Extraction automatique et structurée des données tabulaires de document PDF du BRGM

Travail de fin d'études

Marijan Soric

14 février 2025

Plan

Cadre du stage

Inria Valda (Inria Paris, DI ENS, CNRS) Thèmes : gestion de données complexes, données produites par l'activité humaine.

Inria Cedar (Inria Saclay, LIX, CNRS) Thèmes : analyse de données riches à l'échelle du Cloud.

BRGM (Bureau de Recherches Géologiques et Minières) Service géologique national français : applications des sciences de la Terre pour gérer les ressources et les risques du sol et du sous-sol.

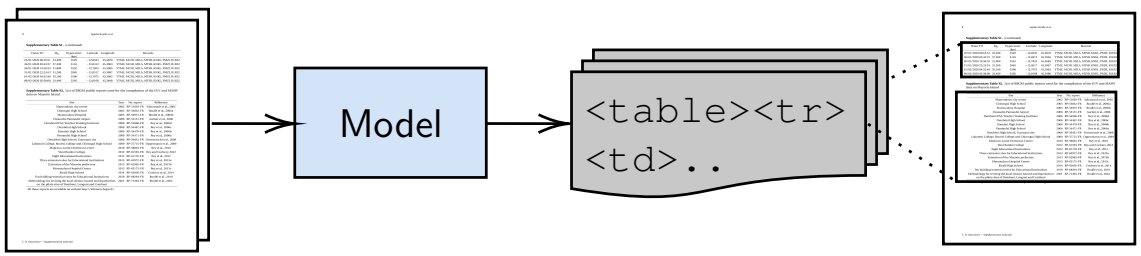
Cadre du travail

- Projet GéolAug collaboration Inria & BRGM
- Aide à la préparation des missions par les géologues, faciliter l'accès à la connaissance
- « Exploitation et structuration des données et de la connaissance »
- Des données hétérogènes :
  - Cartes géologiques, schémas
  - Bases de données
  - Texte
  - Tableaux

Plan

Définition du sujet

Extraction automatique et structurée des données tabulaires de document PDF

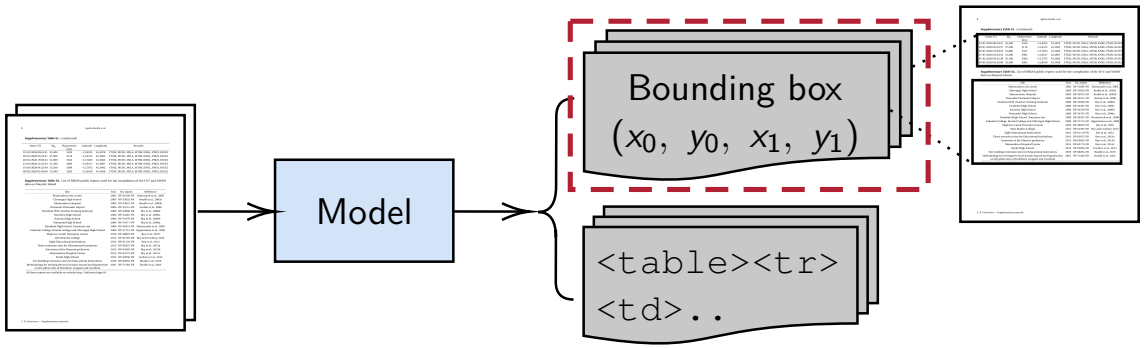


PDF Document

Extracted structured data

Définition des tâches (1/2)

Détection de tableaux : Trouver toutes les tableaux d'un document



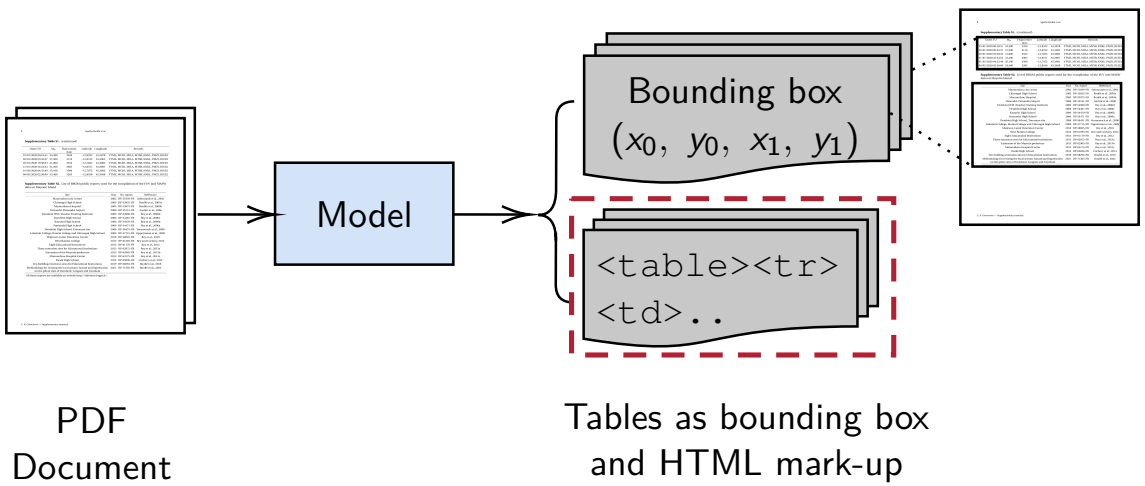
PDF Document

Tables as bounding box and HTML mark-up

Difficulté : diversité des formats de tableaux (avec ou sans bords)

Définition des tâches (2/2)

Reconnaissance de la structure de tableaux : Extraire les données du tableau avec son contenu de manière structurée

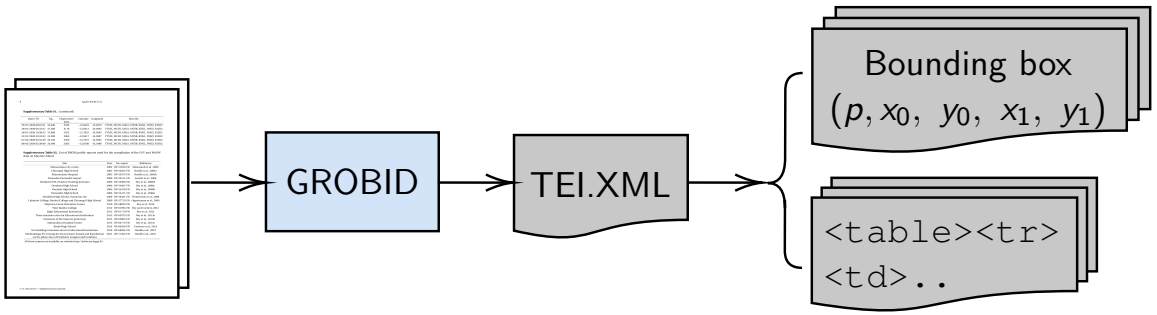


Difficulté : diversité des formats de cellules (vides, alignement...)
Extraction de tableaux : Détection + Structure

Plan

Approche de base (1/2)

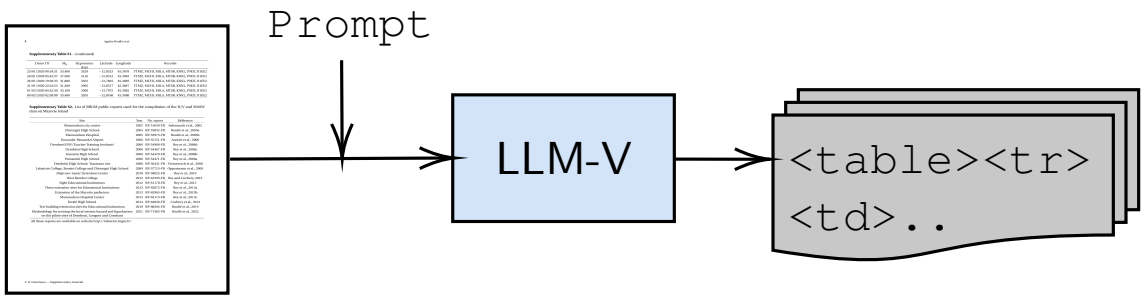
GROBID (Lopez, 2008) logiciel d'analyse de PDF, notamment utilisé dans HAL [1].



1. https://hal.science

Approche de base (2/2)

LLM-Vision GPT-4o-mini avec l'API d'OpenAI.



Note : LLM-Vision ne produit pas de coordonnées.

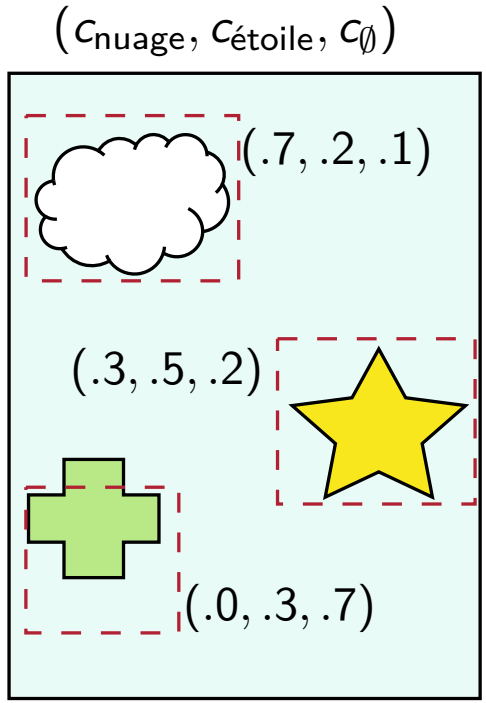
Plan

La détection d'objet

Détection d'objet : détecter la présence d'instance dans une image

Instance. Classe / objet à trouver (comme "table" ou "colonne", "ligne", "cellule").

Détection. Localisation (coins d'un rectangle) et distribution de probabilité sur les étiquettes : score de confiance.



Détection d'objet : instances (nuage, étoile, pas d'objet)

GROBID et LLM-Vision produisent des prédictions sans scores.

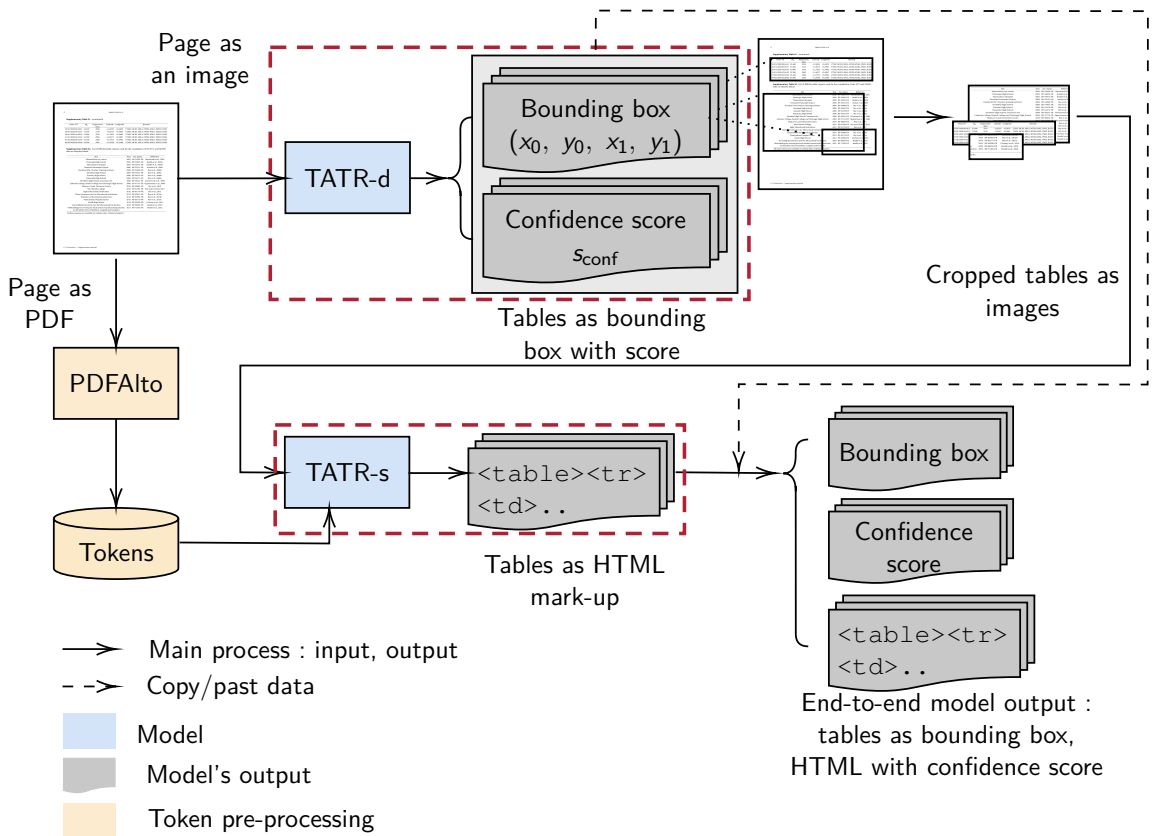
Méthodes d'extractions à deux étapes

On assemble des modèles spécialisés dans chacune des tâches.

Méthodes	**Détection de Tableaux**	**Reconnaissance de la Structure de Tableaux**
TATR-extract (Smock et al., 2022)	TATR-detect	TATR-structure
VGT+TATR-structure	VGT (Da et al., 2023)	TATR-structure
XY+TATR-extract	XY+TATR-detect	TATR-structure

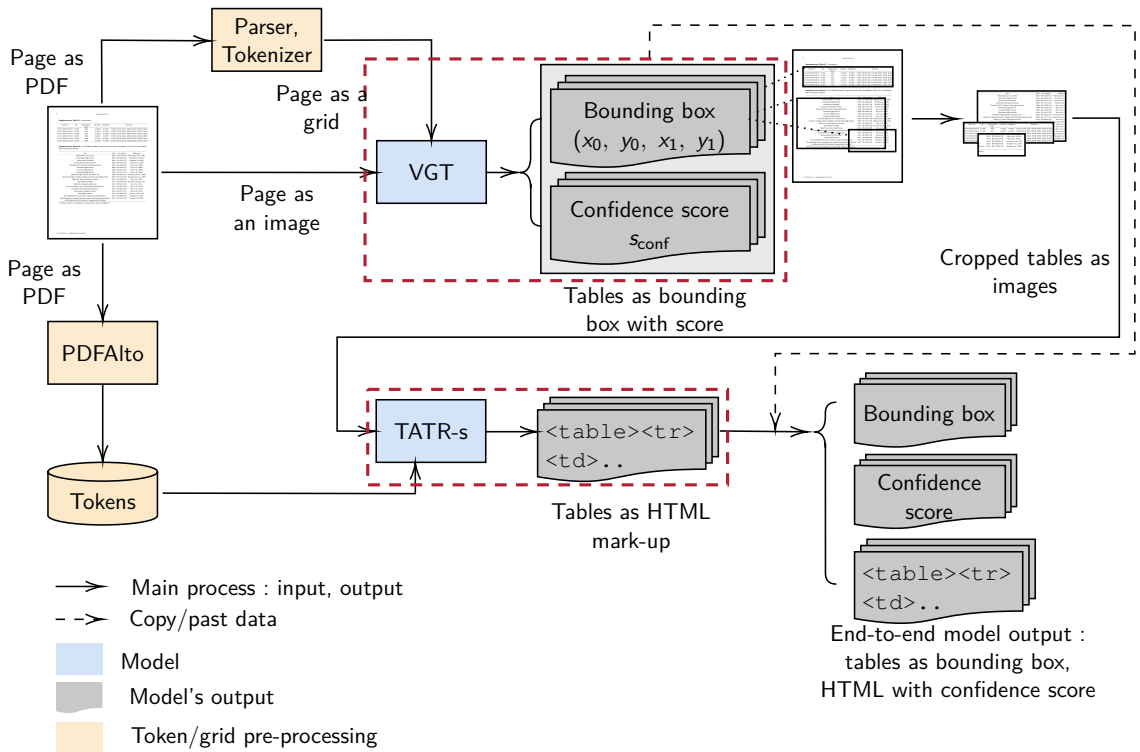
Méthodes d'extractions (1/3)

TATR-extract composé de deux modèles : TATR-detect et TATR-structure, utilisant l'architecture DETR (Carion et al., 2020) (transformeur encodeur-décodeur).



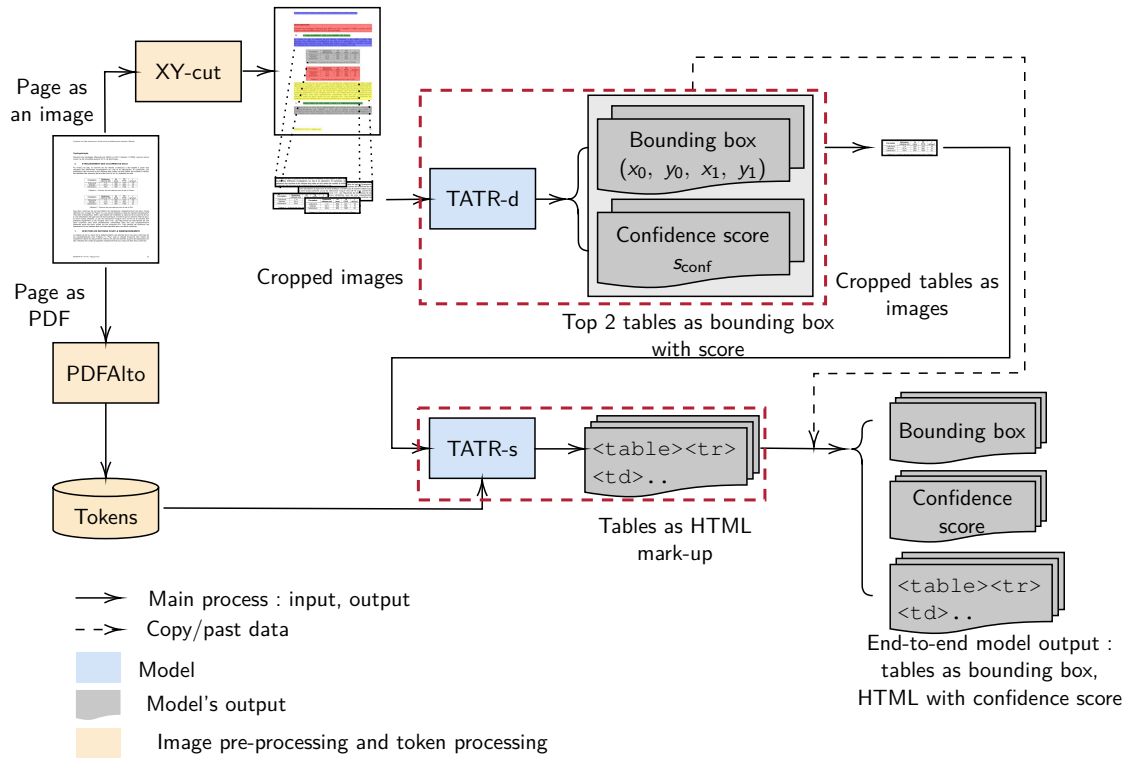
Méthodes d'extractions (2/3)

VGT+TATR-structure utilise VGT pour la détection de tableaux : modèle spécialisé dans la reconnaissance de mise en page (dont DT). VGT est multimodal : exploite le contenu visuel et textuel.



Méthodes d'extractions (3/3)

XY+TATR-extract ajoute un pré-traitement des pages grâce à l'algorithme X-Y cut (Ha et al., 1995)



Plan

Détection de tableaux

Métriques classiques utilisées : Precision, Rappel, basées sur les predictions *positives* (binaire).

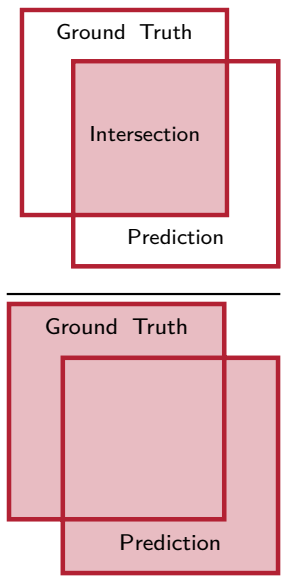
Positive	Prédiction « il y a un tableau »
True Positive (TP)	Tableau correctement détecté
False Positive (FP)	Tableau détecté qui n'est pas un vrai tableau
False Negative (FN)	Vrai tableau non détecté

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

Vrai et faux positif

Nous décidons si une prédiction (positive) est un TP ou FP, à l'aide de l'intersection-over-union (IoU) par rapport à un seuil θ_J.

$$\text{IoU} = \frac{\text{area of overlap}}{\text{area of union}} =$$



Si $\text{IoU} > \theta_J$ alors la prédiction est un TP, sinon un FP.

Métriques

Precision P_{θ_J} mesure à quel point le modèle est précis dans ses prédictions.

Rappel R_{θ_J} mesure à quel point le modèle ne rate pas les vrais tableaux.

Mais ces métriques sont sensibles au choix du seuil θ_J, d'où l'utilisation de métriques d'agrégation $\mathrm{WAvg}(P) = \sum_{\theta_J} w_{\theta_J} P_{\theta_J}$.

Average Precision Aire sous la courbe Precision–Rappel pour les modèles avec scores de confiance.

Calibration du modèle Vérifie que les scores de confiance correspondent à des probabilités.

Plan

Reconnaissance de la structure de tableaux

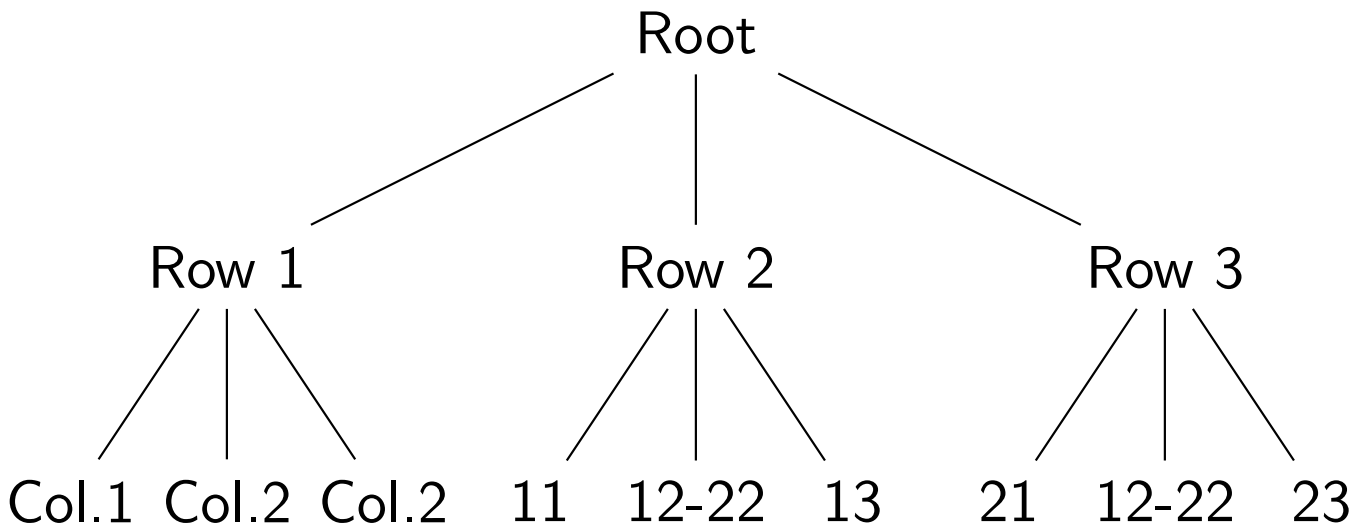
- Les métriques peuvent utiliser :
  - Localisations absolues de la structure (lignes, colonnes, cellules) comme pour la détection de tableaux.
  - Positions relatives des cellules et structure globale comme TEDS (Li et al., 2020) et GriTS (Smock et al., 2023).
- Évaluation des méthodes d'extraction dans leur ensemble : la phase de reconnaissance de structure dépend de la phase de détection.

Métriques (1/2)

TEDS calcule la similarité entre les tableaux représentés sous forme d'arbres.

Col.1	**Col.2**	**Col.3**
11	12-22	13
21		23

Représenté sous forme d'arbre :



Métriques (2/2)

GriTS représentent les tableaux sous forme matricielle et calculent leur similarité sur différents types.

Col.1	**Col.2**	**Col.3**
11	12-22	13
21		23

GriTS Content contenu textuel du tableau

$$\begin{pmatrix} \text{Col.1} & \text{Col.2} & \text{Col.3} \\ 11 & 12-22 & 13 \\ 21 & 12-22 & 23 \end{pmatrix}$$

GriTS Topology structure du tableau (topologie)

$$\begin{pmatrix} (0,0,1,1) & (0,0,1,1) & (0,0,1,1) \\ (0,0,1,1) & (0,0,1,2) & (0,0,1,1) \\ (0,0,1,1) & (0,-1,1,1) & (0,0,1,1) \end{pmatrix}$$

Plan

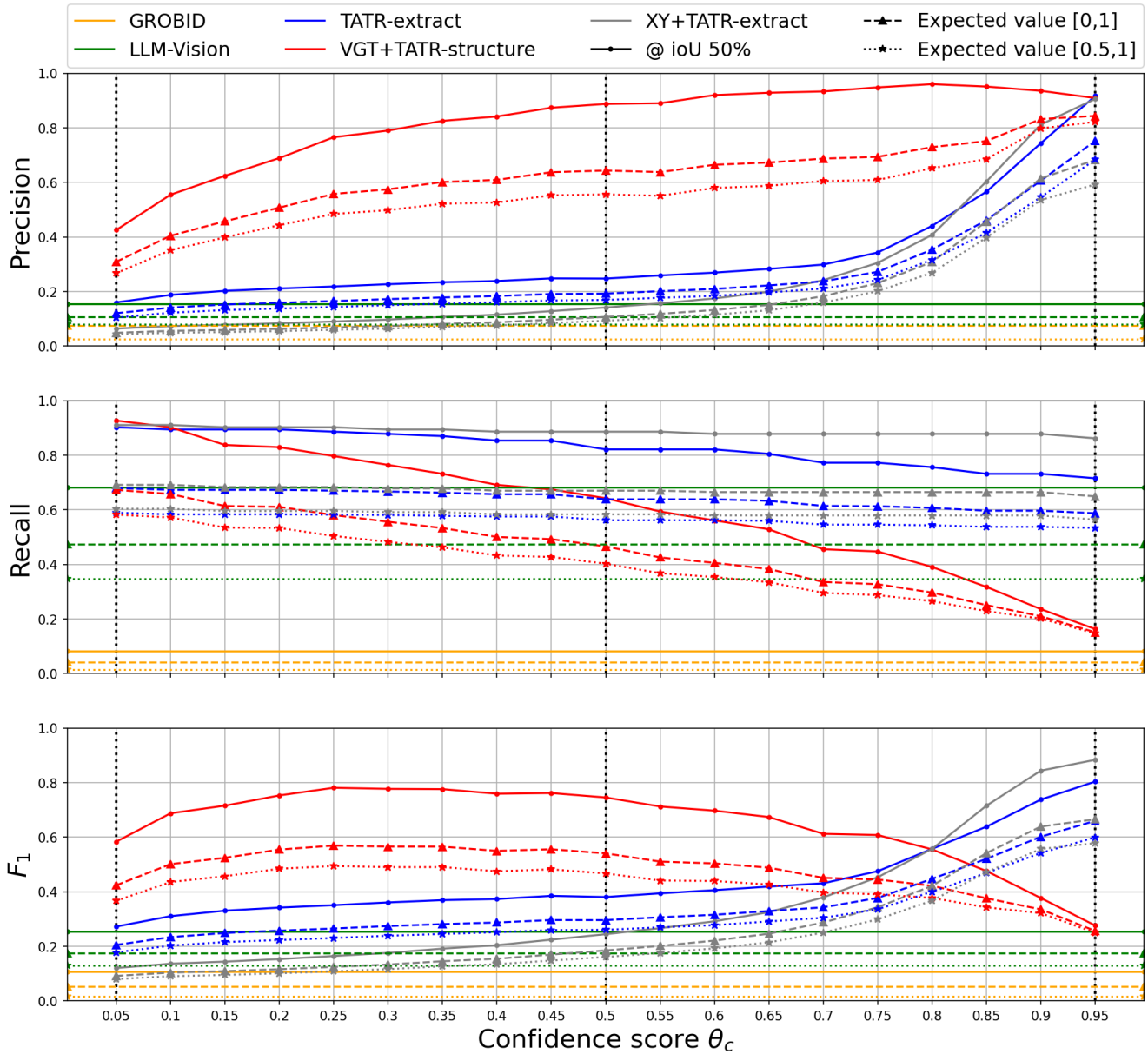
Scores de confiance

On définit un seuil θ_c pour définir les prédictions *positives* des modèles avec scores de confiance.

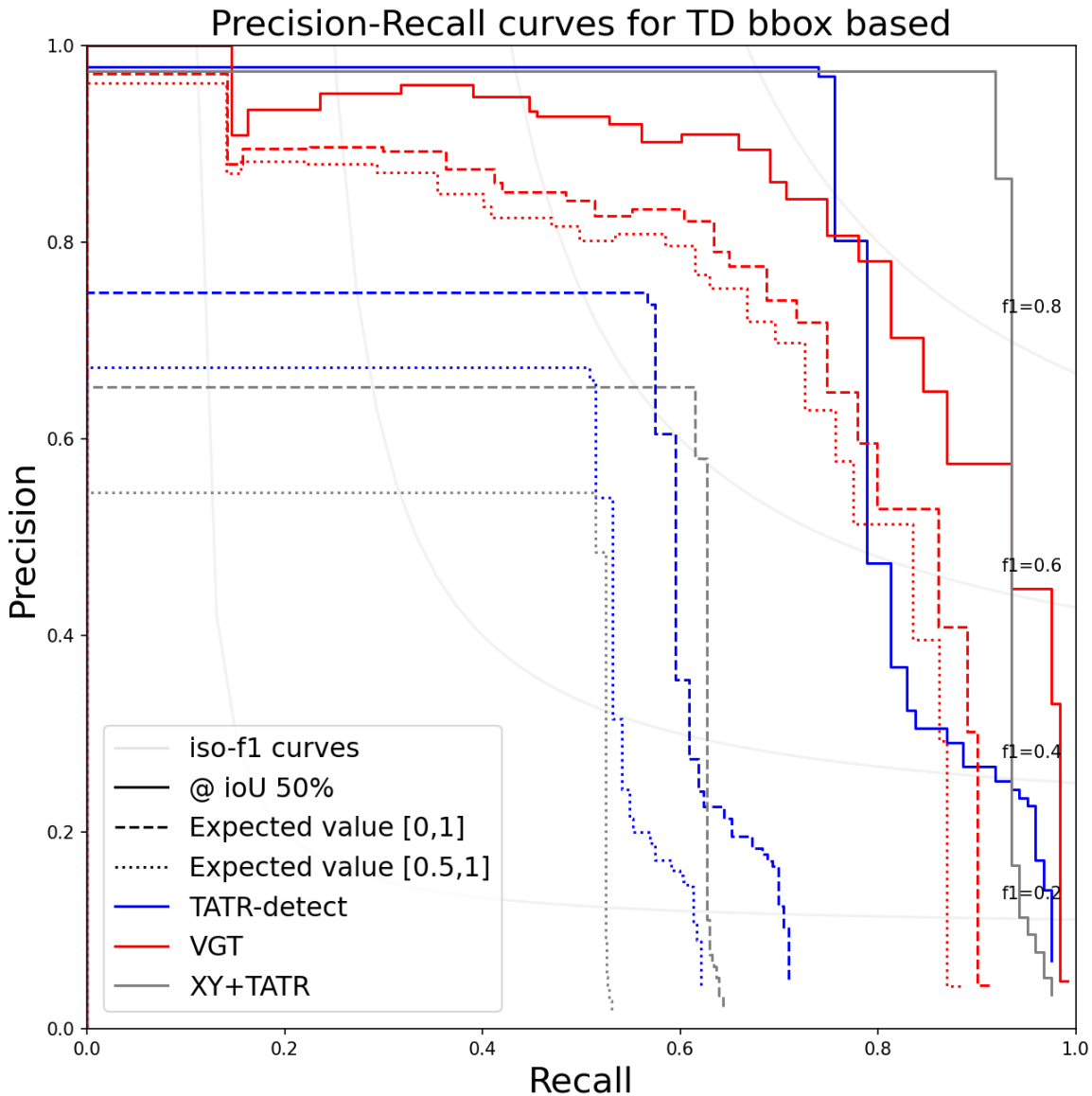
$$\mathcal{P}^+_{\theta_c} := \{\widehat{y} \mid (\widehat{y}, c) \in \mathcal{P},\ c_{\text{table}} > \theta_c\}$$

Ainsi on peut tracer *Precision*, *Rappel*, F_1 en fonction de θ_c

Détection de tableaux (Precision, Rappel, F_1)



Détection de tableaux (Courbes Precision–Rappel)



Exemple : comparaison TATR-detect / VGT

Echantillon	Critères d'exclusion	
SPs3 –12,5 à 15m	$D_{10} > 2mm$	non
	$D_{70} < 74\ \mu m$ et $I_P > 10$	non
	Critères de suspicion	
critères des terrains sableux	Sr voisin de 100 %	oui
	$C_u = \frac{D_{60}}{D_{10}} < 15$	non
	$0,05 < D_{50} < 1,5mm$	non
	$\sigma'_v < 200kPa$	oui
critères des terrains argileux	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

0.365,0.001,0.634

Tableau 33 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 5 issu de SPs3

Echantillon	Critères d'exclusion	
SPs4 – 0 à 3m	$D_{70} < 74\ \mu m$ et $I_P > 10$	non
	Critères de suspicion	
	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

0.838,0. ,0.162

Tableau 34 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 1 issu de SPs4

Echantillon	Critères d'exclusion	
SPs4 – 3 à 6m	$D_{70} < 74\ \mu m$ et $I_P > 10$	oui
	Critères de suspicion	
	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

0.392,0. ,0.608

Tableau 35 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 2 issu de SPs4

0.476,0.004,0.52

Echantillon	Critères d'exclusion	
	$D_{10} > 2mm$	non
	$D_{70} < 74\ \mu m$ et $I_P > 10$	non
	Critères de suspicion	
critères des terrains sableux	Sr voisin de 100 %	oui
	$C_u = \frac{D_{60}}{D_{10}} < 15$	non
	$0,05 < D_{50} < 1,5mm$	non
	$\sigma'_v < 200kPa$	oui
critères des terrains argileux	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

Tableau 36 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 3 issu de SPs4

Table Table (rotated)

(a) TATR-detect (c_{table}, $c_{\text{rot. table}}$, $c_{\emptyset}$)

Evaluation de l'aléa sismique pour 8 sites de futurs établissements scolaires à Mayotte

table 64%

Critères d'exclusion		
	$D_{10} > 2mm$	non
	$D_{70} < 74\ \mu m$ et $I_P > 10$	non
	Critères de suspicion	
critères des terrains sableux	Sr voisin de 100 %	oui
	$C_u = \frac{D_{60}}{D_{10}} < 15$	non
	$0,05 < D_{50} < 1,5mm$	non
	$\sigma'_v < 200kPa$	oui
critères des terrains argileux	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

Tableau 33 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 5 issu de SPs3

table 83%

Echantillon	Critères d'exclusion	
	$D_{70} < 74\ \mu m$ et $I_P > 10$	non
	Critères de suspicion	
	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

Tableau 34 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 1 issu de SPs4

table 6%

table 83%

Echantillon	Critères d'exclusion	
	$D_{70} < 74\ \mu m$ et $I_P > 10$	oui
	Critères de suspicion	
	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

Tableau 35 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 2 issu de SPs4

table 87%

Echantillon	Critères d'exclusion	
	$D_{10} > 2mm$	non
	$D_{70} < 74\ \mu m$ et $I_P > 10$	non
	Critères de suspicion	
critères des terrains sableux	Sr voisin de 100 %	oui
	$C_u = \frac{D_{60}}{D_{10}} < 15$	non
	$0,05 < D_{50} < 1,5mm$	non
	$\sigma'_v < 200kPa$	oui
critères des terrains argileux	$D_{15} > 5\ \mu m$	non
	$w_L < 35\ \%$	non
	$w > 0,9 \cdot w_L$	non
	Sol au-dessus de la droite A du diagramme de plasticité	non

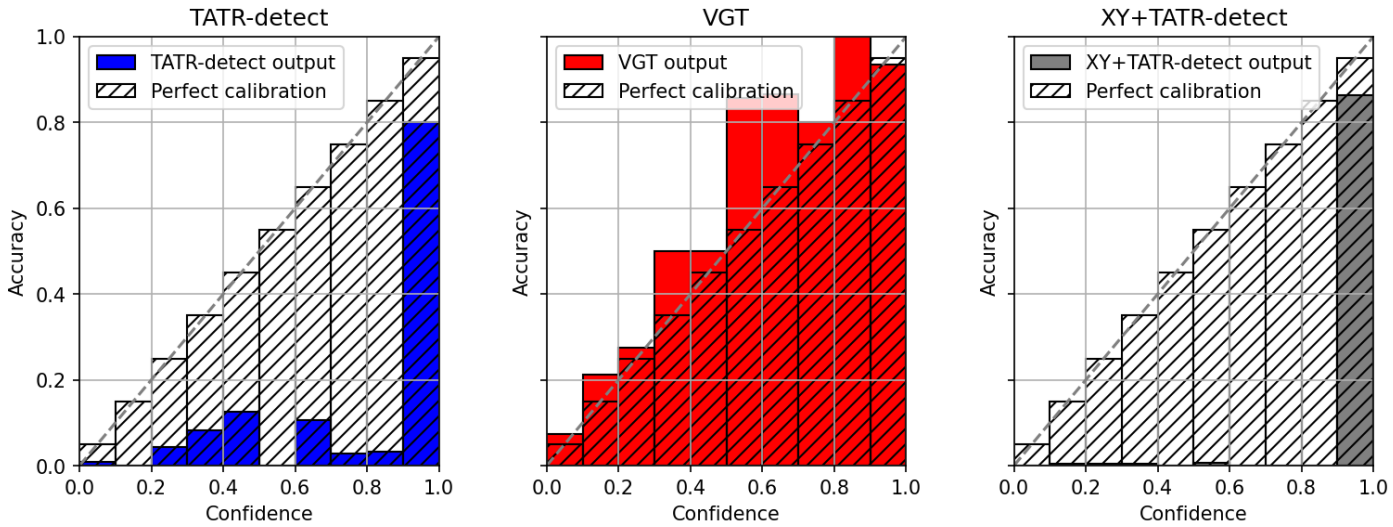
Tableau 36 - Critères de susceptibilité à la liquéfaction de l'échantillon n° 3 issu de SPs4

BRGM/RP-61170-FR – Rapport final. 71

(b) VGT c_{table}

Fiabilité des scores de confiance

Faut-il faire confiance aux scores de confiance des modèles de vision ?



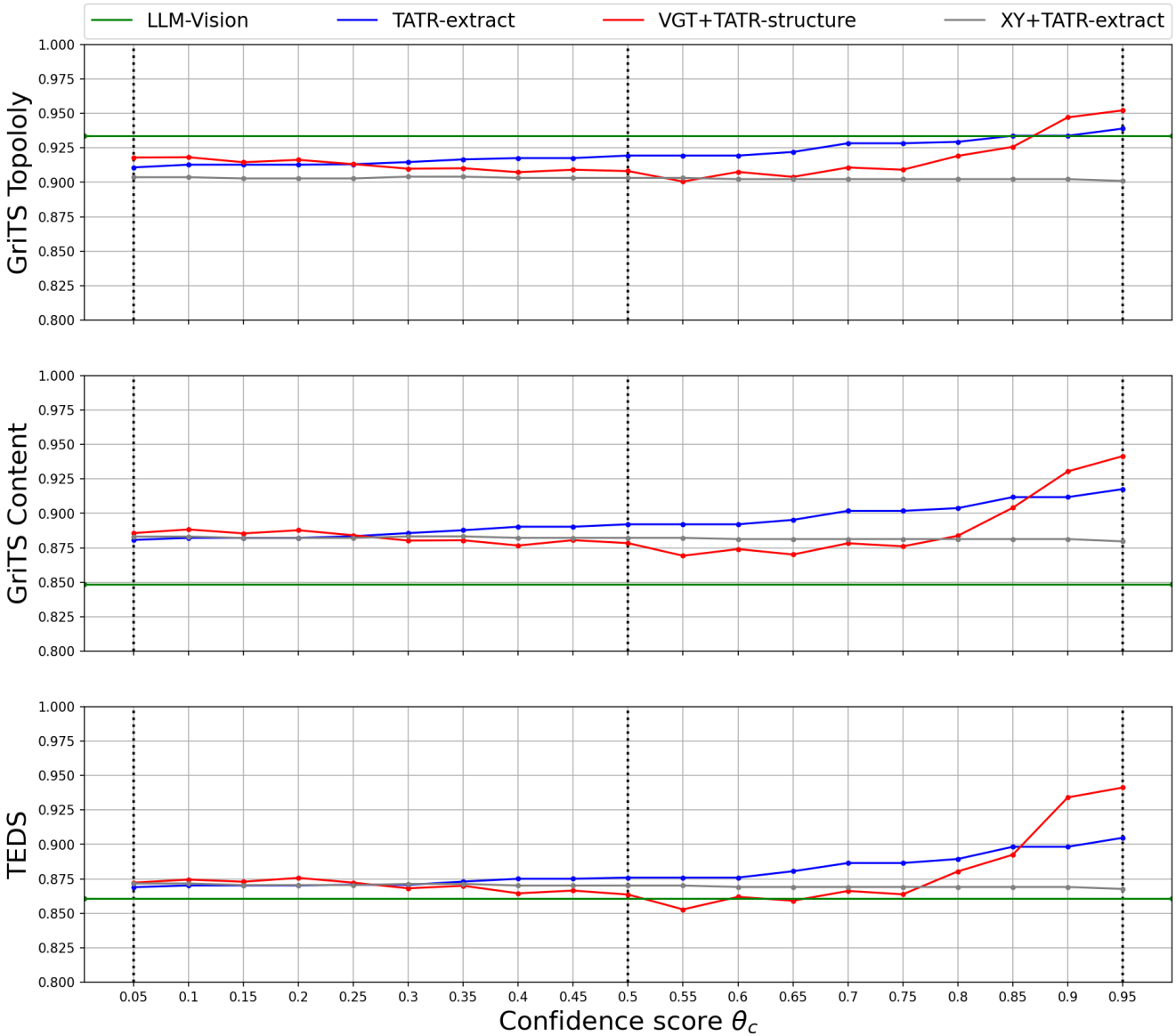
Diagrammes de fiabilité (Niculescu-Mizil & Caruana, 2005) des modèles

Plan

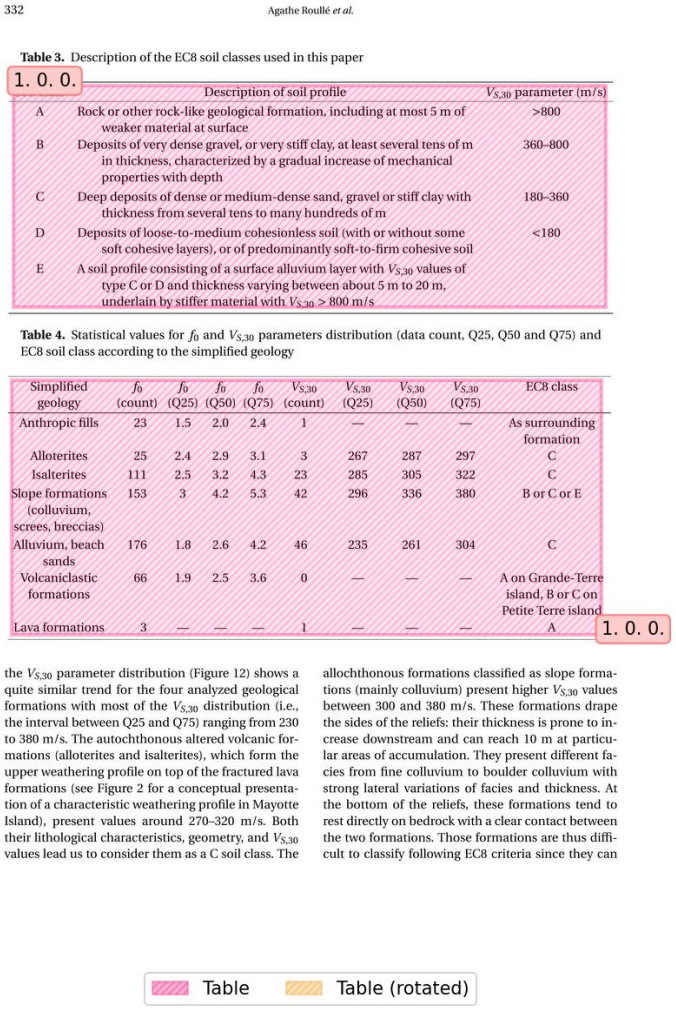
Sur quelles prédictions évaluer la RST ?

- La phase de reconnaissance de structure dépend de la phase de détection pour une evaluation des méthodes d'extraction dans leur ensemble.
- Choix du calcul des score de similarités à partir de l'ensemble des *True Positive* : couple (tableau prédit, tableau de référence).

Reconnaissance de structure (GriTS Top, Cont, TEDS)



Exemple d'extraction avec TATR-extract



332 Agathe Roullé *et al.*

Table 3. Description of the EC8 soil classes used in this paper

	Description of soil profile	$V_{S,30}$ parameter (m/s)
A	Rock or other rock-like geological formation, including at most 5 m of weaker material at surface	>800
B	Deposits of very dense gravel, or very stiff clay, at least several tens of m in thickness, characterized by a gradual increase of mechanical properties with depth	360–800
C	Deep deposits of dense or medium-dense sand, gravel or stiff clay with thickness from several tens to many hundreds of m	180–360
D	Deposits of loose-to-medium cohesionless soil (with or without some soft cohesive layers), or of predominantly soft-to-firm cohesive soil	<180
E	A soil profile consisting of a surface alluvium layer with $V_{S,30}$ values of type C or D and thickness varying between about 5 m to 20 m, underlain by stiffer material with $V_{S,30}$ > 800 m/s	

1. 0. 0.

Table 4. Statistical values for f_0 and $V_{S,30}$ parameters distribution (data count, Q25, Q50 and Q75) and EC8 soil class according to the simplified geology

Simplified geology	f_0 (count)	f_0 (Q25)	f_0 (Q50)	f_0 (Q75)	$V_{S,30}$ (count)	$V_{S,30}$ (Q25)	$V_{S,30}$ (Q50)	$V_{S,30}$ (Q75)	EC8 class
Anthropic fills	23	1.5	2.0	2.4	1	—	—	—	As surrounding formation
Alloterites	25	2.4	2.9	3.1	3	267	287	297	C
Isalterites	111	2.5	3.2	4.3	23	285	305	322	C
Slope formations (colluvium, screes, breccias)	153	3	4.2	5.3	42	296	336	380	B or C or E
Alluvium, beach sands	176	1.8	2.6	4.2	46	235	261	304	C
Volcaniclastic formations	66	1.9	2.5	3.6	0	—	—	—	A on Grande-Terre island, B or C on Petite Terre island
Lava formations	3	—	—	—	1	—	—	—	A

1. 0. 0.

the $V_{S,30}$ parameter distribution (Figure 12) shows a quite similar trend for the four analyzed geological formations with most of the $V_{S,30}$ distribution (i.e., the interval between Q25 and Q75) ranging from 230 to 380 m/s. The autochthonous altered volcanic formations (alloterites and isalterites), which form the upper weathering profile on top of the fractured lava formations (see Figure 2 for a conceptual presentation of a characteristic weathering profile in Mayotte Island), present values around 270–320 m/s. Both their lithological characteristics, geometry, and $V_{S,30}$ values lead us to consider them as a C soil class. The allochthonous formations classified as slope formations (mainly colluvium) present higher $V_{S,30}$ values between 300 and 380 m/s. These formations drape the sides of the reliefs: their thickness is prone to increase downstream and can reach 10 m at particular areas of accumulation. They present different facies from fine colluvium to boulder colluvium with strong lateral variations of facies and thickness. At the bottom of the reliefs, these formations tend to rest directly on bedrock with a clear contact between the two formations. Those formations are thus difficult to classify following EC8 criteria since they can

Table Table (rotated)

(a) TATR-detect

Table 3. Description of the EC8 soil classes used in this paper

Soil class	Description of soil profile	$V_{S,30}$ parameter (m/s)
A	Rock or other rock-like geological formation, including at most 5 m of weaker material at surface	>800
B	Deposits of very dense gravel, or very stiff clay, at least several tens of m in thickness, characterized by a gradual increase of mechanical properties with depth	360–800
C	Deep deposits of dense or medium-dense sand, gravel or stiff clay with thickness from several tens to many hundreds of m	180–360
D	Deposits of loose-to-medium cohesionless soil (with or without some soft cohesive layers), or of predominantly soft-to-firm cohesive soil	<180
E	A soil profile consisting of a surface alluvium layer with $V_{S,30}$ values of type C or D and thickness varying between about 5 m to 20 m, underlain by stiffer material with $V_{S,30}$ > 800 m/s	

Table 4. Statistical values for f_0 and $V_{S,30}$ parameters distribution (data count, Q25, Q50 and Q75) and

Data cell Column header cell Projected row header cell

(b) TATR-structure

Soil class	Description of soil profile	V S,30 parameter (m/s)
A	Rock or other rock-like geological formation, including at most 5 m of weaker material at surface	>800
B	Deposits of very dense gravel, or very stiff clay, at least several tens of m in thickness, characterized by a gradual increase of mechanical properties with depth	360–800
C	Deep deposits of dense or medium-dense sand, gravel or stiff clay with thickness from several tens to many hundreds of m	180–360
D	Deposits of loose-to-medium cohesionless soil (with or without some soft cohesive layers), or of predominantly soft-to-firm cohesive soil	<180
E	A soil profile consisting of a surface alluvium layer with V S,30 values of type C or D and thickness varying between about 5 m to 20 m, underlain by stiffer material with V S,30 > 800 m/s	

(c) Tableau extrait (HTML)

Conclusion

- Utilisation de méthodes d'extraction bout-en-bout pour le BRGM.
- Amélioration des méthodes existantes via modification des traitements de données.
- Résultats globalement bons sur le jeu de données (F_1-score, métriques RST).
- Nécessité un choix d'utilisation :
  - Utiliser un seuil θ_c pour constituer un ensemble de positive.
  - Faire confiance au score de confiances.

La suite / Perspectives

Le stage :

- Exploiter les tableaux extraits via des requêtes.
- Sémantiser, enrichir les tableaux par leur contexte (légende).
- Jeu de données plus grand.

La thèse :

- S'intéresser à d'autres types de données (texte, schéma).
- Exploiter l'hétérogénéité des données avec des méthodes multi-modales.
- Situer la connaissance dans ses dimensions spatiales et temporelles.


Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. https://arxiv.org/abs/2005.12872


Da, C., Luo, C., Zheng, Q., & Yao, C. (2023). Vision Grid Transformer for Document Layout Analysis. https://arxiv.org/abs/2308.14978


Ha, J., Haralick, R., & Phillips, I. (1995).Recursive X-Y cut using bounding boxes of connected components. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *2*, 952-955 vol.2. https://doi.org/10.1109/ICDAR.1995.602059


Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., & Li, Z. (2020, mai). TableBank : Table Benchmark for Image-based Table Detection and Recognition. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Éd.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (p. 1918-1925). European Language Resources Association. https://aclanthology.org/2020.lrec-1.236/


Lopez, P. (2008). GROBID.


Niculescu-Mizil, A., & Caruana, R. (2005).Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625-632. https://doi.org/10.1145/1102351.1102430

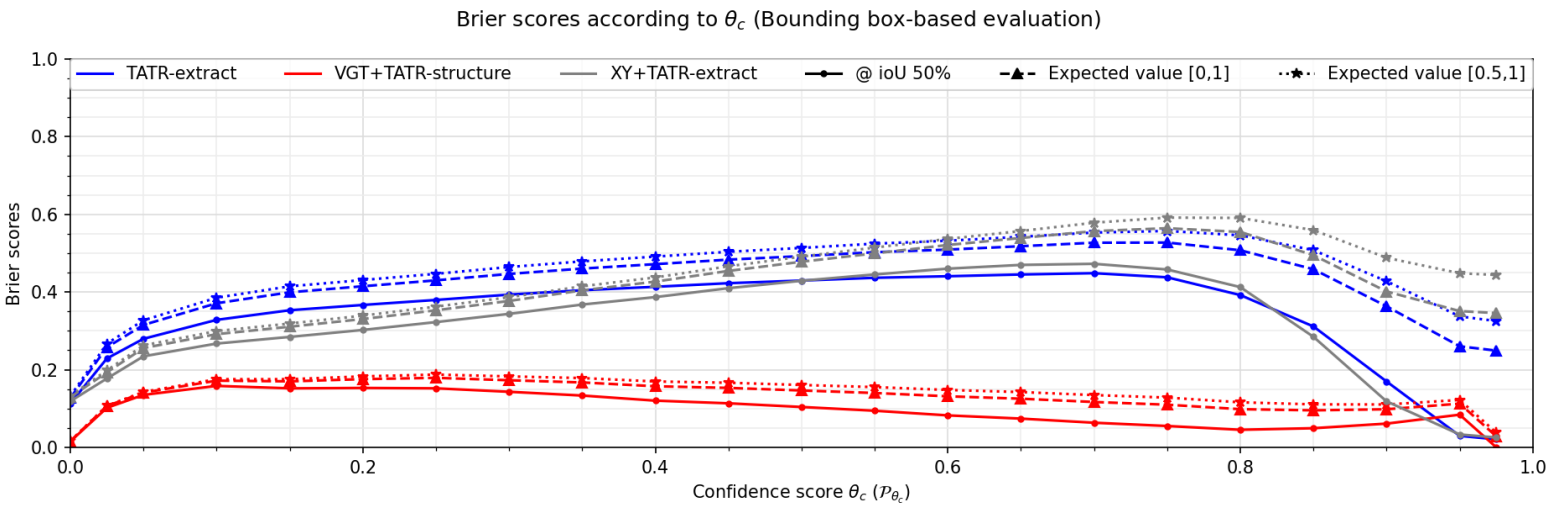
 Smock, B., Pesala, R., & Abraham, R. (2022).PubTables-1M : Towards Comprehensive Table Extraction From Unstructured Documents. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4634-4642.

 Smock, B., Pesala, R., & Abraham, R. (2023). GriTS : Grid table similarity metric for table structure recognition. https://arxiv.org/abs/2203.12555

Annexes

Brier score

Calcule d'une *accuracy* : proportion de prédictions correctement classifiées à partir des scores de confiance. Brier score (1D) est la moyenne quadratique des erreurs.

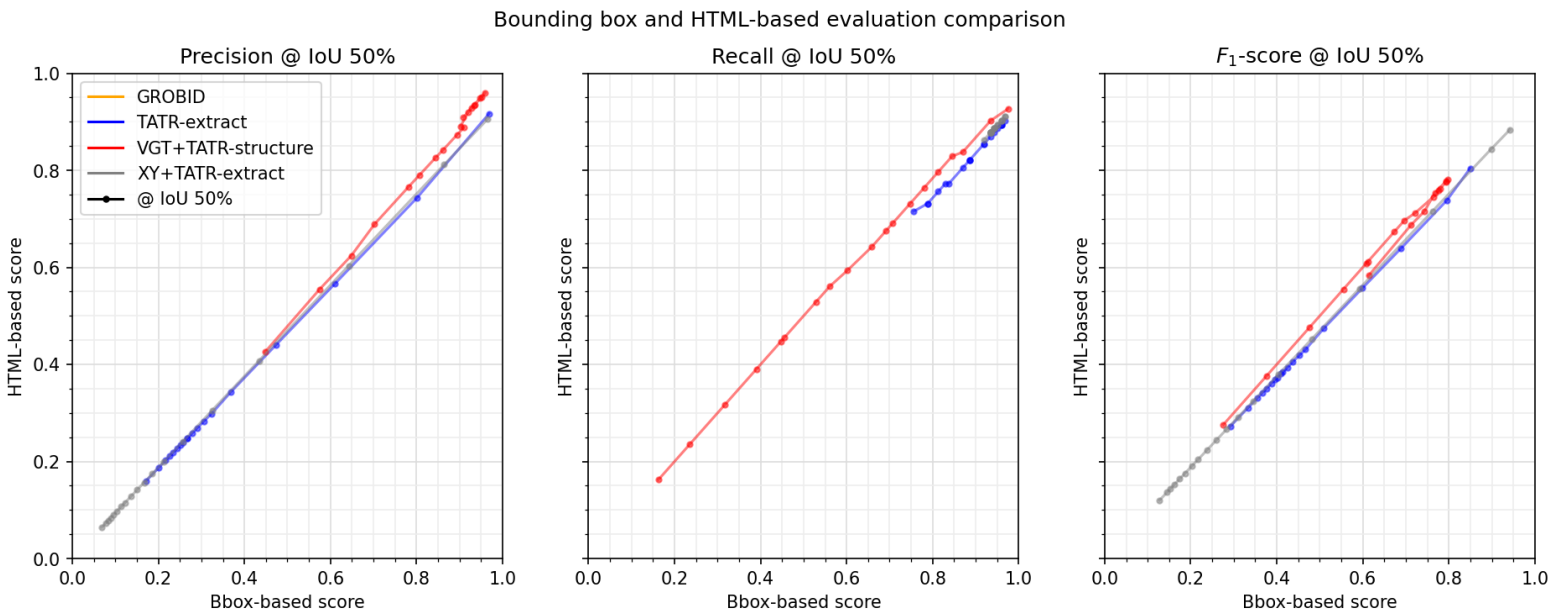


Algorithme X-Y cut

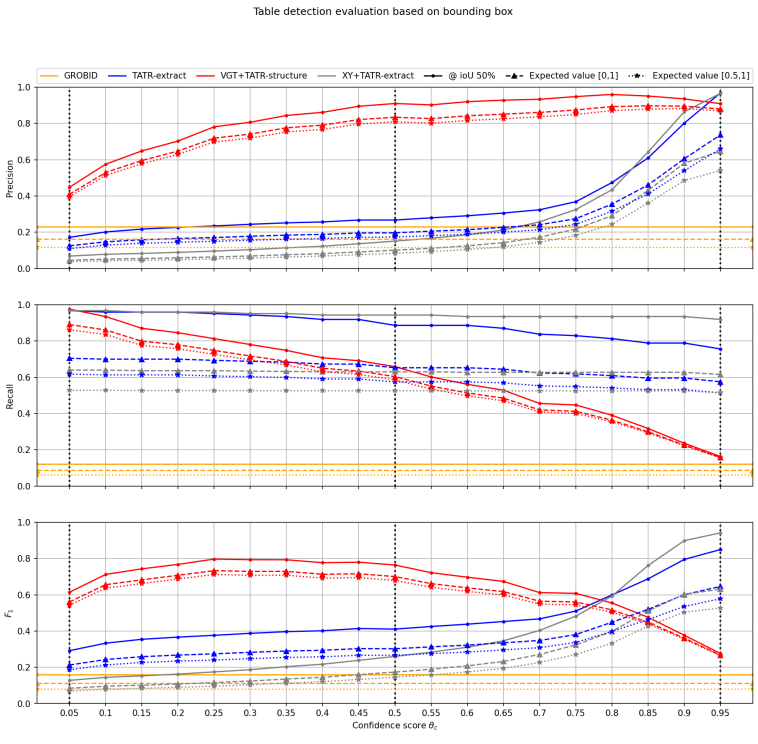
1. Projeter les pixels noirs selon les axes X et Y de l'image (ou sous-image).
2. À partir du profil, repérer les espacements entre paragraphes.
3. Créer des sous-images.
4. Répéter les étapes 1 à 3 jusqu'à la rencontre d'un critère d'arrêt.

Détection de tableaux avec HTML et coordonnées de tableaux (1/2)

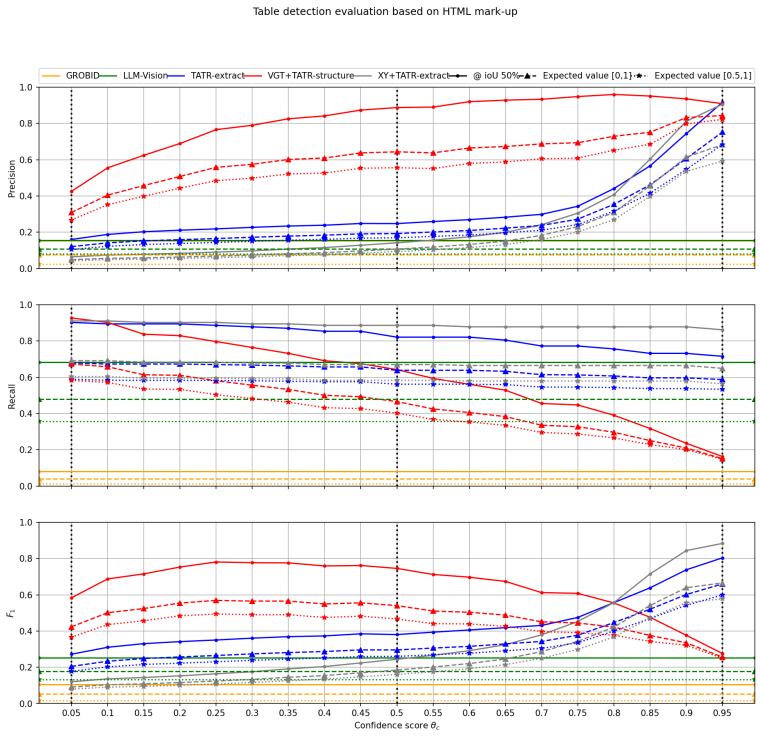
Représentation des tableaux par des 2-grammes multi ensembles (avec chaîne de deux caractères qui ne sont pas des balises).



Détection de tableaux avec HTML et coordonnées de tableaux (2/2)



(a) DT avec coordonnées



(b) DT avec tableaux en HTML