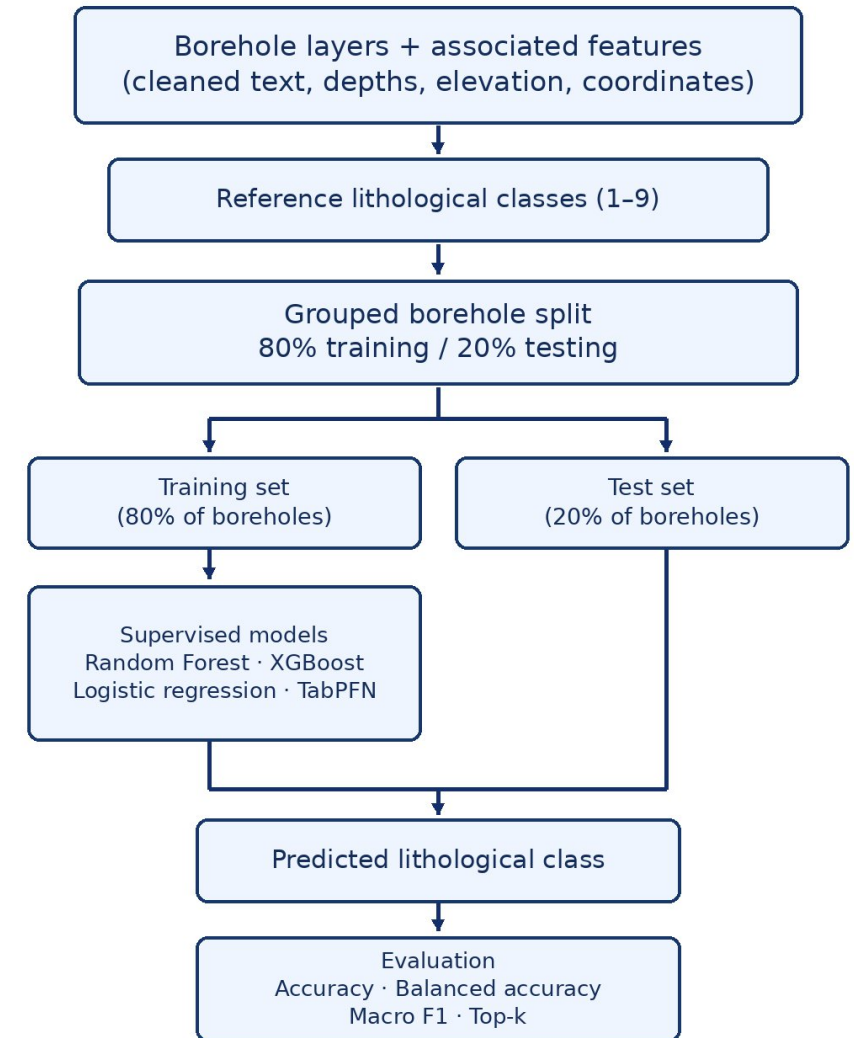
2. Supervised Classification – Method

Objective : Automatically predict the lithological class of an unknown layer.

- Four models compared : XGBoost, Random Forest, TabPFN and logistic regression.
- Train/test split by borehole : 80% / 20%.

Model	Principle
TabPFN	Pretrained Transformer for tabular data
Random Forest	Ensemble of decision trees
XGBoost	Sequentially built trees that correct previous errors
Logistic Regression	Probabilistic linear model

Supervised classification workflow



All layers from a borehole remain in the same split.

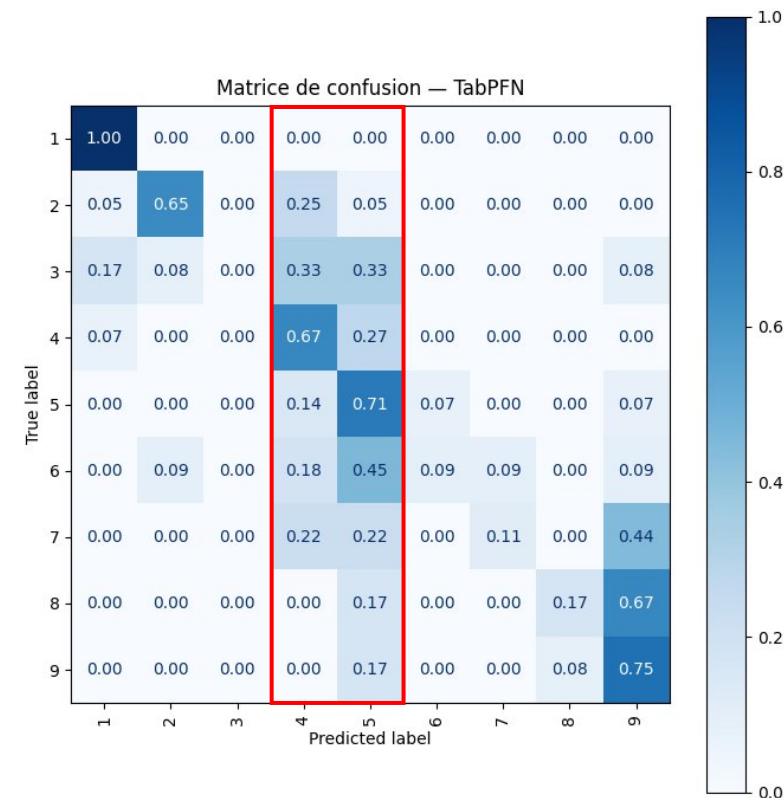
2. Supervised Classification – Results

Modèle	Accuracy	Balanced Accuracy	F1 Macro
TabPFN	54.6 %	46.1 %	41.8 %
Random Forest	53.8 %	45.5 %	43.5 %
XGBoost	52.1 %	45.6 %	45 %
Régression logistique	35.3 %	31.9 %	27.3 %

- TabPFN achieves the highest accuracy. Performance is relatively similar across the three leading models.

Modèle	Top-1 accuracy	Top-2 accuracy	Top-3 accuracy
TabPFN	54.6%	69.7%	81.5%
Random Forest	53.8%	70.6%	79.8%
XGBoost	52.1%	64.7%	79.0%
Logistic Regression	35.3%	61.3%	80.7%

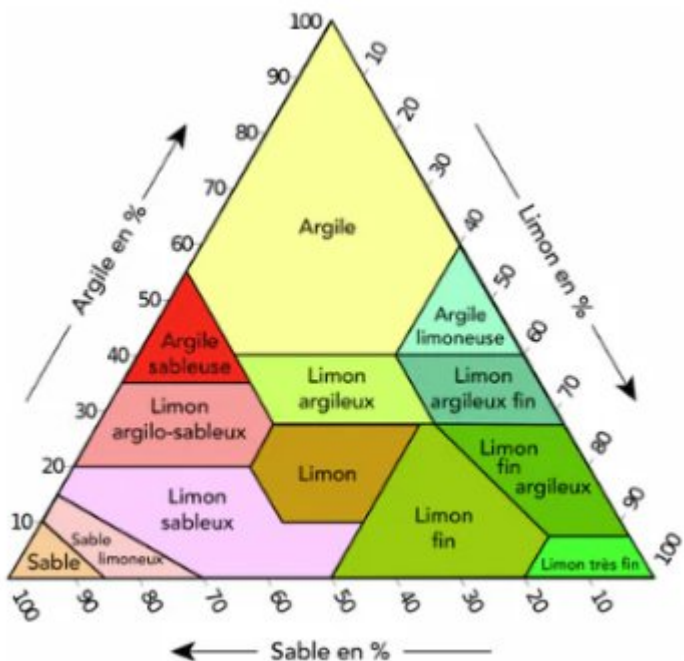
- Top-1 accuracy ranges from approximately 50% to 55% for the three best-performing models. Considering the three most probable classes, the true class is included in about 80% of cases.



- Classes 1, 2, 4, 5 and 9 are recognized relatively well.
- Class imbalance and the lithological similarity of some categories affect performance.

Classe	1	2	3	4	5	6	7	8	9	Total
Effectif	106	88	48	87	87	44	39	26	62	587

2. Supervised Classification – Results



- Aim: predict clay, silt and sand percentages using Random Forest regression.
- $R^2 = 0.23$

Step	Model	Accuracy	Balanced acc.	Macro F1
1 m	TabPFN	52.2%	40.2%	36.8%
	Random Forest	47.2%	36.9%	35.0%
	XGBoost	44.5%	35.5%	35.3%
	Logistic reg.	27.8%	26.6%	21.9%
0.5 m	TabPFN	52.6%	40.2%	37.0%
	Random Forest	47.0%	36.6%	35.6%
	XGBoost	41.4%	32.5%	32.5%
	Logistic reg.	28.1%	27.6%	22.9%
0.1 m	TabPFN	51.2%	39.6%	36.8%
	Random Forest	46.1%	36.3%	35.6%
	XGBoost	38.9%	30.7%	30.6%
	Logistic reg.	27.9%	27.6%	22.8%

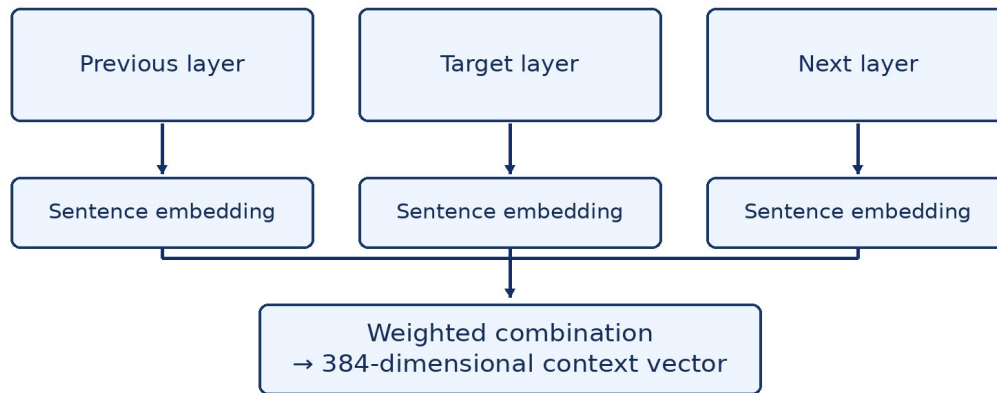
Dataset	Observations
Original data	587
1 m discretization	4,596
0.5 m discretization	8,946
0.1 m discretization	43,969

3. Automatic Clustering – Method (BERTopic)

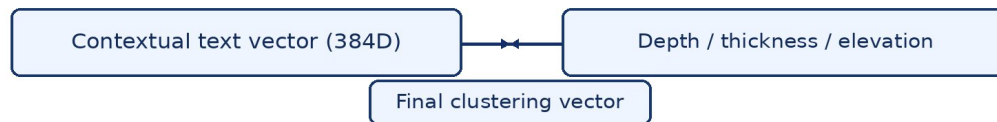
Objective : Automatically identify coherent lithological groups from layer descriptions.

- No predefined lithological classes.
- Layer representations are enriched with **vertical context** and **depth/elevation variables**.

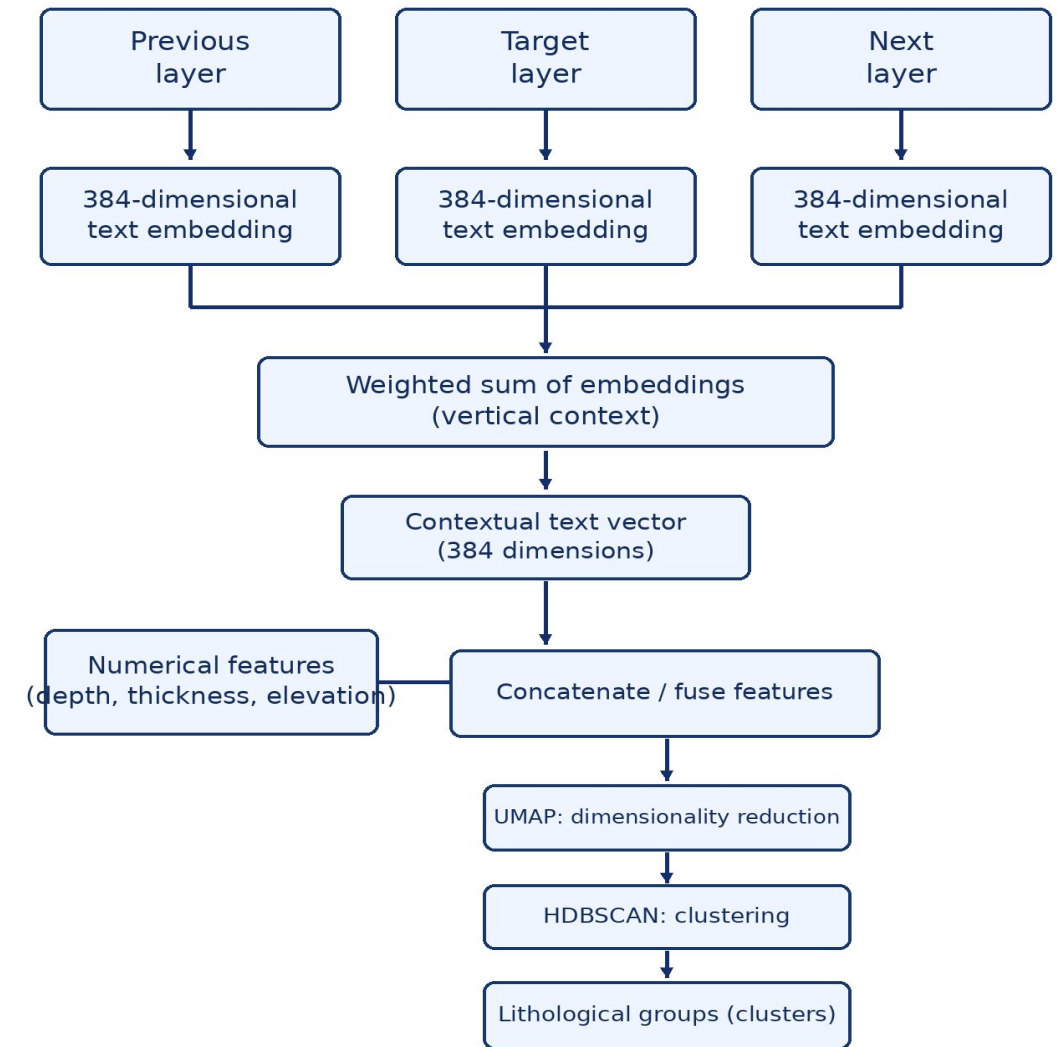
1. Text representation



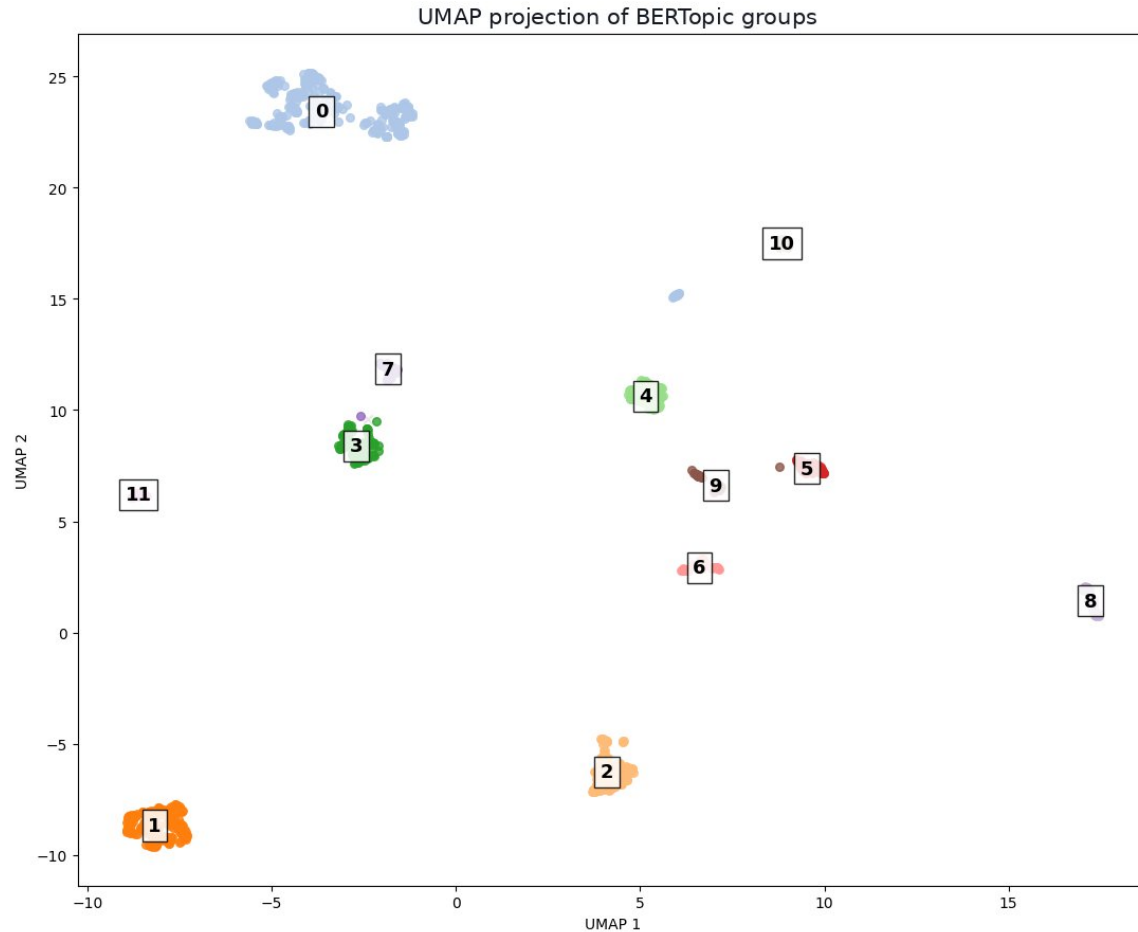
2. Feature fusion



Context-aware lithological clustering

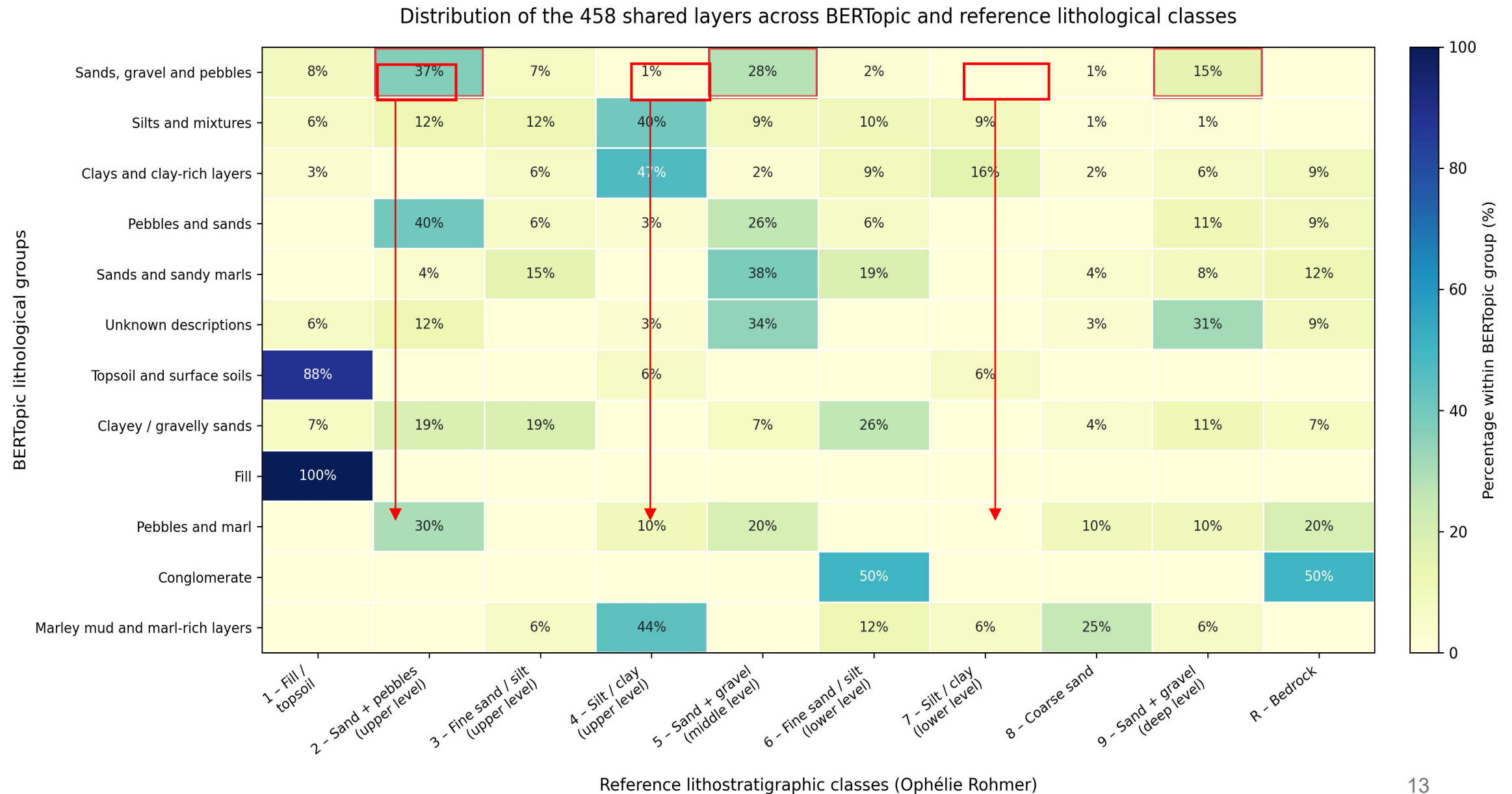


3. Automatic Clustering – Lower Var Valley



Cluster	Layers	Lithological interpretation
-1	1	Noise / unassigned
0	344	Gravelly sands and pebbles
1	190	Silts and silty mixtures
2	145	Clays and clay-rich layers
3	106	Pebbles and sands
4	94	Sands and sandy marls
5	62	Unknown descriptions
6	54	Topsoil
7	53	Clayey or gravelly sands
8	53	Made ground / fill
9	49	Pebbles and marl
10	34	Conglomerate
11	29	Marly mud

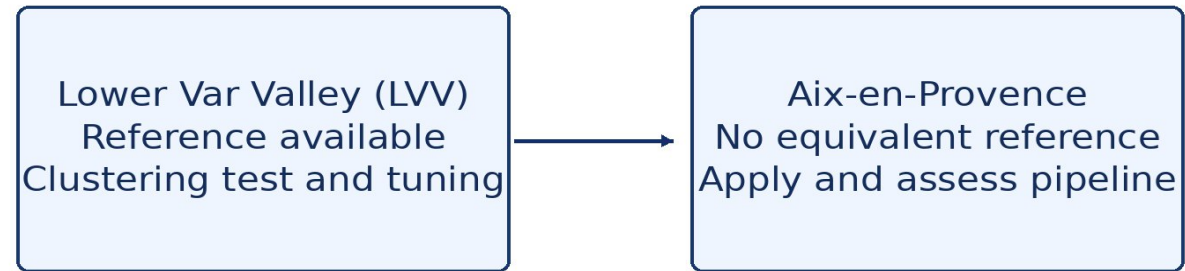
3. Automatic Clustering – Comparison with Reference (LVV)



3. Automatic Clustering – From the LVV to Aix-en-Provence

Why the Lower Var Valley (LVV)?

- The Lower Var Valley has an existing reference classification.
- It allows the clustering method to be tested, tuned and evaluated.
- The resulting groups can be compared with that reference.



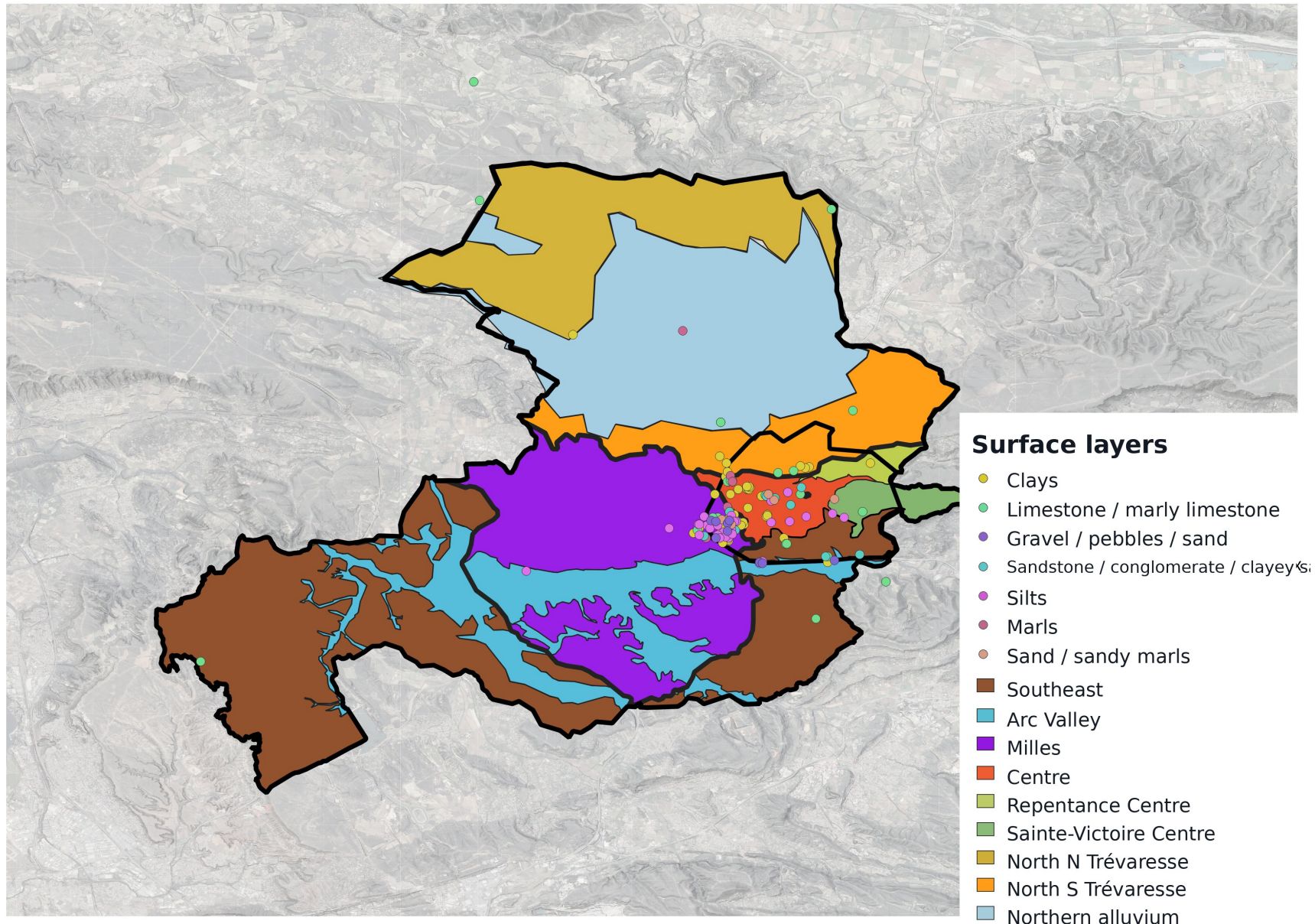
LVV = method evaluation → Aix-en-Provence = application

What about applying the method to Aix-en-Provence?

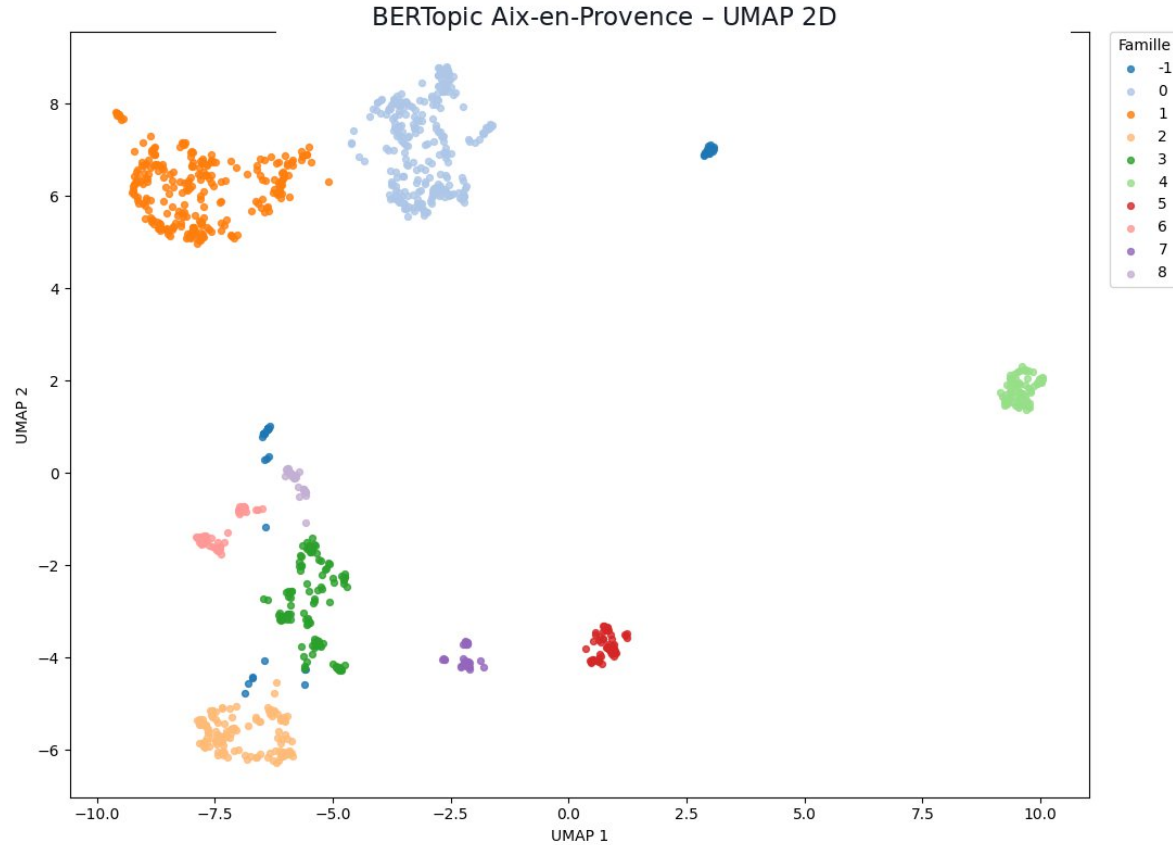
- No equivalent reference classification is available.
- The method evaluated in the LVV is therefore transferred to Aix.
- Evaluation focuses more on the geological coherence of groups and clustering stability.

3. Automatic Clustering – Second Geological Map by Thibaut Dunet

- Simplified geological map of Aix-en-Provence (source: Thibaut Dunet, 2026).



3. Automatic Clustering – Aix-en-Provence Case Study



Cluster	Layers	Lithological interpretation
-1	53	Noise
0	273	Clays
1	272	Limestones and marly limestones
2	138	Marls
3	130	Sandstones, conglomerates, clayey sands
4	81	Silts
5	65	Topsoil
6	57	Sands and sandy marls
7	42	Fill
8	30	Gravels, pebbles and sands

4. Conclusion

Limitations

- Highly heterogeneous lithological descriptions and a limited dataset.
- **Irregular spatial distribution of observations.**
- Insufficient **constraints** to build a reliable 3D geological model.

Future Work

- Refine selected groups through a **second clustering step**.
- Revisit the data **cleaning process**.
- Incorporate additional geological information (e.g. **formation ages and graphical lithological logs**).

Layer description	Code	Geology (1/50000 ^e geological map)
Filling, concrete, Topsoil	1	Anthropogenic deposits
Sand + pebbles + gravels	2	Recent alluvium
Fine sand. Silty sand or clay sand.	3	Deltaic deposits
Silt, clay (sometimes muddy)	4	Muddy deposits and pro-delta
Sand + pebbles + gravels	5	Recent alluvium. Fluvial deposits
Fine sand. Silty sand or clay sand	6	Deltaic deposits
Silt, clay (sometimes muddy)	7	Muddy deposits and pro-delta
Coarse sand	8	Ancient alluvium. Fluvial deposits
Sand + pebbles + gravels	9	Ancient alluvium. Fluvial deposits
Bedrock	R	Pliocene, Cretaceous, Jurassic, Trias

Source: Rohmer et al. (2020).

Profondeur	Formation	Lithologie	Lithologie	Stratigraphie	Altitude
0.70	Fz	~ ~ ~ ~	Terre végétale.	Holocène	9.90
0.90		~ ~ ~ ~	Argile et sable fin. Alluvions du Var.		9.70
1.95		~ ~ ~ ~	Graviers et galets.		8.65
2.85		~ ~ ~ ~	Galets.		7.75
3.00		~ ~ ~ ~	Sable fin.		7.60
6.50		~ ~ ~ ~	Galets et sable fin.		4.10
8.45		~ ~ ~ ~	Galets et graviers.		2.15
9.50		~ ~ ~ ~	Graviers et peu de galets.		1.10
10.50		~ ~ ~ ~	Graviers et peu de sable.		0.10
13.80		~ ~ ~ ~	Sable fin et peu de graviers.		-3.20
14.75		~ ~ ~ ~	Galets et peu de sable.		-4.15
16.10		~ ~ ~ ~	Galets et sable fin.		-5.50
18.60		~ ~ ~ ~	Galets, graviers et sable.		-8.00
18.90		~ ~ ~ ~	Argile.		-8.30
20.35		~ ~ ~ ~	Galets, graviers et sable.		-9.75
22.10	Fy-z	~ ~ ~ ~	Graviers et sable agglomérés.	Würm à Holocène	-11.50
24.60		~ ~ ~ ~	Graviers et galets.		-14.00
24.80		~ ~ ~ ~	Graviers et sable.		-14.20
25.80		~ ~ ~ ~	Argile, sable et galets.		-15.20
27.50		~ ~ ~ ~	Galets, sable et graviers.		-16.90
28.00		~ ~ ~ ~	Sable très comprimé.		-17.40
28.40		~ ~ ~ ~	Sable fin.		-17.80
32.45		~ ~ ~ ~	Graviers, galets et sable aggloméré.		-21.85
			Graviers et galets.		

5. References

Borehole Data Classification

Ghorbanfekr, Kerstens & Dirix (2025b) — GEOBERTje, borehole description classification.
Kim (2022b) — automated lithofacies classification.
Pawley et al. (2024) — depth to bedrock, borehole data and spatial variables.

Classification Methods

Hollmann et al. (2023) — TabPFN.
Biau & Scornet (2016) — Random Forest.
Chen & Guestrin (2016) — XGBoost.
James et al. (2023) — logistic regression and statistical learning.
Pedregosa et al. (2011) — scikit-learn.

Clustering / Text Representations

Grootendorst (2022) — BERTopic.
Reimers & Gurevych (2019) — Sentence-BERT.
McInnes, Healy & Melville (2018) — UMAP.
McInnes, Healy & Astels (2017) — HDBSCAN.
Chango et al. (2022) — feature fusion.
Warrens & van der Hoef (2022) — Adjusted Rand Index.

Annexes

Model type: pretrained Transformer neural network for tabular data. TabPFN = Tabular Prior-Data Fitted Network.

Performs well on small to medium-sized tabular datasets.

Few hyperparameters to tune.

A recent approach, quite different from tree-based methods.

Useful when the number of observations is relatively limited.

How does TabPFN work?

Transformer pretrained on a large number of synthetic tabular classification problems.

For a new task, it receives the training dataset and the observations to be predicted.

Its pretraining allows it to directly output class membership probabilities without training a neural network from scratch.

An attention mechanism captures relationships between observations, features and classes.

Transformer: a neural network architecture based on attention, enabling the model to identify which information is most relevant for prediction.

- Multiclass problem: 9 lithological classes
- Relatively small tabular dataset
- Output: probability for each class
- Highest accuracy: 54.6%
- **Top-3 accuracy: 81.5%**

Model type : an ensemble of decision trees.

A decision tree repeatedly splits the data (e.g. depth ≤ 5 m? Then elevation, etc.). Each split aims to make the resulting groups as homogeneous as possible in terms of class labels.

A single tree can overfit easily. Random Forest therefore builds many different trees :

Data $\rightarrow$ Tree₁, Tree₂, Tree₃, ..., Tree₅₀₀ $\rightarrow$ Aggregate predictions

Each tree is trained on:

- A different bootstrap sample of the observations
- A random subset of features

The final prediction aggregates the results from all trees.

Advantages :

- strong performance on tabular data;
- captures nonlinear relationships;
- robust to noise;
- limited sensitivity to feature scaling;
- handles complex feature interactions.

Parameters :

- **n_estimators = 500**
- **class_weight = "balanced"**
- **random_state = 42**

Model type : an ensemble of decision trees trained using gradient boosting.

Key difference from Random Forest:

Random Forest builds trees relatively independently, then combines their predictions.
XGBoost builds trees sequentially:

Tree₁ → errors → Tree₂ → residual errors → Tree₃ → ...

Each new tree attempts to correct the errors (or residuals) of previous trees.

- Particularly effective on tabular data
- Models complex nonlinear relationships
- Includes several regularization mechanisms to reduce overfitting

Parameters

- **n_estimators = 300**
- **max_depth = 5**
- **learning_rate = 0.05**
- **subsample = 0.9**
- **colsample_bytree = 0.9**

Model type : probabilistic linear statistical model.

Principle

- Computes a linear combination of input features
- Transforms this combination into class membership probabilities
- For multiclass classification, outputs one probability for each lithological group
- Selects the class with the highest probability

$X_1, X_2, \dots, X_p \rightarrow \text{linear combination} \rightarrow P(C_1), P(C_2), \dots, P(C_k) \rightarrow \text{predicted class}$

- Simple and fast
- Easy to interpret
- Serves as a baseline to assess whether more complex models provide added value

Limitation

Decision boundaries are essentially linear, making this model less suitable if relationships between input features and lithology are highly complex.

Ghorbanfekr, Kerstens & Dirix (2025b)

→ Automated classification of geological borehole descriptions with GEOBERTje.

Kim (2022b)

→ Identification/classification of lithofacies from well data using machine learning.

Grootendorst (2022)

→ BERTopic.

Pawley et al. (2024)

→ Spatial and depth information in geological borehole datasets.

Chango et al. (2022)

→ Feature fusion / multimodal data.

GEOBERTje — Ghorbanfekr, Kerstens & Dirix (2025b)

Automate the classification of free-text geological borehole descriptions.

- Based on the Dutch **BERTje** architecture;
- approximately **109 million parameters**;
- adapted to the geological domain using **Masked Language Modeling**;
- approximately **300,000 Dutch borehole descriptions**;
- then **fine-tuned on about 3,000 annotated** descriptions for lithological classification.

The model predicts ;

- primary lithology;
- secondary lithology;
- tertiary lithology.

For example, “fine clayey sand with some gravel” can be decomposed into fine sand + clay + gravel.

Same initial challenge: transforming heterogeneous free-text geological descriptions into structured lithological information.

Kim (2022b) — Lithofacies Classification

Automatically identify lithofacies in the Eagle Ford and Austin Chalk formations (oil and gas-bearing geological formations in Texas).

Data

- core data;
- thin sections / imagery;
- geophysical well logs.

Four lithofacies were first defined geologically using properties such as texture, mineralogy and sedimentary structures.

Kim then trains a one-dimensional convolutional neural network (**1D-CNN**) using five logs:

- gamma ray;
- density;
- neutron porosity;
- deep resistivity;
- sonic log.

The model learns patterns in the vertical sequence of logs to predict lithofacies. It is evaluated by cross-validation and a blind-well test.

The paper shows that lithology/lithofacies can be inferred automatically from well data and that the vertical structure contains valuable information.

Pawley et al. (2024)

Map depth to bedrock in Alberta, Canada.

Approach

- Use of borehole lithological logs;
- use of a language model to expand the number of known bedrock-depth measurements (i.e. boreholes with an identified bedrock depth);
- machine learning models: **XGBoost, Random Forest and Cubist**;
- addition of spatial features: coordinates, proximity, neighbouring observations and spatial-lag variables.

Main finding

Adding spatial-lag features improves predictions compared with using geographic coordinates alone.

This supports combining text with complementary information, particularly **depth and spatial context**, to characterize borehole data.

B3 – Evaluation Metrics

Accuracy : overall proportion of correct predictions.

Balanced accuracy : mean of the recall scores calculated for each class. Gives equal weight to every class, even rare ones.

Recall : proportion of correctly identified observations among those truly belonging to a given class.

Precision : proportion of observations that truly belong to a class among those predicted as that class.

F1 score : harmonic mean of precision and recall; it is high only when both are high.

Macro F1 : unweighted average of the F1 scores across classes. Balances precision and recall without favouring larger classes.

Top-k accuracy : a prediction is correct if the true class is among the k highest-probability classes.

B3 – Evaluation Metrics

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K F1_k$$

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^K \text{Rappel}_k$$

TP : true positives **FP** : false positives **FN** : false negatives **K** : number of classes (here K = 9).

UMAP (Uniform Manifold Approximation and Projection) : Dimensionality reduction that preserves the local data structure as much as possible. It reduces vector dimensionality before clustering to make the task easier for HDBSCAN.

- The final vector for each layer combines textual information and numerical features.
- UMAP identifies neighbouring layers in this high-dimensional feature space.
- It then projects the data into a lower-dimensional space to facilitate clustering.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) : Density-based clustering algorithm that automatically identifies dense groups and can label some observations as noise. Unlike k-means, the number of groups need not be specified beforehand.

- It identifies observations located in sufficiently dense regions.
- It builds a hierarchy of groups at different density levels.
- It retains the most stable groups as clusters.
- Observations that cannot be assigned confidently are labelled as noise (-1).

UMAP

- **n_neighbors**

Number of neighbours used to construct the local data structure.
→ Small: more local structure; larger: more global structure.

- **n_components**

Number of dimensions in the reduced space used for clustering.
→ Here, HDBSCAN uses a 5-dimensional representation.

- **min_dist**

Controls how closely points can be placed in the reduced space.
→ A low value yields more compact groups.

- **metric = "cosine"**

Measures vector similarity based on orientation.
→ Well suited to text embeddings.

- **random_state = 42**

Fixes randomness to ensure reproducible dimensionality reduction.

HDBSCAN

- **min_cluster_size**

Minimum number of observations required to form a cluster.
→ Prevents very small clusters.

- **min_samples**

Controls how strictly observations must belong to a dense region.
→ Higher values make HDBSCAN more conservative, possibly increasing noise.

- **metric = "euclidean"**

Distance used to compare observations in the UMAP-reduced space.

- **cluster_selection_method = "eom"**

Excess of Mass: selects the most stable clusters from the HDBSCAN hierarchy.

UMAP

- `n_neighbors = 35`
- `n_components = 5`
- `min_dist = 0.1`
- `metric = "cosine"`
- `random_state = 42`

HDBSCAN

- `min_cluster_size = 25`
- `min_samples = 10`
- `metric = "euclidean"`
- `cluster_selection_method = "eom"`

Feature weights

- Target layer: 1.00
- Previous layer: 0.50
- Next layer: 0.50
- Depth feature block: 0.20
- Elevation feature block: 0

UMAP

- `n_neighbors = 50`
- `n_components = 5`
- `min_dist = 0.1`
- `metric = "cosine"`
- `random_state = 42`

HDBSCAN

- `min_cluster_size = 25`
- `min_samples = 10`
- `metric = "euclidean"`
- `cluster_selection_method = "eom"`

Feature weights

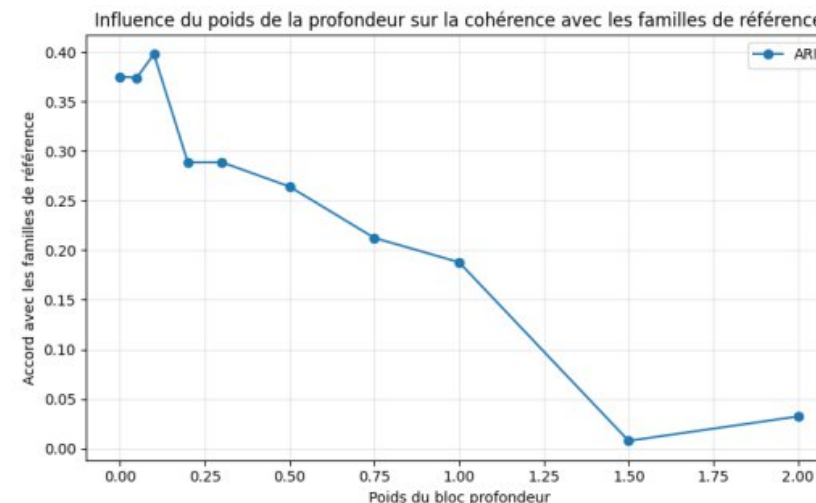
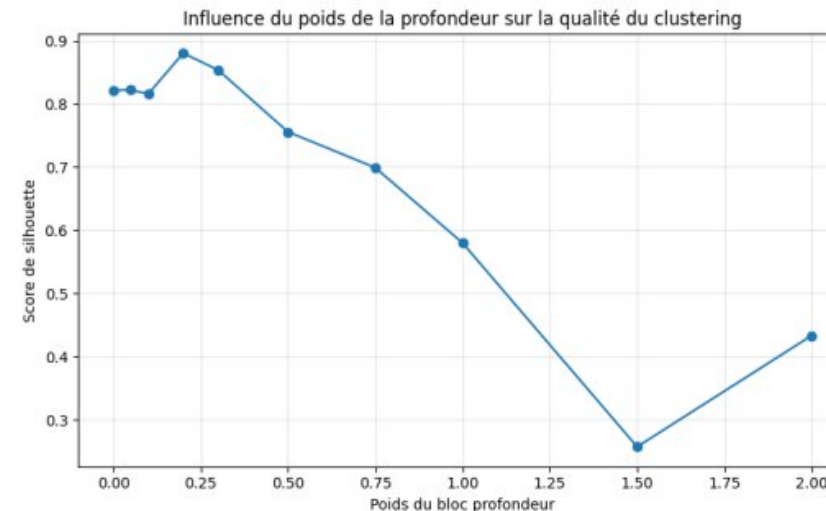
- Target layer: 1.00
- Previous layer: 0.25
- Next layer: 0.25
- Depth feature block: 0.20
- Elevation feature block: 0.10

B6 – ARI vs Silhouette Score and Random Seeds

ARI (Adjusted Rand Index) : Compares **two partitions** of the same dataset. $ARI \approx 1$ indicates identical partitions. Used here to compare clustering with the reference classification or to measure stability across runs.

Silhouette score : Measures **internal clustering quality** without reference labels. Close to 1: compact, well-separated groups. Negative: potentially misassigned observations.

Random seed	Clusters	Noise (%)	ARI vs 42
0	10	7.62	0.941
10	8	2.37	0.950
20	11	9.99	0.912
30	10	7.10	0.916
42	9	4.65	1.000



$$P_{ij} = \frac{N_{ij}}{N_i} \times 100$$

where:

P_{ij} : percentage of layers in BERTopic group i assigned to reference group j ;

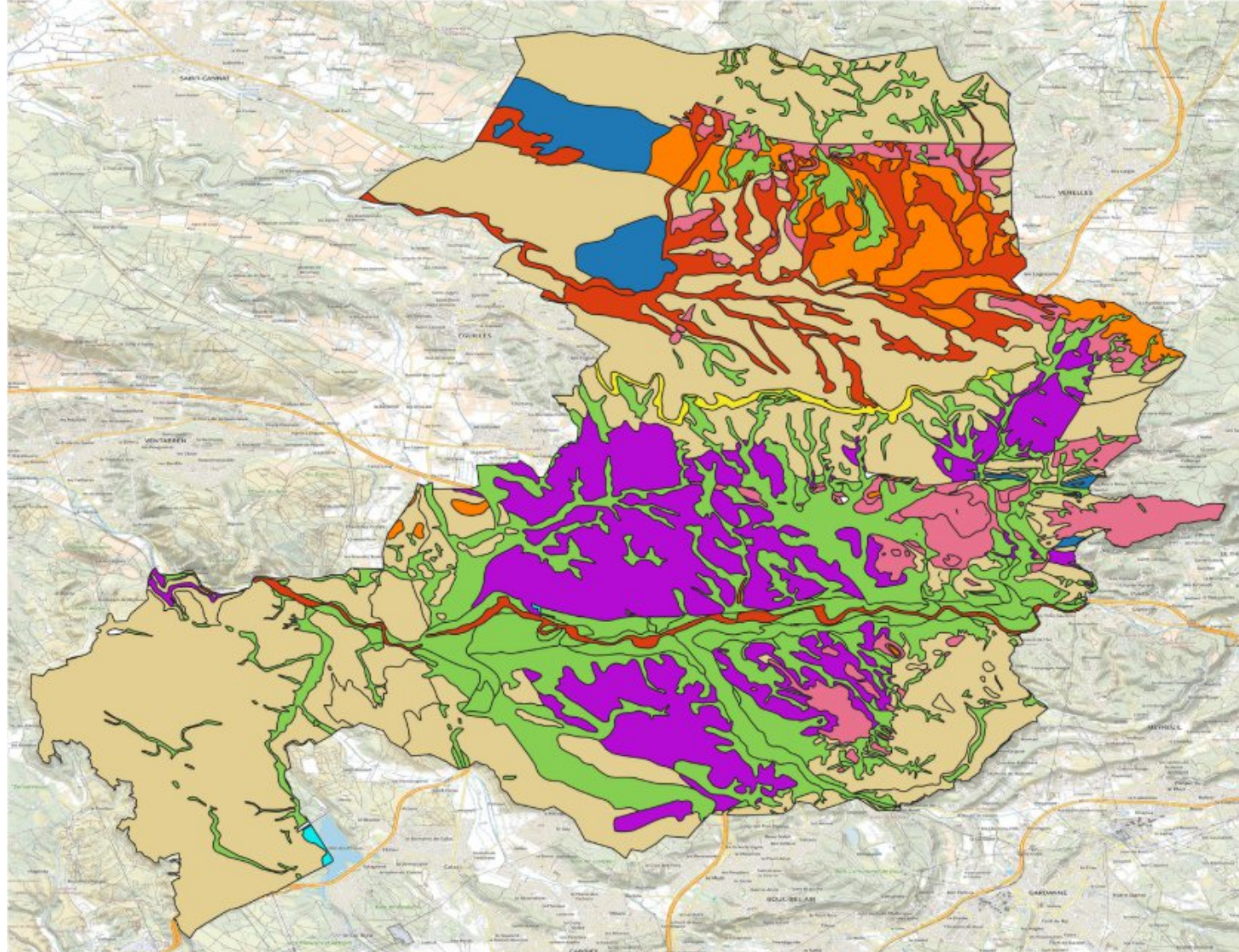
N_{ij} : number of shared layers belonging to both BERTopic group i and reference group j ;

N_i : total number of shared layers belonging to BERTopic group i .

For each BERTopic group i :

$$\sum_j P_{ij} = 100 \text{ \%}.$$

B8 – Second Geological Map by Thibaut Dunet



Legend:

-  Limestones
-  Tortonian
-  Glacial deposits
-  Quaternary
-  Clays
-  Marls
-  Conglomerates
-  Figons silica sands
-  Drainage network