Journée Scientifique et Technique du CFMS du 29 janvier 2020

« Machine Learning et Big Data en Géotechnique»

Utilisation d’un algorithme de classification par machine learning pour la caractérisation géomécanique des sols

Marie-Cécile Febvey

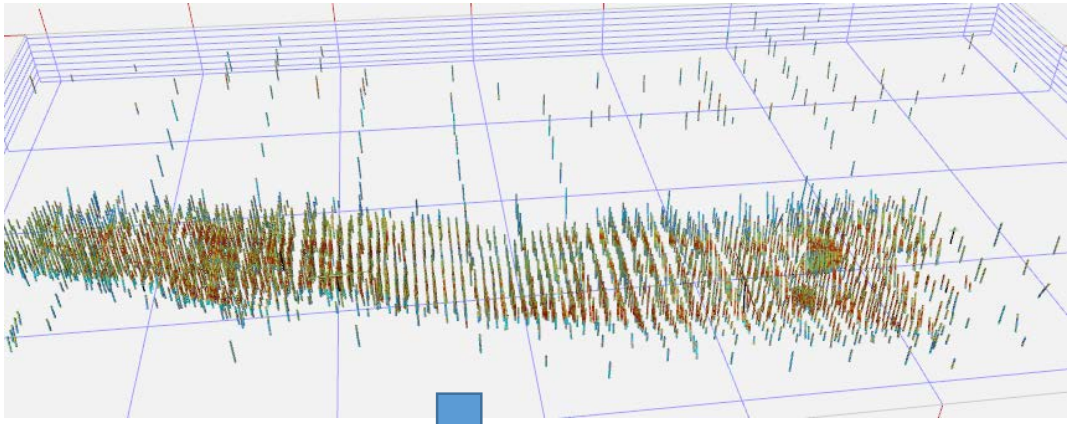
Geovariances

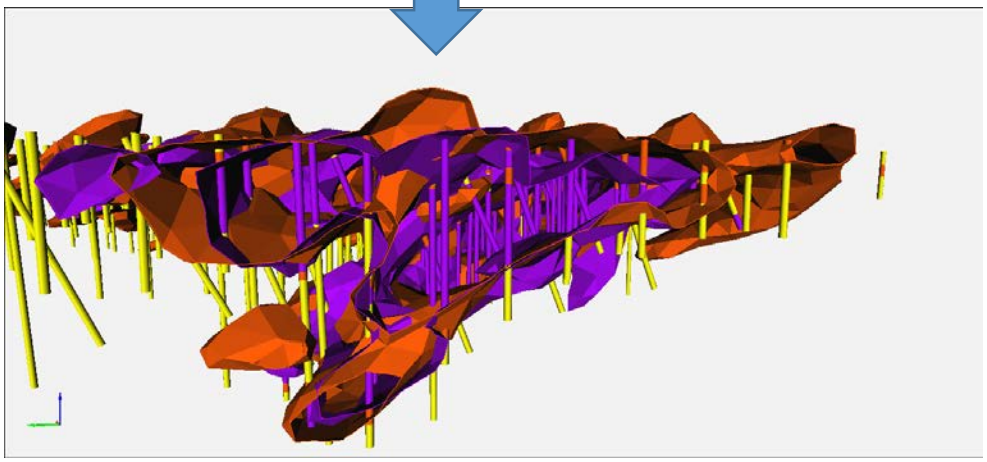
Where no one has gone before

Objectifs

- Données de plus en plus abondantes et difficiles à traiter
- Besoin de méthodes innovantes pour analyser les données et les regrouper
- Regroupement= cluster, première étape avant une modélisation du sous-sol

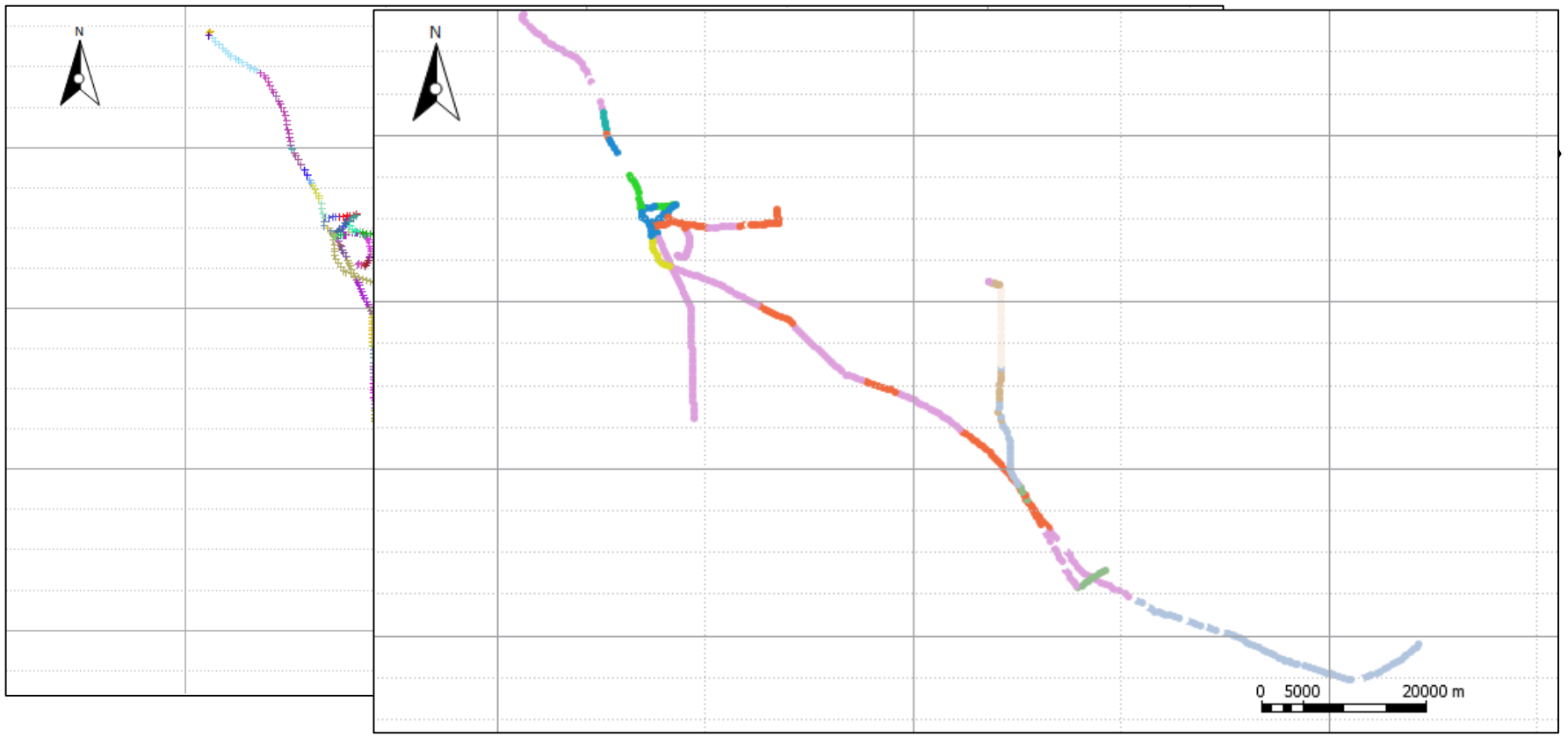
Objectifs





Variable
Teneur 1
Teneur 2
Teneur 3
Mineral 1
Mineral 2
Altération

Objectifs



Classification d'échantillons

- Basée sur la méthode de **classification (ou clustering) géostatistique hiérarchique (GHC);**

- Méthode d'apprentissage non supervisée utilisée dans des méthodes de **Machine Learning**

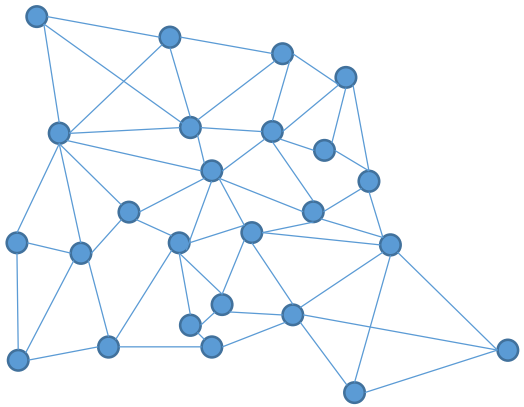
- Méthode iterative qui aggrège les échantillons et les "groupes" (cluster) entre eux en fonction de la distance de corrélation qui décrit leur similarités;

Classification d'échantillons

- Pour aggréger deux groupes ensemble, leur distance est comparée dans une matrice de **dissimilarités** qui contient toutes les distances possibles entre les différents groupes. La paire de groupe avec la plus petite distance est regroupée

- **Répétabilité, traçabilité, objectivité.**

Classification d'échantillons



1. Création d'un réseau
2. Calcul des dissimilarités entre les échantillons groupés
3. Regroupement progressif des échantillons/groupes les moins dissemblables

Paramètres

- Plusieurs variables peuvent être considérées en entrée (coordonnées, variables catégorielles etc)

- Chaque variable a un poids assigné afin de définir son importance dans la définition du groupe (cluster)

Fonction de dissimilarité

- Distance euclidienne pondérée

$$d_{i,j} = \sum_{v}^{Nv} (\theta^{v}{}_{i,j} \times w^{v}) + \theta^{c}{}_{i,j} \times w^{c}$$

$\theta^{v}{}_{i,j} = Contribution\ de\ la\ variable\ v\ pour\ la\ paire\ de\ variables\ i,j$
$\theta^{c}{}_{i,j} = Contribution\ des\ coordonnées\ pour\ la\ paire\ d'échantillons\ i,j$
$w^{v} = Poids\ assigné\ à\ la\ variable\ v$
$w^{c} = Poids\ assigné\ aux\ coordonnées$
$Nv = Nombre\ de\ variables\ (sans\ les\ coordonnées)$
$\boldsymbol{d_{i,j} = Dissimilarités\ entre\ les\ paires\ d'échantillons\ i,j}$

Variable continue: distance entre deux échantillons/groupes (cluster)

- La dissimilarité entre deux échantillons (Z_i,Z_j) d'une variable continue est :

$$\theta^v{}_{i,j} = (Z^n{}_i - Z^n{}_j)^2$$

$$Z^n{}_i = \frac{Z_i - m}{\sigma}$$

$m = Moyenne$
$\sigma = Ecart\ Type$
$Z_i = valeur\ de\ la\ variable\ i$
$Z^n{}_i = Z_i\ normalisé$

Variable discrète: distance entre deux échantillons/groupes (cluster)

- Coefficient de dissimilarité pour chaque variable catégorielle (valeur de la variable discrète)

- Pour les variables discrètes: matrice de dissimilarités pour chaque variable catégorielle

$$Md = \begin{pmatrix} \alpha_{AA} & \alpha_{AB} & \alpha_{AC} \\ & \alpha_{BB} & \alpha_{BC} \\ & & \alpha_{CC} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 5 \\ 1 & 0 & 10 \\ 5 & 10 & 0 \end{pmatrix}$$

Les coefficients de la diagonale sont zéros;
Si $\alpha_{AB} > \alpha_{AC}$ et $\alpha_{AB} / \alpha_{AC} = 5$, cela signifie que A est 5 fois plus dissimilaire de B qu'il ne l'est de C.

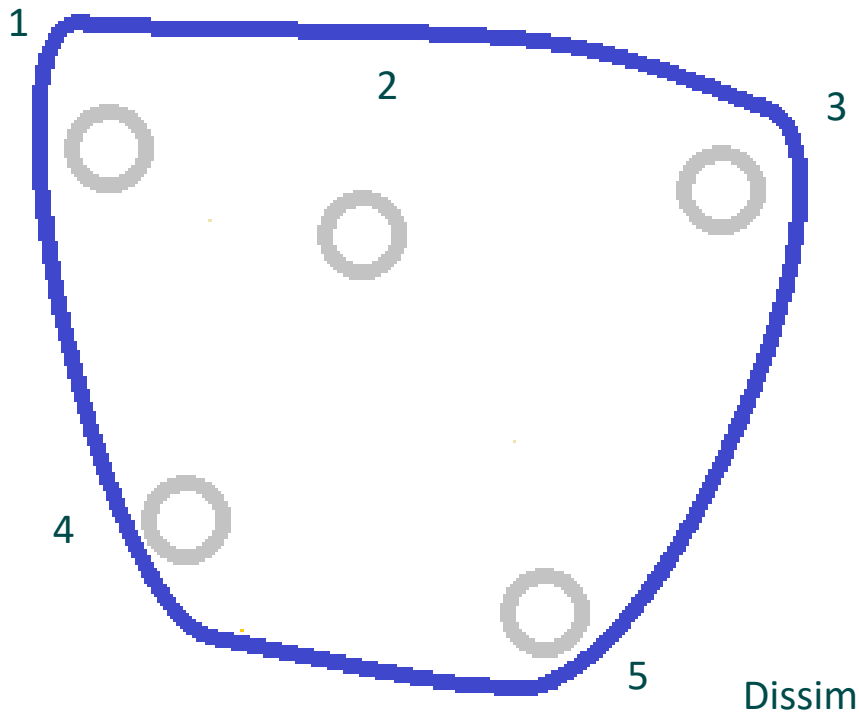
Distance entre deux groupes

- En multivariables, toutes les distances sont combinées en utilisant une somme pondérée

- $D^2_{Multi}(i,j) = \sum_1^n w_{Z_n} D^2_{Z_n}(i,j) + \sum_1^m w_{Cat_m} D^2_{Cat_m}(i,j) + w_{Coor} D^2_{Coor}(i,j)$
- Le poids affecté aux coordonnées est déduit des proportions données en entrée

	1	2	3	4	5
1	0	4.5	2.5	2	3
2		0	4	5	6
3			0	3	1.5
4				0	3.5
5					0

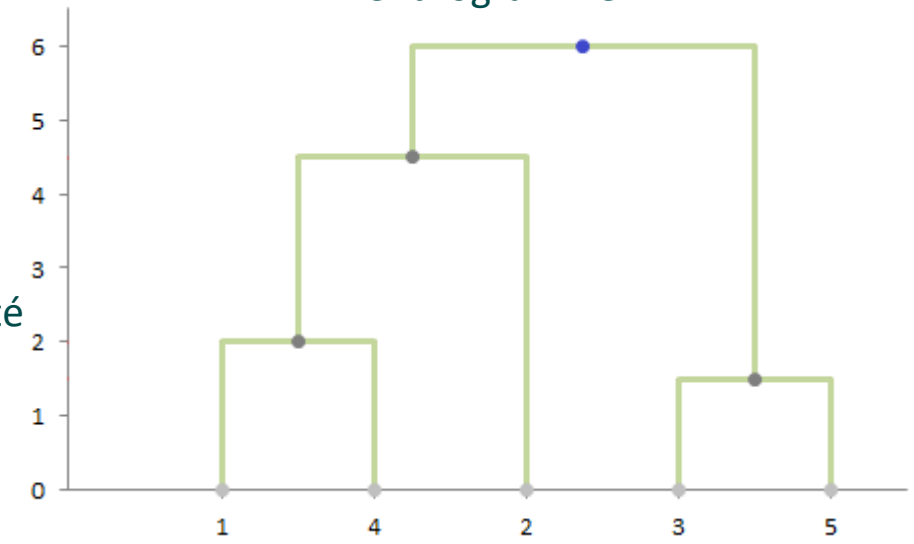
Quelles règles?



Matrice de dissimilarité

	1-2-3-4-5
1-2-3-4-5	0

Dendrogramme



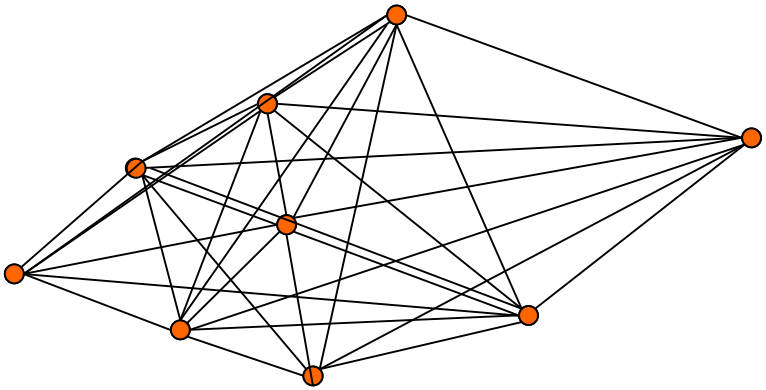
Graphique de connectivité

- Lors de la première itération, il y autant de groupes que d'échantillons. Regarder toutes les paires demande beaucoup de temps;

- Un sous-ensemble de paires peut être considéré dans un voisinage donné

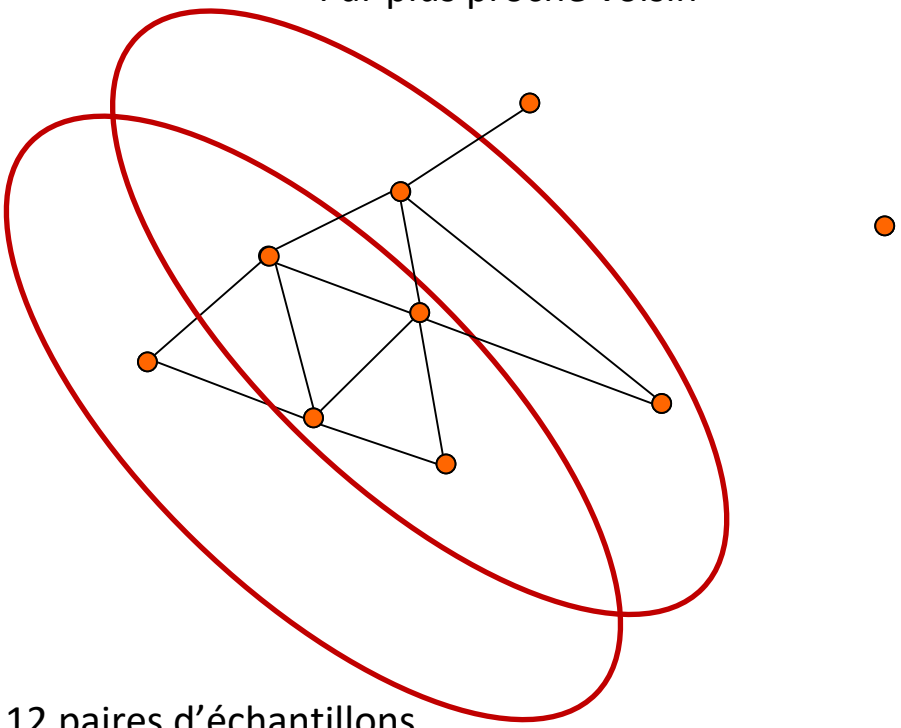
Graphique de connectivité

Cas Exhaustif



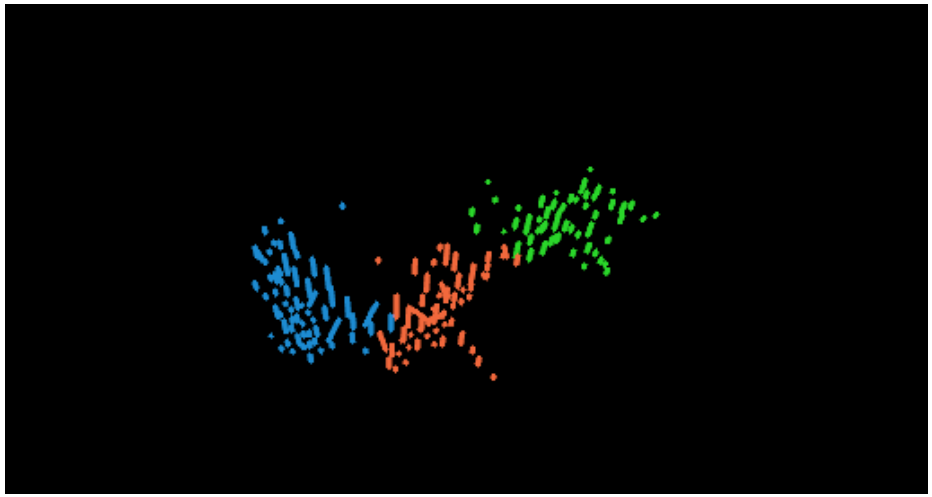
36 paires d'échantillons

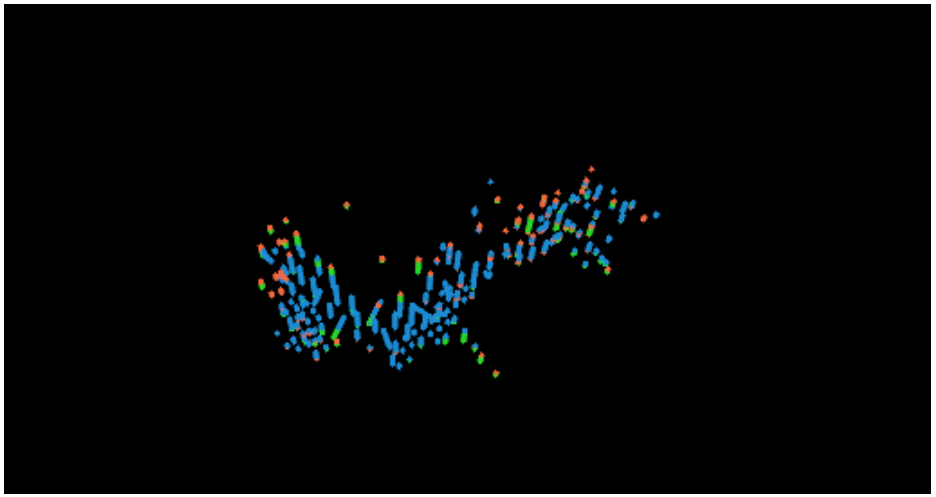
Par plus proche voisin



12 paires d'échantillons

Paramètres





- Poids coordonnées: 100%
- Poids teneur: 0%

- Poids coordonnées : 0%
- Poids teneur: 100%

Mise à jour – Big Data

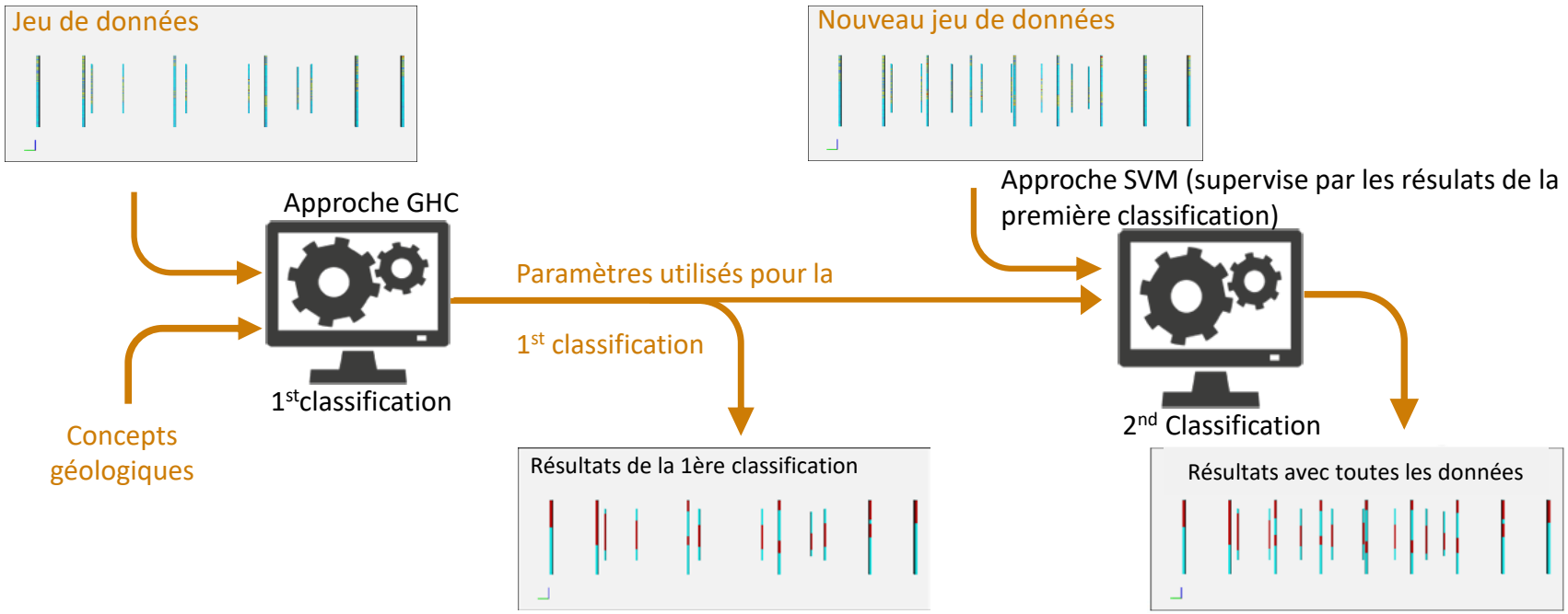
- Lorsqu'on ajoute des données, il est possible de les grouper en comparant le nouveau regroupement avec l'ancien

- Pour définir un groupe, il n'est pas nécessaire de garder l'ensemble des échantillons mais seulement une petite partie de ceux-ci

- L'algorithme de classification peut ainsi être amélioré en ne considérant qu'un sous ensemble d'échantillon, en constituant un regroupement sur ce sous-ensemble et en l'appliquant dans un second temps sur l'ensemble des échantillons

Mise à jour – Big Data

Combinaison de deux algorithmes de classification: Geostatistical Hierarchical Clustering (GHC) et Support Vector Machine (SVM).

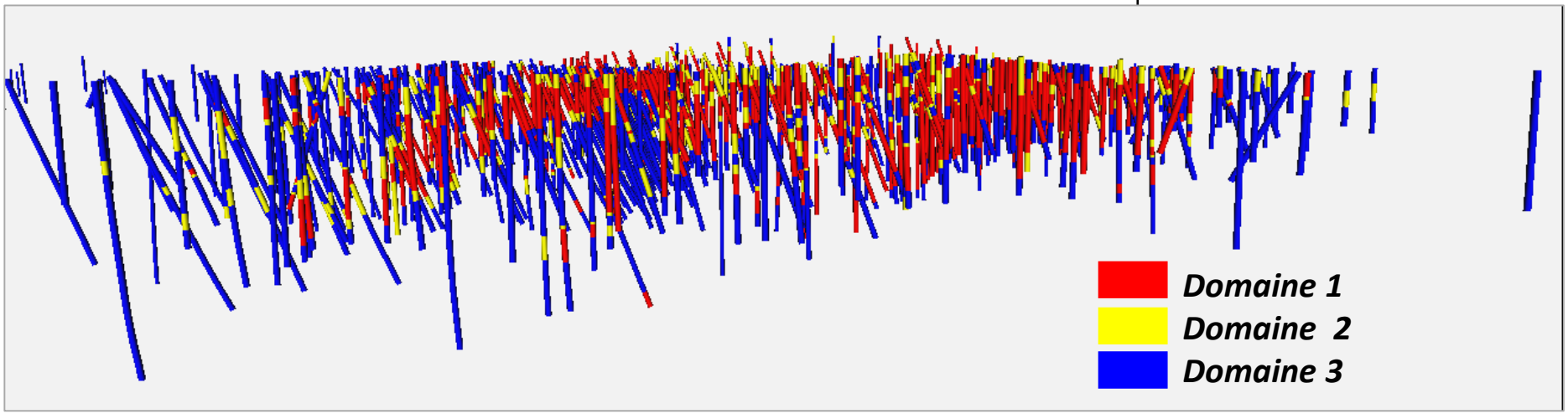


Exemple minier

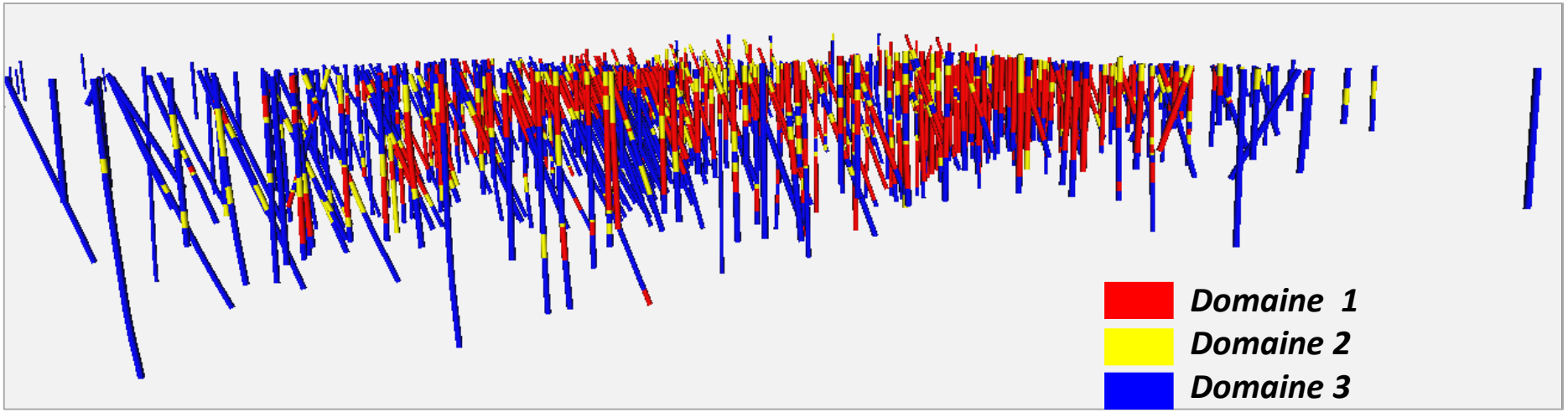
- Gisement multi éléments

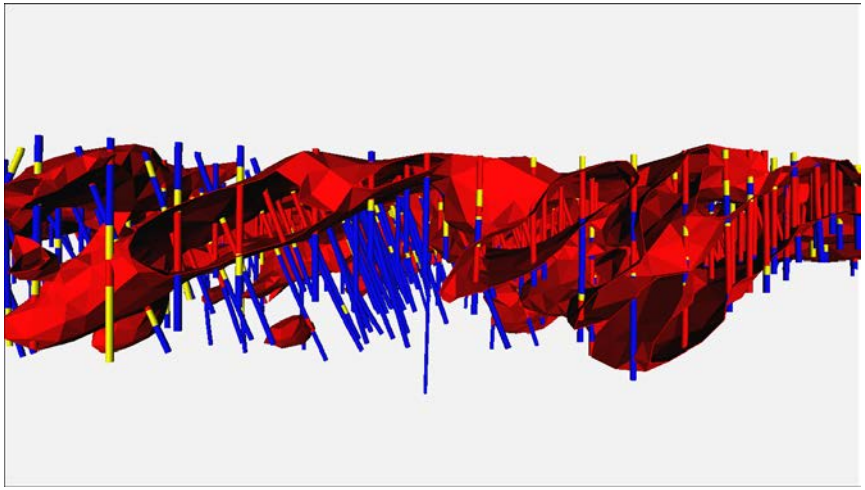
- 2 000 forages & 100.000 échantillons;
  - 10 teneurs; 15 mesures spectrales;
  - 1 variable catégorielle

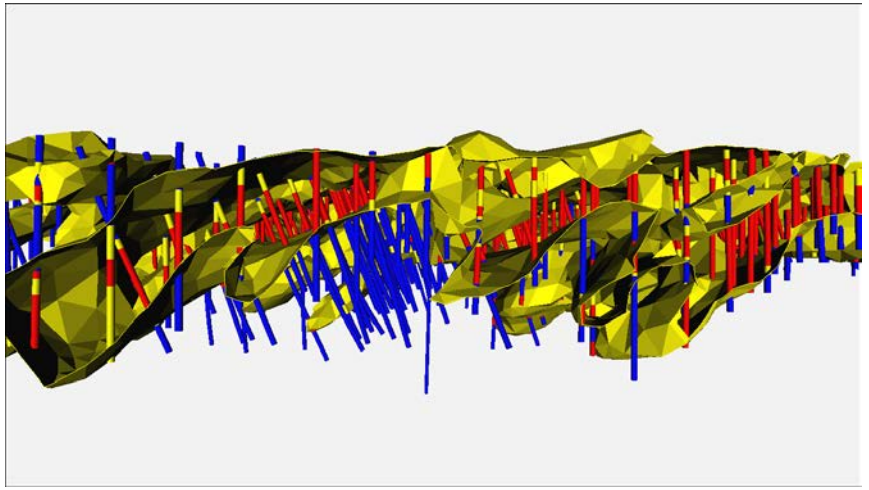
Variable	Weight
Teneur 1	10
Teneur 2	5
Teneur 3	1
Mineral 1	1
Mineral 2	1
Altération	1



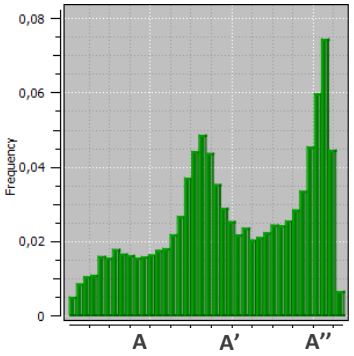
Exemple minier







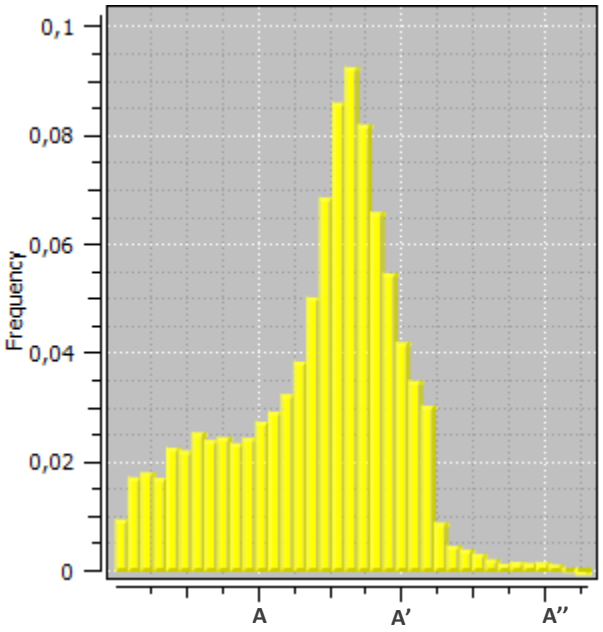
Première campagne



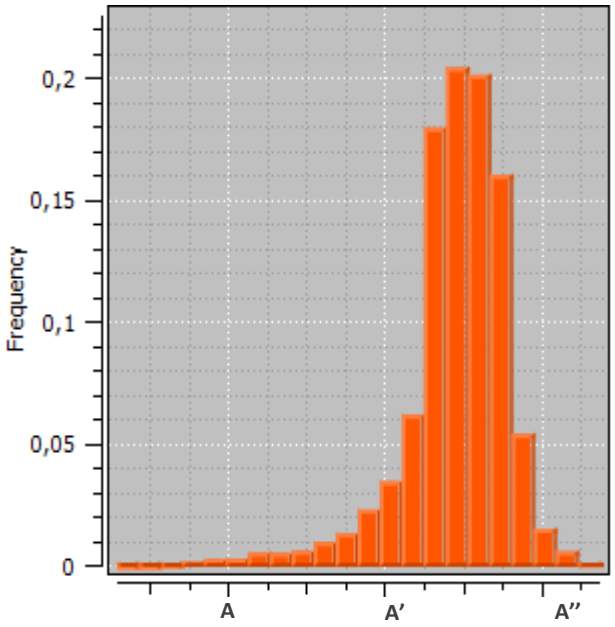
Distribution globale

Résultats:

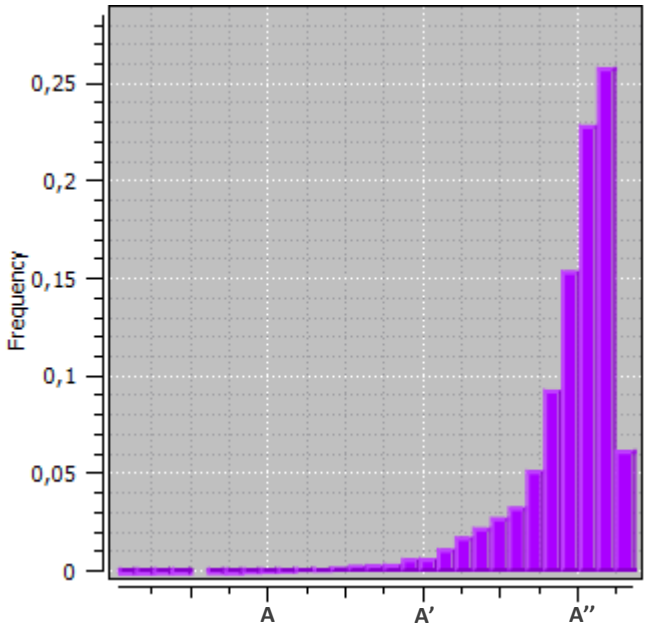
Domaine 3



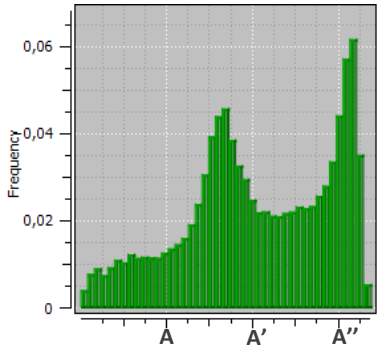
Domaine 2



Domaine 3



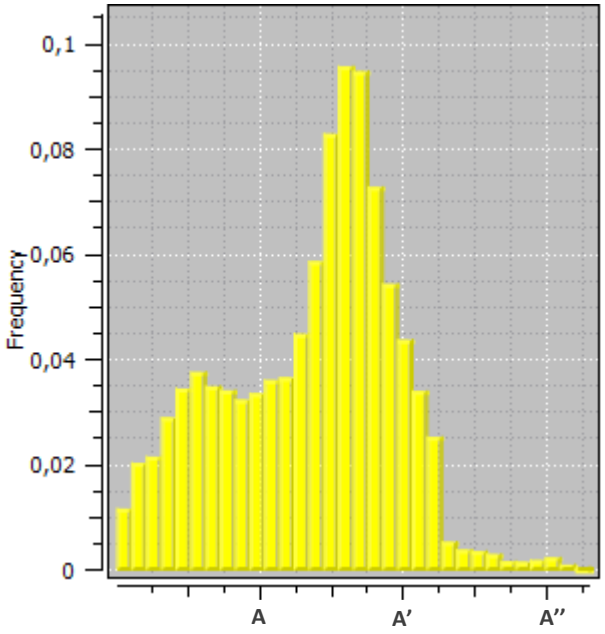
Deux campagnes



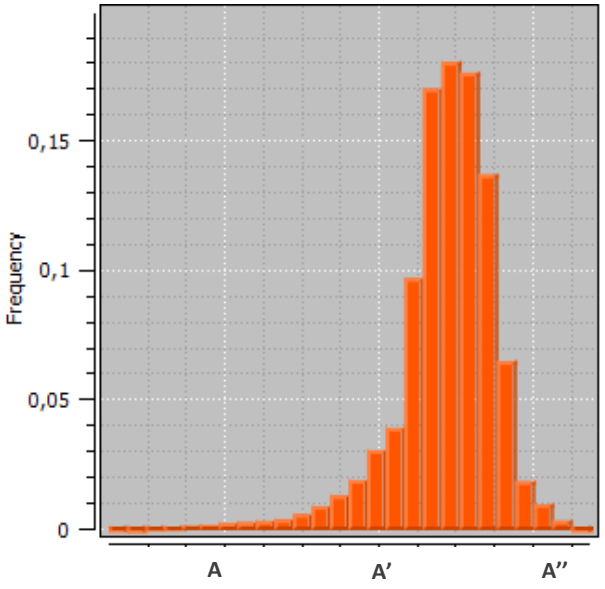
Distribution globale

Résultats:

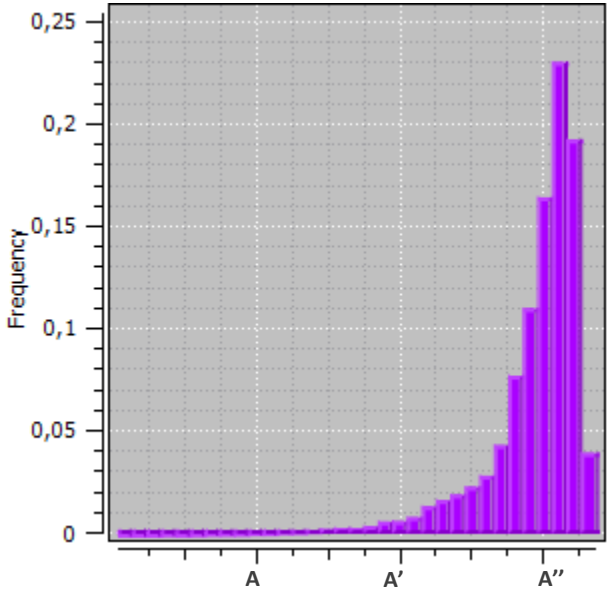
Domaine 3



Domaine 2

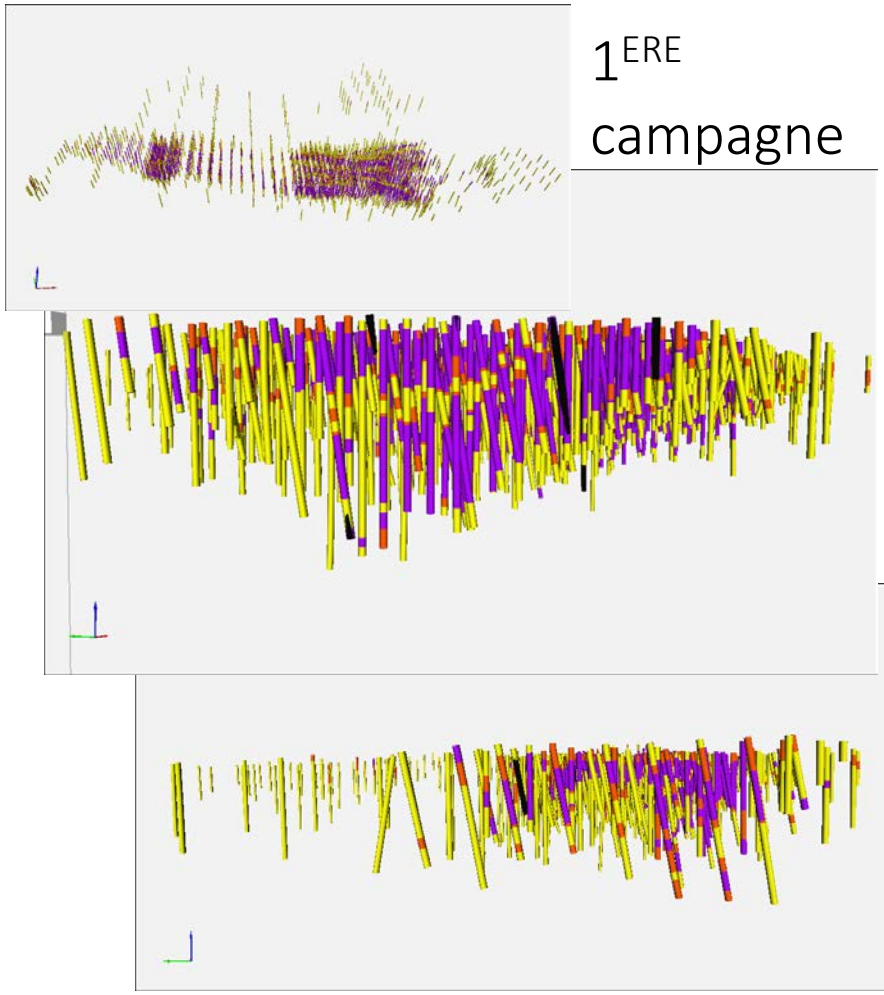


Domaine 1

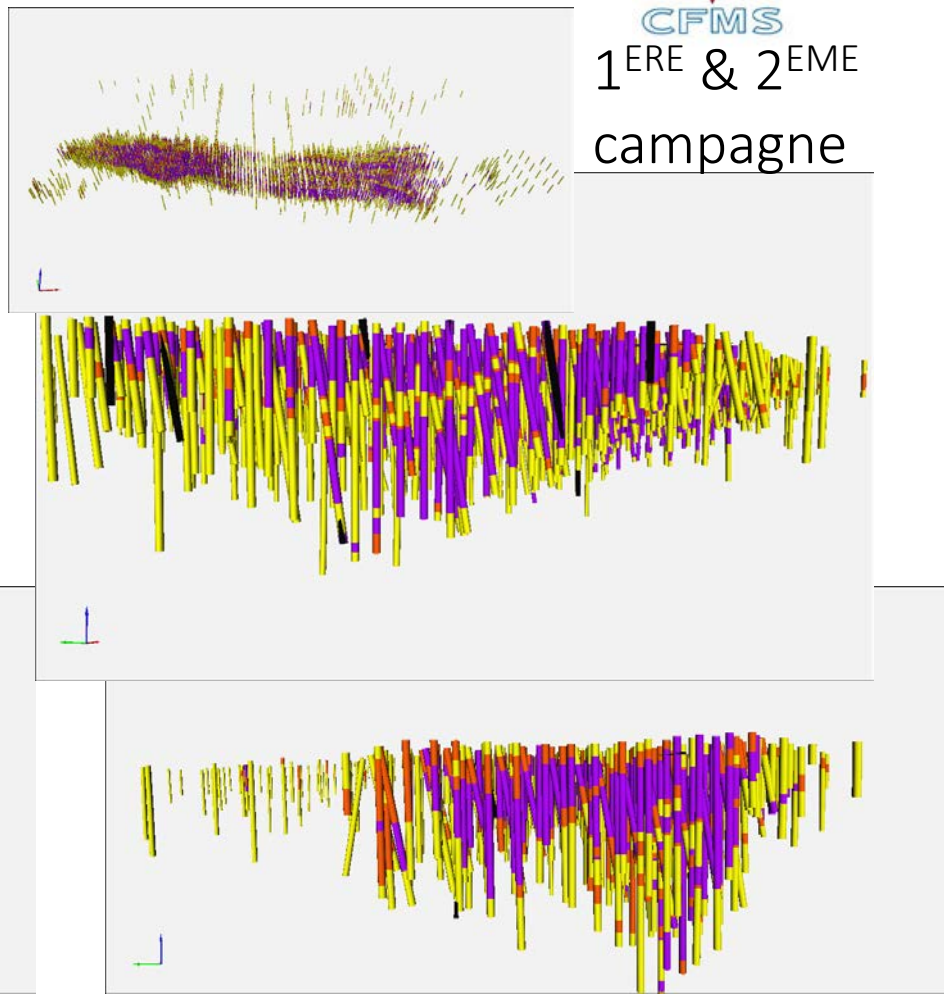


COMPARISON RESULTATS

1ERE campagne

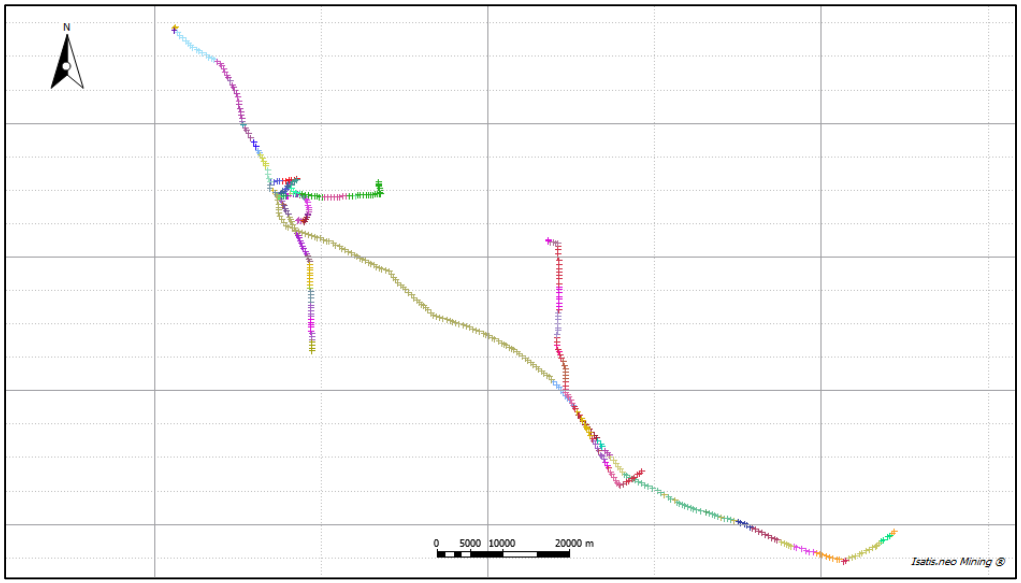


1ERE & 2EME campagne



Exemple: classification d'échantillons de sol

- 546 échantillons
- 57 lithologies différentes
- 11 « groupes » lithologiques
- 5 variables d'intérêt



geotech
Type de sol

- Silt sableux
- Sable limoneux et argileux
- Sable lgrt argileux
- Tourbe et sable lâche
- Limon argilo-sableux
- Limon et silt
- Sable limoneux
- Grès
- Calcaire
- Schiste
- Silt argileux
- Argile plastique
- Argile silteuse
- Craie peu résistante
- Argile moynt résistante
- Argile peu/moynt résistante
- Argile peu résistante
- Argile résistante
- Limons/argiles sableuses
- Sable
- Argiles marneuses
- Calcaire marneux
- Argiles plus ou moins sableuse
- Calcaire tendre
- Calcaire, limons, craie
- Argiles et craie
- Limons argileux
- Calcaires
- Argiles
- Calcaires / marnes
- Marno-calcaires
- Limon argilo-crayeux
- Limons sablo-argileux
- Limon argilo-silteux
- Mort terrain
- Craie compacte
- Limon sableux
- Limon argileux à sableux
- Limon et argile
- Limon crayeux compact
- Craie
- Silt limoneux
- Limon silteux
- Craie beige
- Silt
- Craie sableuse
- Craie limoneuse
- Argile limoneuse
- inconnu
- Limon
- Limon lgrt argileux
- Craie altérée
- Limon très crayeux
- Argile
- Limon argileux
- Limon crayeux
- Sable argileux
- N/A

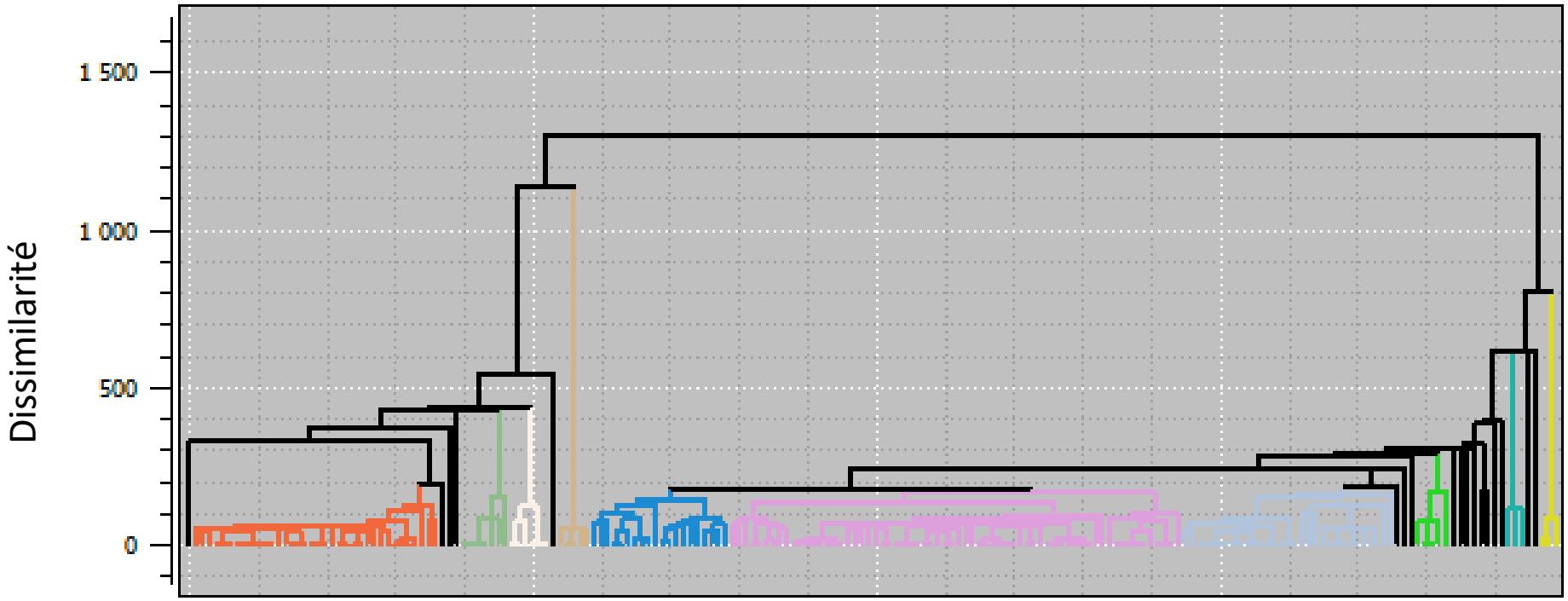
Cas d'étude: classification d'échantillons de sol

		c' (kPa)		cu (kPa)		phi		pl (kPa)	
	Total Count	Moyenne	Ecart Type	Moyenne	Ecart Type	Moyenne	Ecart Type	Moyenne	Ecart Type
Argile	15	8.09	2.47	40.00	13.54	27.73	0.86	808.18	394.34
Argile limoneuse	5	15.00	0.00	55.00	0.00	25.00	0.00	1135.00	0.00
Argile moynt résistante	8	10.00	0.00	20.00	0.00	30.00	0.00	700.00	0.00
Argile peu résistante	4	3.00	0.00	5.00	0.00	20.00	0.00	250.00	0.00
Argile peu/moynt résistante	28	5.00	0.00	10.00	0.00	25.00	0.00	500.00	0.00
Argile plastique	3	15.00	0.00	67.00	0.00	25.00	0.00	670.00	0.00
Argile résistante	4	12.00	0.00	25.00	0.00	30.00	0.00	1000.00	0.00
Argile silteuse	31	13.23	2.71	47.45	17.25	23.55	2.27	772.26	415.93
Argiles	16	8.00	0.00	25.00	0.00	28.00	0.00	529.38	67.13
Argiles et craie	9	8.00	0.00	25.00	0.00	28.00	0.00	600.00	0.00
Argiles marneuses	6	8.00	0.00	25.00	0.00	28.00	0.00	800.00	0.00
Argiles plus ou moins sableuse	30	10.00	0.00	20.00	0.00	28.00	0.00	800.00	0.00

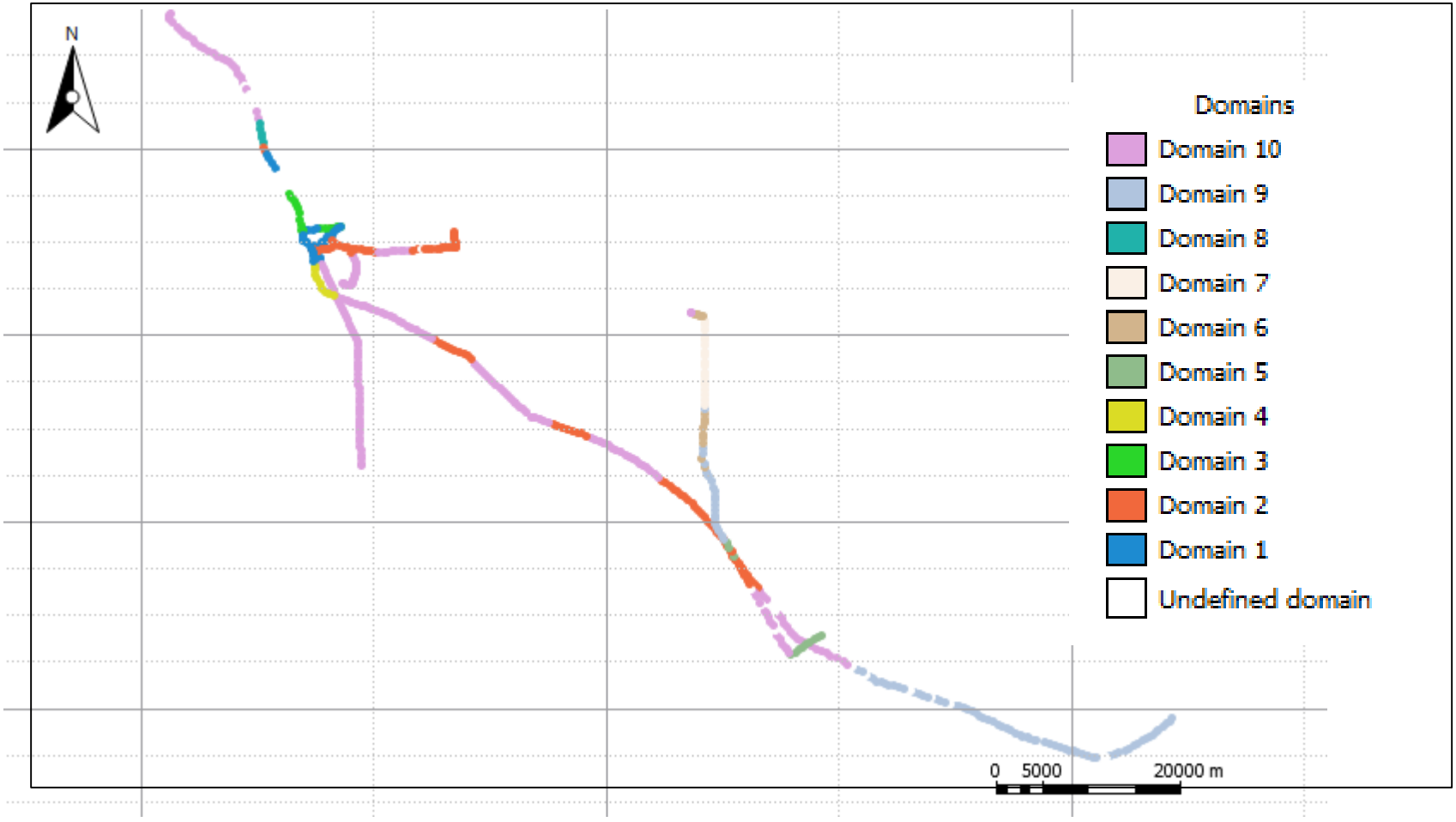
- Comment regrouper mes échantillons?
- Argile avec argile? Limon argileux avec argile limoneuse? Mais moyenne sur c' très différente!

Cas d'étude: classification d'échantillons de sol

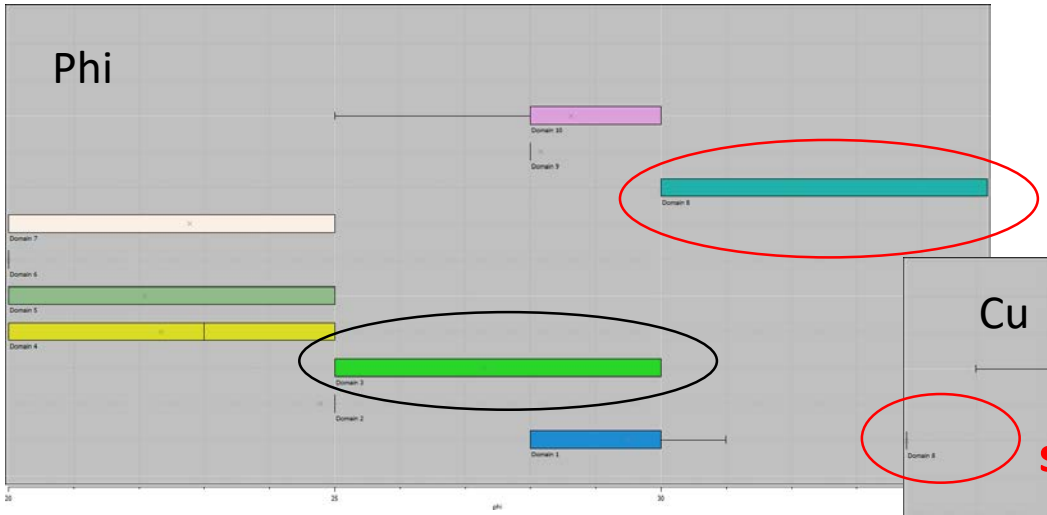
- Algorithme de classification
- Plusieurs tests
- 10 domaines

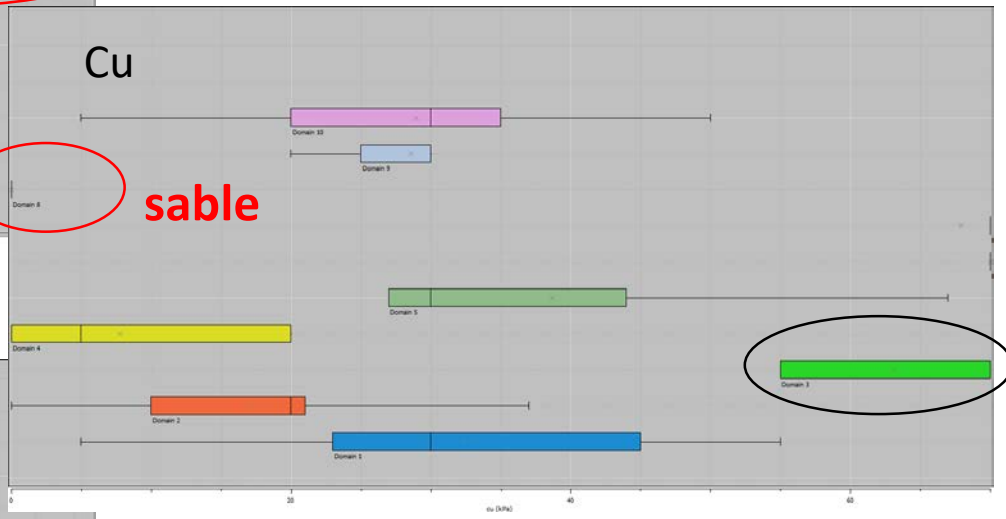


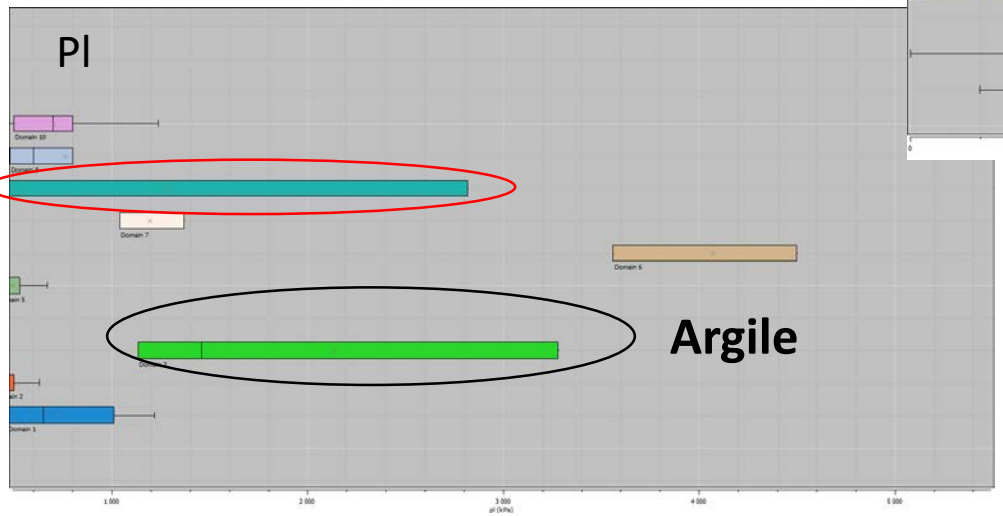
Cas d'étude: classification d'échantillons de sol



Cas d'étude







Résultats

- Exemple « géotechnique »: sans connaître les paramètres, on arrive à un résultat cohérent

- Idée d'application: pouvoir trouver des zones de faiblesse au sein d'une même lithologie pour appliquer des designs particuliers par zone?

Résultats

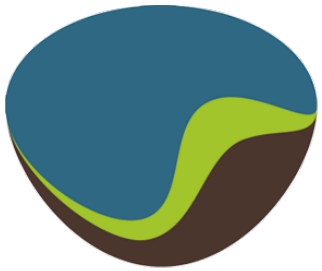
- Algorithme de classification : une classification objective, rapide et reproductible

- Mise à jour du regroupement par SVM: amélioration de la classification lorsque de nouvelles données arrivent

Merci pour votre attention

Marie-Cécile Febvey

febvey@geovariances.com



Geovariances

Where no one has gone before