Classification de cartes de décisions pour des contextes opérationnels de crise nucléaire

François MICHELON

Ecole Centrale de Nantes - Phimeca Paris

Intitulé

"Élaborer une procédure d'aide à la décision intégrant de manière fidèle et intelligible les informations apportées par des modèles physiques et probabilistes."

Intitulé

"Élaborer une procédure d'aide à la décision intégrant de manière fidèle et intelligible les informations apportées par des modèles physiques et probabilistes."

- accident rejet atmosphère
- prise de décisions, que faire ?
- ingé simulations, entrées incertaines donc variation panel de cartes
- synthèse info sans perte info, presentation non-spécialiste

Ce stage a donc porté sur la mise au point de méthodes de tris automatisés et de sélections de simulations, pour être capable de faire ressortir un panel pertinent et complet de sorties.

Modèle physique

Paramètres :

- les données météorologiques de la zone impactée
- les séries temporelles de rejet à la source
- modèle de la dispersion atmosphérique
- vitesse de dépôt.

IRSN

Figure – Institut de Radioprotection et de Sûreté Nucléaire

Modèle physique

Paramètres :

- les données météorologiques de la zone impactée
- les séries temporelles de rejet à la source
- modèle de la dispersion atmosphérique
- vitesse de dépôt.

Figure – Institut de Radioprotection et de Sûreté Nucléaire

L'IRSN a développé un modèle mathématique permettant de simuler dans le temps et l'espace la dispersion des particules radioactives à partir d'un point d'émission.

Ce modèle nécessite la connaissance de nombreux paramètres, tous soumis à un certain degré d'incertitude

Le but de la catégorisation est de dégager des sous-groupes représentatifs des simulations obtenues. Une fois le clustering effectué, on cherche aussi à évaluer les sous-groupes donc à étudier les valeurs des entrées ayant générées les différences entre clusters et à évaluer les populations pouvant être affectées par l'accident simulé.

Doses et mesures de protection

Concepts de dose :

- Dose absorbée
- Dose inhalation
- Dose efficace

Concepts de dose :
- Dose absorbée
- Dose inhalation
- Dose efficace

Les effets sur la santé dépendent de plusieurs paramètres :

- la quantité d'énergie transmise par les rayonnements et leurs natures ;
- les modalités d'exposition (interne - par ingestion notamment - ou externe) ;
- l'organe ou le tissu atteint (poumons, peau...).

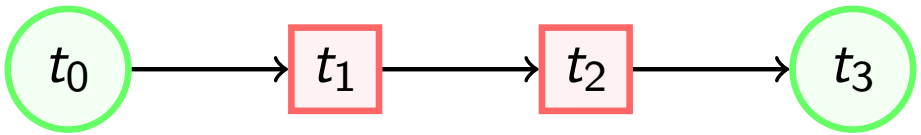
Tout d'abord, on calcule la dose absorbée qui est l'énergie déposée par unité de masse par un rayonnement ionisant. Ensuite, pour prendre en compte l'influence de deux paramètres - le type de tissu ou d'organe touché et le type de rayonnement – on calcule deux doses :

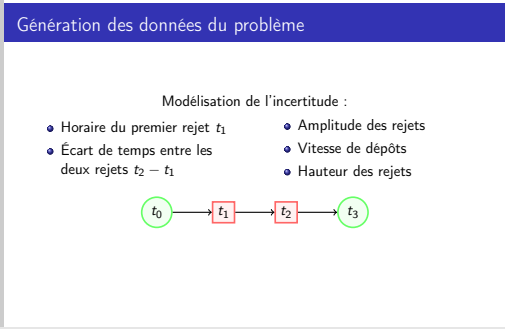
- la première, appelée dose inhalation (en Sievert, Sv), prend en compte les effets des rayonnements des radionucléides inhalés ;
- la seconde, appelée dose efficace (en Sievert, Sv), prend en compte le type de tissu ou d'organe touché.

Génération des données du problème

Modélisation de l'incertitude :

- Horaire du premier rejet t_1
- Écart de temps entre les deux rejets $t_2 - t_1$
- Amplitude des rejets
- Vitesse de dépôts
- Hauteur des rejets

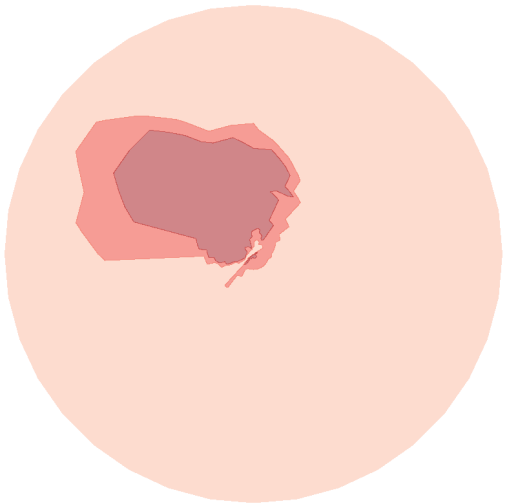




Pour obtenir des cartes diverses, l'accident simulé est basé sur un scénario fictif et grave. On simule sur 3 jours avec un pas de temps d'une heure et avec 2 rejets de particules nocives d'intensités variées et d'instants d'émission différents. On se réduit toutefois à une météo unique pour l'ensemble des calculs.

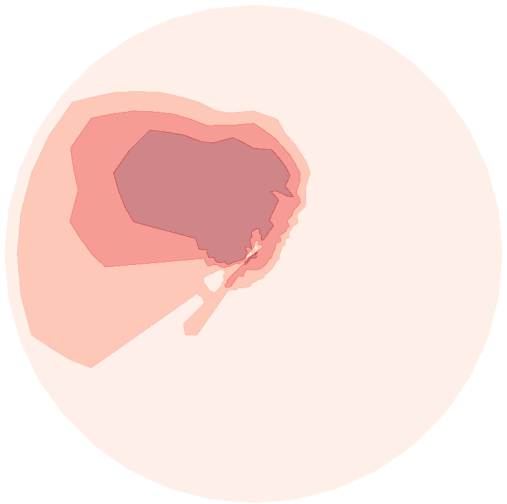
Sur avis d'expert, on modélise l'incertitude de l'horaire du premier rejet, de l'amplitude des rejets, des vitesses de dépôts des particules et de la hauteur des rejets dans l'atmosphère par des lois Gaussiennes et l'écart de temps entre les deux rejets est modélisé par une loi Lognormale. Et pour chacun des 100 échantillons de paramètres obtenus, on lance les calculs 3 fois en faisant varier le modèle de dispersion atmosphérique. On obtient alors 300 simulations représentatives d'un potentiel accident.

Création des cartes de décisions

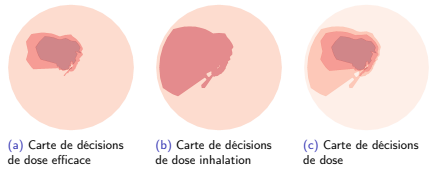


(a) Carte de décisions de dose efficace

(b) Carte de décisions de dose inhalation



(c) Carte de décisions de dose



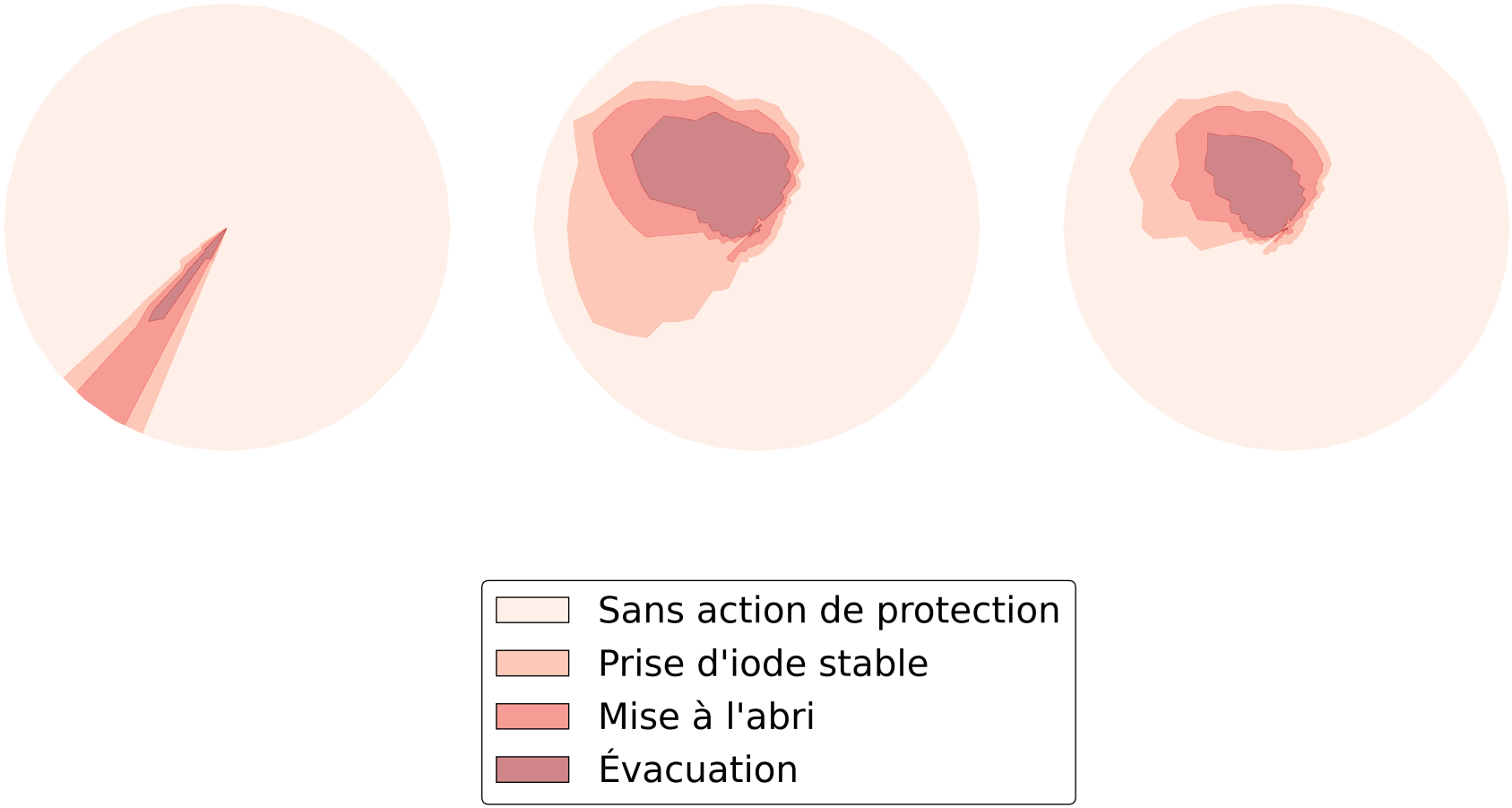
(a) Carte de décisions de dose efficace (b) Carte de décisions de dose inhalation (c) Carte de décisions de dose

A partir du jeu de données calculé, on crée les cartes de doses suivantes centrées autour de la zone d'émission

Les contours correspondant aux décisions "prise d'iode", "mise à l'abri" et "évacuation" coïncident aux zones où la dose (efficace ou inhalation) dépasse les niveaux recommandés par l'IRSN.

Pour chaque simulation, on regroupe l'information de dose efficace et de dose inhalation au cours du temps en une carte contenant les zones de décisions associées aux valeurs de doses à la fin de la simulation. La surface du panache de dose étant croissante en temps, on se contente de travailler sur la pire situation pour chaque carte, qui correspond à l'instant final. On se contente de l'information de dépassement de seuil, ainsi deux régions avec une dose efficace de 60 mSv et 1000 mSv sont équivalentes.

Diversité des cartes de décisions



Diversité des cartes de décisions

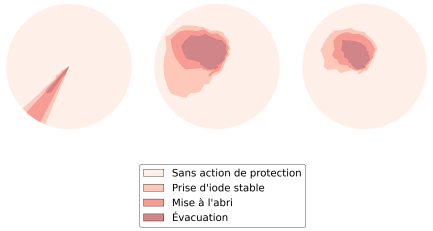


Figure – Cartes de décisions de dose

La diversité des cartes de décisions est importante. Deux cartes peuvent être très proches comme complètement opposées ou posséder une ressemblance globale sans pour autant être similaire.

C'est là que rentre en jeu le clustering. Comme énoncé dans l'introduction, il s'agit alors de résumer l'information du panel de cartes. Ainsi, on s'attachera à regrouper les cartes par paquet de la meilleure façon possible et à développer des outils permettant de visualiser les caractérisques de chaque catégorie. La suite du rapport se concentrera sur les différentes méthodes de catégorisation mises en œuvre lors du stage.

Introduction de la méthode

$$\mathcal{X} = \{x_1, x_2, ..., x_n\} \in \left(\mathbb{R}^d\right)^n$$

$$\mathcal{Y} = \{y_1, y_2, ..., y_n\} \in \left(\mathbb{R}^3\right)^n$$

$$avec \quad d >> 3$$

$\mathcal{X} = \{x_1, x_2, ..., x_n\} \in (\mathbb{R}^d)^n$
$\mathcal{Y} = \{y_1, y_2, ..., y_n\} \in (\mathbb{R}^3)^n$
avec $d >> 3$

La technique tSNE permet de visualiser un jeu de données de haute dimension, comme c'est le cas dans notre étude, en les plongeant en un jeu de données de dimension 2 ou 3. Dans cette présentation, on se referera à $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$ pour l'espace de faible dimension et à y_i comme le plongement de x_i, point appartenant au jeu de données de haute dimension $\mathcal{X} = \{x_1, x_2, ..., x_n\}$.

Cette méthode permet notamment de révéler la struture de données complexes sans trop la simplifier. En effet, l'algorithme tSNE est une méthode non-linéaire préservant la proximité des points lors du plongement en faible dimension. Etant donné que l'on cherche à caractériser la proximité de simulation potentiellement avec des panaches couvrant des zones qui ne s'intersectent pas, il est alors compréhensible qu'une méthode linéaire ne serait pas suffisante pour obtenir une classification adéquate.

Développements mathématiques

p_{ij} = la similarité du point x_i au point x_j

q_{ij} = la similarité du point y_i au point y_j

p_{ij} = la similarité du point x_i au point x_j

q_{ij} = la similarité du point y_i au point y_j

La première étape de l'algorithme est de convertir les distances Euclidiennes des points de hautes dimensions en des probabilités conditionnelles répresentant leur similarité. La similarité du point x_i au point x_j réprésente la probabilité conditionnelle $p_{j|i}$, que x_i choisisse x_j comme son voisin, dans le cas où les voisins sont choisis proportionellement à leur densité de probabilité sous une loi Gaussienne centrée en x_i.

Pour les points x_j dont la distance Euclidenne à x_i est faible, $p_{j|i}$ est grand.

Probabilité jointe : afin que chaque point x_i ait de l'importance dans la fonction de coût.

Développements mathématiques

p_{ij} = la similarité du point x_i au point x_j

q_{ij} = la similarité du point y_i au point y_j

Deux distributions : P et Q

$\rightarrow$ mesure de dissimilarité

p_{ij} = la similarité du point x_i au point x_j

q_{ij} = la similarité du point y_i au point y_j

Deux distributions : P et Q

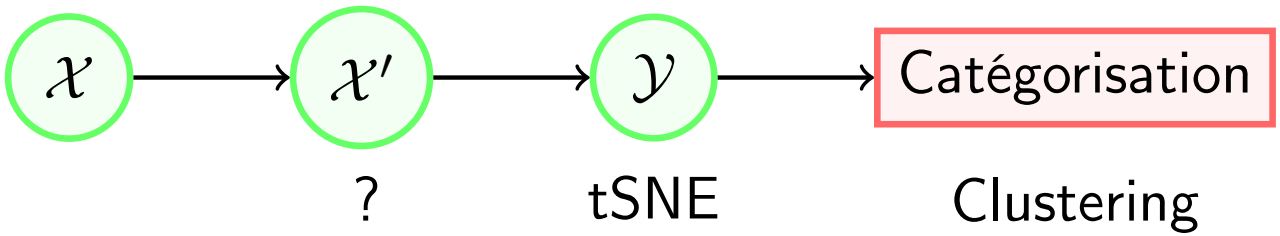
$\rightarrow$ mesure de dissimilarité

On calcule alors des probabilités conditionelles similaires que l'on note $q_{j|i}$. Cette dernière possède une fonction de coût avec un gradient plus simple que dans la méthode SNE. On utilise des lois de Student.

Le choix de la distribution de Student est aussi justifié par la queue de sa distribution plus importante qu'une Gaussienne, ce qui permet de mieux séparer les points x_i, x_j avec une similarité moyenne dans $\mathcal{Y}$.

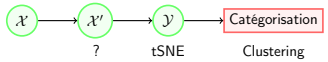
Les points y_i sont alors calculés en minimisant la divergence de Kullback-Leibler de P par rapport Q qui est une mesure de dissimilarité entre les distributions P et Q.

Méthode numérique



$$\mathcal{X}' = \{x'_1, x'_2, ..., x'_n\} \in (\mathbb{R}^p)^n$$

$$avec \quad d >> p > 3$$

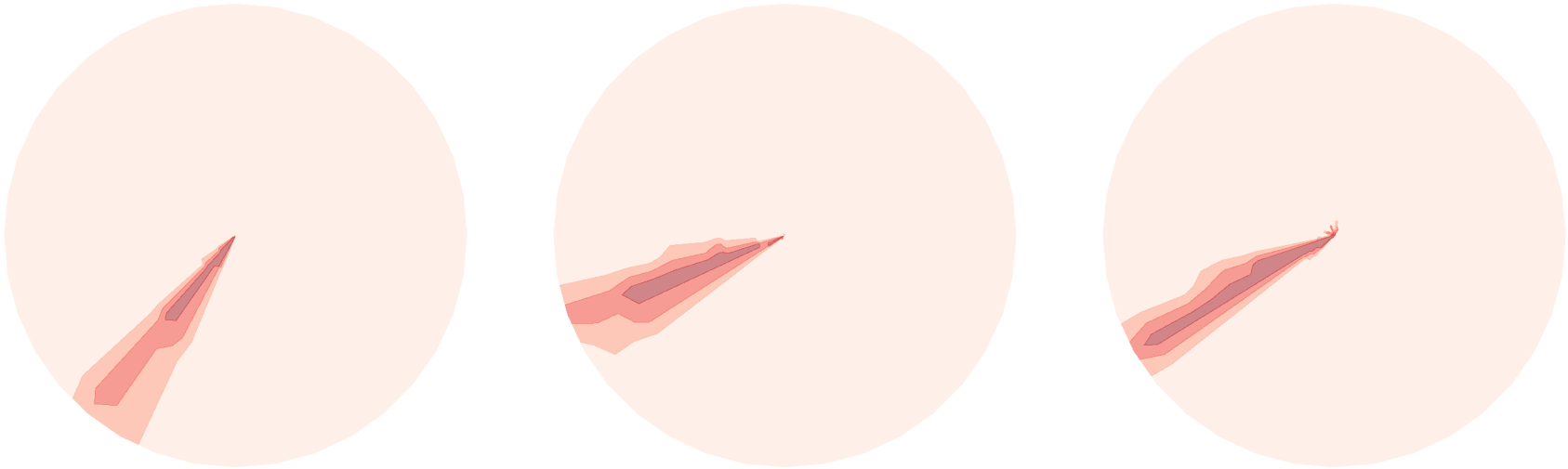


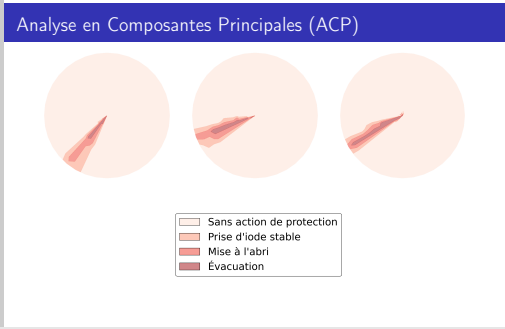
$\mathcal{X}' = \{x'_1, x'_2, ..., x'_n\} \in (\mathbb{R}^p)^n$

avec $d >> p > 3$

Cette fonction nécessite toutefois un traitement des données en amont. En effet, il est nécesssaire de réduire la dimension du jeu de données avant d'appeler la fonction, sinon le temps de calcul serait trop important. Il est indiqué dans la documentation de la fonction scikit-learn que la dimension des données en entrée soit au maximum équivalente au nombre d'échantillons. Dans notre cas, il faudrait passer d'une dimension de $36 \times 40 = 1440$ à environ 300. Nous allons maintenant développer différentes méthodes de traitement des données.

Analyse en Composantes Principales (ACP)





L'Analyse en Composante Principale ou ACP est une méthode linéaire permettant de résumer l'information d'un jeu de données en réduisant le nombre de variables.

L'un des problèmes majeurs de l'ACP est la linéarité de la méthode. En effet, les panaches obtenus dans les différentes simulations peuvent être de formes similaires mais être orientés différemment. On aimerait donc que la catégorisation puisse faire des simulations 2 et 3 des voisins proches et qu'elles soient plus éloignées de la simulation 1.

les composantes des simulations obtenues après ACP sont des combinaisons linéaires des composantes avant transformations ainsi cette méthode n'est pas adéquate pour rendre compte de la complexité des cartes.

Modèles Auto-Associatifs (MAA)

limite de l'ACP : $\rightarrow$ différence d'orientation des panaches de deux simulations proches

Compromis : critère métrique de l'ACP VS critère topologique de proximité

limite de l'ACP : → différence d'orientation des panaches de deux simulations proches

Compromis : critère métrique de l'ACP VS critère topologique de proximité

Les modèles auto-associatifs (MAA) sont une extension non-linéaire de l'Analyse en Composantes Principales (ACP) basée sur des approximations successives d'ensemble de points par des variétés différentiables de dimension croissante. Une variété, au sens mathématique du terme, est un espace topologique localement euclidien. De la même manière que les droites et les plans sont des espaces vectoriels de dimension 1 et 2 respectivement, les courbes et les surfaces sont des variétés de dimension 1 et 2 respectivement.

Compromis entre le critère métrique de l'ACP et un critère topologique.

Or la distance Euclidienne implique que peu importe la différence d'orientation des panaches de deux simulations proches, si les zones recouvertes par les panaches sont disjointes alors ces deux simulations seront considérées comme étant très éloignées.

Auto-encodeur

$$E_\phi : \mathcal{X} \to \mathcal{F} \quad D_\theta : \mathcal{F} \to \mathcal{X} \tag{1}$$

$$L(\theta^*, \phi^*) = \min_{\theta,\phi} \frac{1}{n} \sum_{x \in \mathcal{X}} \| x - D_\theta(E_\phi(x)) \|_2^2 \tag{2}$$

$$E_\phi : \mathcal{X} \to \mathcal{F} \quad D_\theta : \mathcal{F} \to \mathcal{X} \tag{1}$$

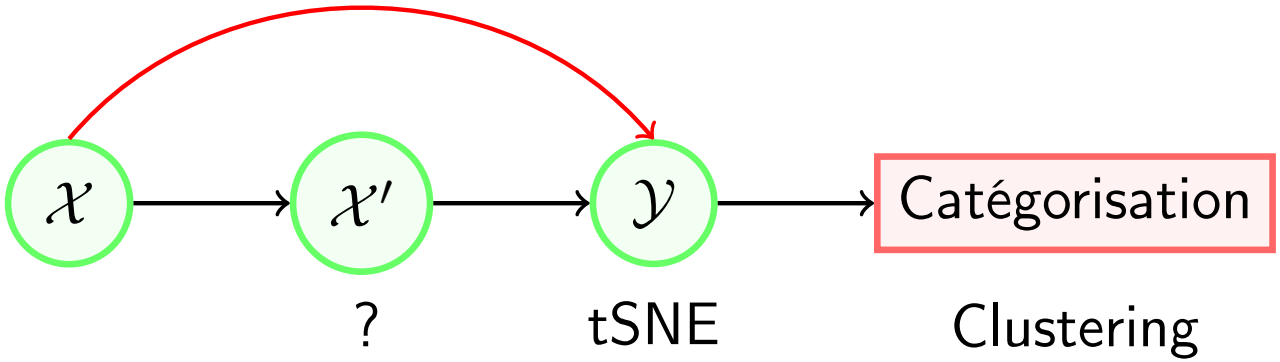
$$L(\theta^*, \phi^*) = \min_{\theta,\phi} \frac{1}{n} \sum_{x \in \mathcal{X}} \| x - D_\theta(E_\phi(x)) \|_2^2 \tag{2}$$

Un auto-encodeur se compose toujours de deux parties, l'encodeur et le décodeur qui peuvent être définies comme des transitions E_ϕ et D_θ.

Pour évaluer la qualité d'un auto-encodeur, on définit une densité de probabilité de référence μ_{ref} sur $\mathcal{X}$ générant les entrées du jeu de données. On cherche donc la valeur des paramètres minimisant la fonction de coût.

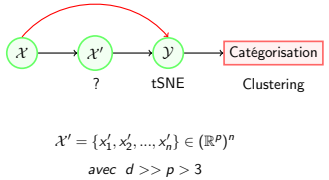
Etant donné que les entrées du problème sont des cartes qui peuvent être vues comme des images, il va de soi d'utiliser un réseau convolutif. La classification d'images implique la reconnaissance de patterns, utiliser un réseau traditionnel dans ce cas nécessite un temps de calcul important, car chaque neurone est ajusté individuellement. Tandis que pour des réseaux convolutifs, l'utilisation de filtres permet d'exploiter les différences locales, sans demander une puissance de calcul importante.

Distance de Wasserstein



$$\mathcal{X}' = \{x'_1, x'_2, ..., x'_n\} \in (\mathbb{R}^p)^n$$

$$avec \quad d >> p > 3$$



Nous avons vu que la méthode tSNE nécessite le calcul de la distance Euclidenne entre l'ensemble des couples de points x_i et x_j. Or cette distance est sujette à la "malédiction de la dimension" et qu'elle ne permet de refléter la proximité souhaitée entre deux simulations. Il faut alors déterminer une distance permettant de s'affranchir de ces limitations.

Barycentres de Wasserstein

$p_1, p_2, \cdots, p_n$ distributions et la catégorisation $\mathcal{C} = \bigsqcup_{k=1}^{K} \mathcal{C}_k$.

$$\forall k \in \{1, 2, \cdots, K\}, \quad p_k^* = \underset{p}{argmin} \sum_{p_k \in \mathcal{C}_k} W(p, p_k) \tag{3}$$

$p_1, p_2, \cdots, p_n$ distributions et la catégorisation $\mathcal{C} = \bigsqcup_{k=1}^{K} \mathcal{C}_k$.

$$\forall k \in \{1, 2, \cdots, K\}, \quad p_k^* = \underset{p}{argmin} \sum_{p_k \in \mathcal{C}_k} W(p, p_k) \qquad (3)$$

On a développé quatres méthodes de clustering différentes : ACP, auto-encodeurs, MAA et Wasserstein. Afin de déterminer la procédure optimale en période de crise, il reste maintenant à comparer ces méthodes. Or on cherche à faire du clustering non-supervisé avec des simulations complexes, i.e. on ne peut déterminer avec certitude si une simulation est bien classée ou non. Il nous reste donc à développer des indices permettant d'évaluer les caractéristiques des méthodes de clustering.

On cherche à obtenir une représentation moyenne pour chacun des sous-groupes. Etant donné la non-linéarité des données, il convient alors qu'effectuer une moyenne empirique ne conviendrait pas. Si l'on considère les simulations comme des densité de probabilité, on peut alors calculer le barycentre de Wasserstein de chaque sous-groupe pour obtenir une représentation simplifiée d'une catégorisation.

aucun soucis normalisation car sous-groupe proche mais temps calculs long

Indice de Davies Bouldin

$$T_k = \frac{1}{|\mathcal{C}_k|} \sum_{x \in \mathcal{C}_k} W(x, x_k) \tag{4}$$

$$S_{k,l} = W(x_k, x_l) \tag{5}$$

$$D = \frac{1}{K} \sum_{k=1}^{K} \max_{l \neq k} \frac{T_l + T_k}{S_{k,l}} \tag{6}$$

$$T_k = \frac{1}{|C_k|} \sum_{x \in C_k} W(x, x_k) \quad (4)$$

$$S_{k,l} = W(x_k, x_l) \quad (5)$$

$$D = \frac{1}{K} \sum_{k=1}^{K} \max_{l \neq k} \frac{T_l + T_k}{S_{k,l}} \quad (6)$$

L'indice de Davies-Bouldin permet d'évaluer si une catégorisation donnée possède des clusters compacts et distants les uns des autres.

Pour un sous-groupe $\mathcal{C}_k$ d'une catégorisation $\mathcal{C} = \bigsqcup_{i=1}^{K} \mathcal{C}_i$ donnée, on définit ainsi l'homogénéité T_k d'un cluster comme la moyenne des distances de chacun des points contenus dans ce cluster au barycentre x_k :

Nous avions dit plus haut vouloir aussi que les clusters soient loin les uns des autres. Pour quantifier cela, on définit la séparation de deux clusters comme la distance entre leurs centroïdes :

Nous avons maintenant deux critères à optimiser. Pour nous faciliter la tâche, nous pouvons les regrouper en un seul critère, l'indice de Davies-Bouldin. L'idée de cet indice est de comparer les distances intraclusters (c'est l'homogénéité), que l'on veut faibles, aux distances interclusters (la séparation), que l'on veut grandes.

Indice de Silhouette

$$a(x) = \frac{1}{|\mathcal{C}_k| - 1} \sum_{u \in \mathcal{C}_k,\ u \neq x} W(u, x) \tag{7}$$

$$b(x) = \min_{l \neq k} \frac{1}{|\mathcal{C}_l|} \sum_{u \in \mathcal{C}_l} W(u, x) \tag{8}$$

$$s(x) = \frac{b(x) - a(x)}{max(a(x), b(x))} \tag{9}$$

Indice de Silhouette

$$a(x) = \frac{1}{|\mathcal{C}_k| - 1} \sum_{u \in \mathcal{C}_k,\, u \neq x} W(u, x) \quad (7)$$

$$b(x) = \min_{l \neq k} \frac{1}{|\mathcal{C}_l|} \sum_{u \in \mathcal{C}_l} W(u, x) \quad (8)$$

$$s(x) = \frac{b(x) - a(x)}{max(a(x), b(x))} \quad (9)$$

Etant donné qu'il existe des simulations intermédiaires, des simulations proches de deux sous-groupes de simulations distincts, on peut s'attendre à ce que leur catégorisation dans l'un des sous-groupes ne soit pas une tâche aisée. Afin de tester si ces simulations intermédiaires sont bien classées ou non, on définit l'indice de Silhouette

On calcule la distance moyenne de x à tous les autres points du cluster $\mathcal{C}_k$ auquel il appartient :

On calcule la plus petite valeur que pourrait prendre a(x), si x était assigné à un autre cluster :

Si x a été correctement assigné, alors $a(x) < b(x)$. le coef est donc d'autant plus proche de 1 que l'assignation de x à son cluster est satisfaisante. Pour évaluer un clustering, on peut calculer son coefficient de silhouette moyen.

Indice de Rand ajusté

$$\mathcal{C} = \bigsqcup_{k=1}^{K} \mathcal{C}_k \quad et \quad \mathcal{C}' = \bigsqcup_{k=1}^{K} \mathcal{C}'_k$$

	groupés dans $\mathcal{C}$	**séparés dans $\mathcal{C}$**
groupés dans $\mathcal{C}'$	a	b
séparés dans $\mathcal{C}'$	c	d

$$RI(\mathcal{C}, \mathcal{C}') = \frac{a+b}{a+b+c+d} \tag{10}$$

Indice de Rand ajusté

$$\mathcal{C} = \bigsqcup_{k=1}^{K} \mathcal{C}_k \quad et \quad \mathcal{C}' = \bigsqcup_{k=1}^{K} \mathcal{C}'_k$$

	groupés dans $\mathcal{C}$	**séparés dans $\mathcal{C}$**
groupés dans $\mathcal{C}'$	a	b
séparés dans $\mathcal{C}'$	c	d

$$RI(\mathcal{C}, \mathcal{C}') = \frac{a+b}{a+b+c+d} \tag{10}$$

On peut ensuite chercher à évaluer la stabilité d'une méthode. L'intérêt de se crirtère est double. Premièrement, on peut comparer les quatres méthodes entre elles. Deuxièmement, ce critère est pertinent pour choisir le nombre de clusters d'une catégorisation donnée

Afin d'appréhender la définition de l'indice de Rand ajusté, on va rapidement développer l'indice de Rand simple et comprendre ses limites. On peut alors comprendre l'indice de Rand comme la probabilité qu'une paire de simulation choisies au hasard soit classée de la même façon dans les deux catégorisations $\mathcal{C}$ et $\mathcal{C}'$.

Afin de corriger cet indice pour qu'il ne soit plus influencé par la potentielle différence de cardinal entre les sous-groupes.

Indice d'information mutuelle ajustée

$$H(\mathcal{C}) = -\sum_{i=1}^{K} P_{\mathcal{C}}(i) \log P_{\mathcal{C}}(i) \tag{11}$$

$$P_{\mathcal{C}}(i) = \frac{|\mathcal{C}_i|}{n} \quad P_{\mathcal{C},\mathcal{C}'}(i,j) = \frac{|\mathcal{C}_i \cap \mathcal{C}'_j|}{n}$$

$$H(\mathcal{C}) = -\sum_{i=1}^{K} P_{\mathcal{C}}(i) \log P_{\mathcal{C}}(i) \quad (11)$$

$$P_{\mathcal{C}}(i) = \frac{|\mathcal{C}_i|}{n} \quad P_{\mathcal{C},\mathcal{C}'}(i,j) = \frac{|\mathcal{C}_i \cap \mathcal{C}'_j|}{n}$$

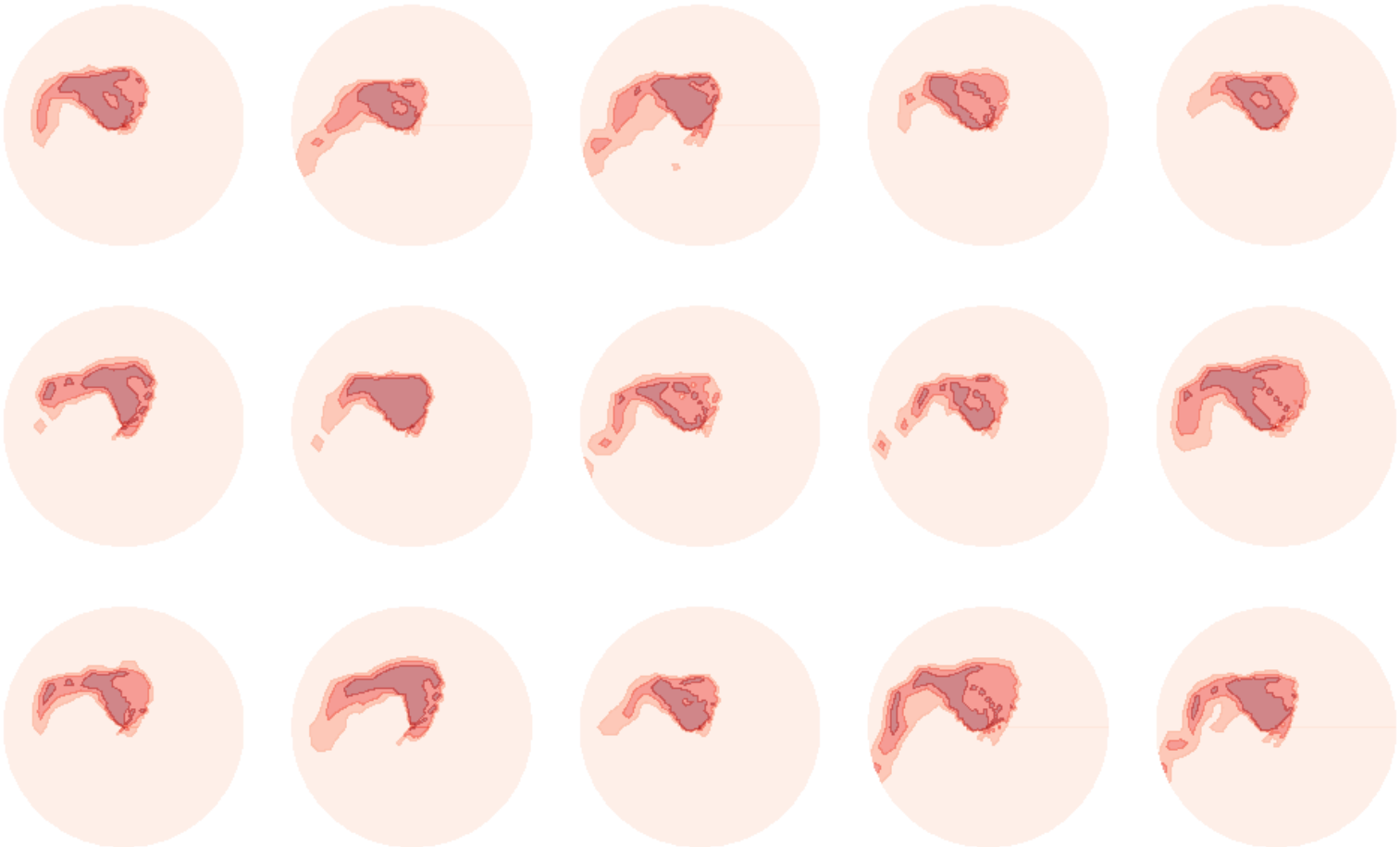
L'indice d'information mutuelle ajusté permet d'évaluer lui aussi la stabilité des méthodes en comparant deux clusterings.

$H(\mathcal{C})$ est positif et prend la valeur 0 quand il n'y a pas d'incertitude concernant l'appartenance d'une simulation à un sous-groupe, soit quand $K = 1$. On définit l'indice d'information mutuelle de la façon suivante :

Cet objet quantifie l'information similaire entre les deux catégorisations. On peut l'utiliser comme une mesure de similarité entre deux clusterings. Comme l'indice de Rand, l'indice d'information mutuelle dépend du nombre de clusters. On calcule alors un indice ajusté à l'indice d'information mutuelle :

De même que l'indice de Rand ajusté, l'indice d'information mutuelle ajusté prend la valeur 0 pour deux catégorisations déterminées au hasard et 1 pour deux partitions égales à des permutations de sous-groupes près.

Problème simplifié - Classe 0

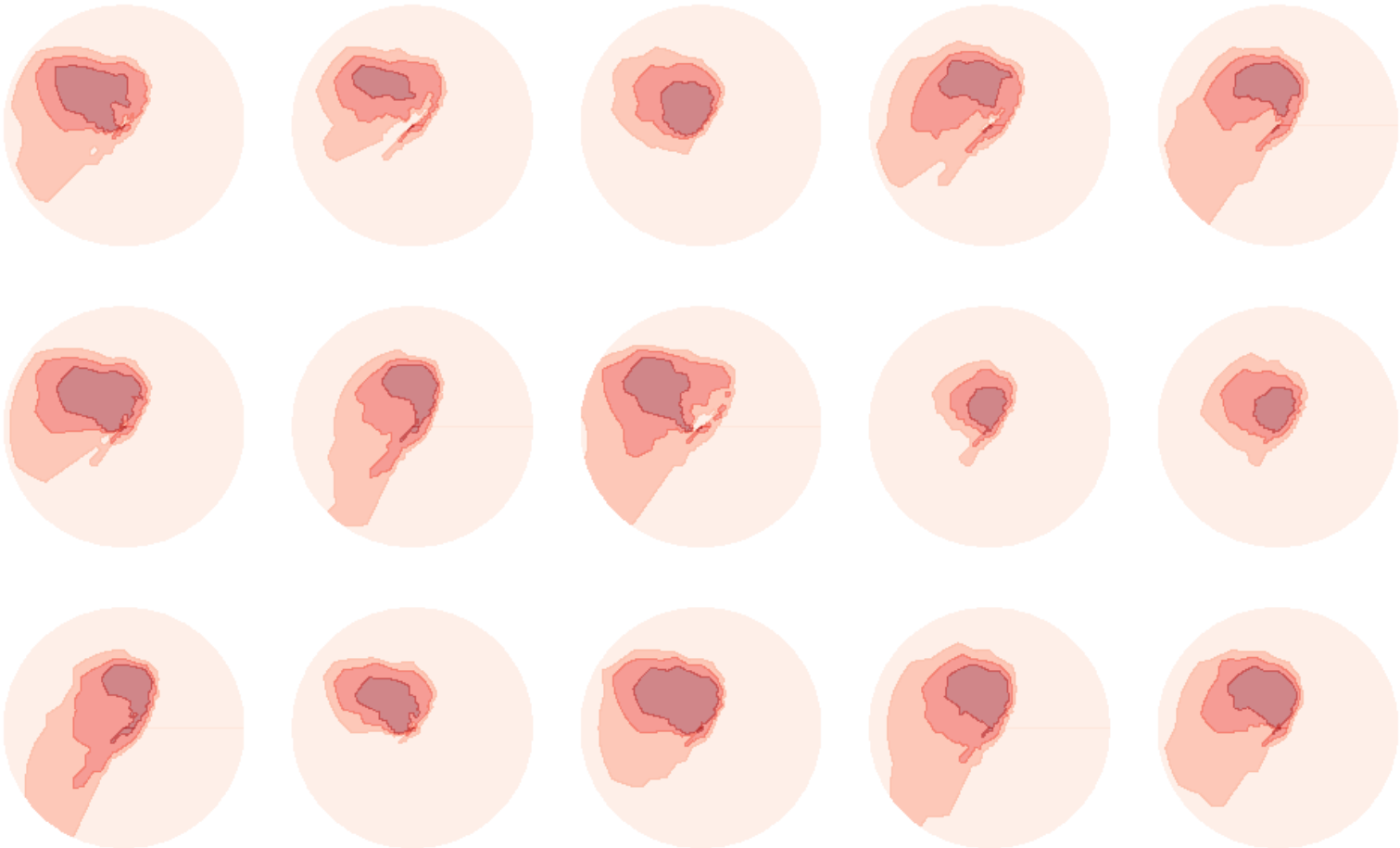


Etant donné la diversité des cartes et les problèmes de cartes intermédiaires, il est compliqué de comparer les modèles visuellement. C'est pourquoi on s'attachera à mettre en place les indices de comparaisons ainsi que les résultats des méthodes sur un problème simplifié.

On développe dans ce problème simplifié un exercice de clustering supervisé. On s'attache alors à étiqueter un échantillon de simulations. On définit trois classes.

Problème simplifié - Classe 1

Problème simplifié - Classe 2



Problème simplifié - Résultats

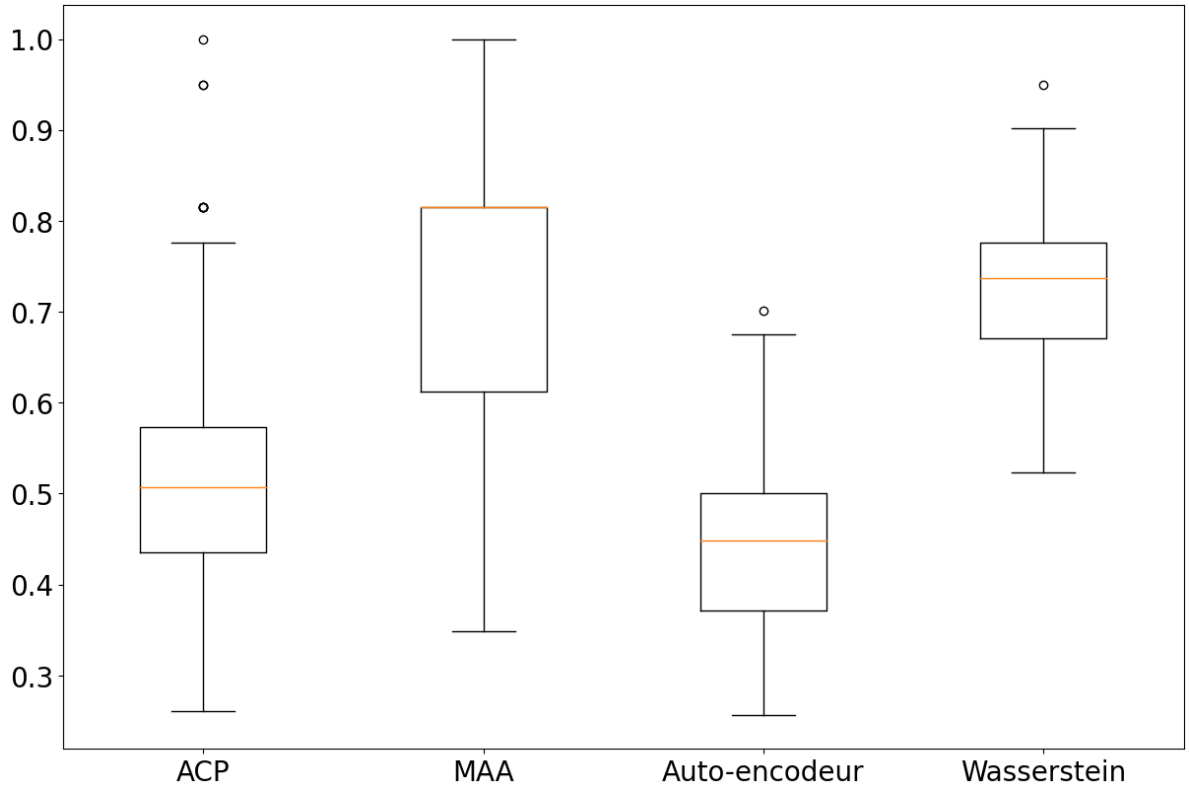


Figure – Boîtes à moustache de l'indice de Rand ajusté par cluster pour un nombre de sous-groupes fixé à 3

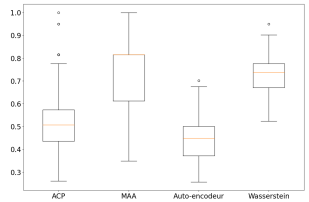


Figure – Boîtes à moustache de l'indice de Rand ajusté par cluster pour un nombre de sous-groupes fixé à 3

On procède alors à 100 catégorisations sur l'ensemble des 75 simulations de l'échantillon en utilisant chacun des quatre modèles avec un nombre de clusters fixé à trois. Pour chaque catégorisation, on calcule alors l'indice de Rand ajusté entre la catégorisation et l'étiquetage réalisé au préalable. On observe alors que les méthodes MAA et distance de Wasserstein produisent les catégorisations les plus satisfaisantes aux regards du problème. La moyenne des scores obtenus est légèrement supérieure avec la méthode MAA qu'avec la méthode distance de Wasserstein cependant cette dernière possède une variance plus faible.

Indices de Davies Bouldin

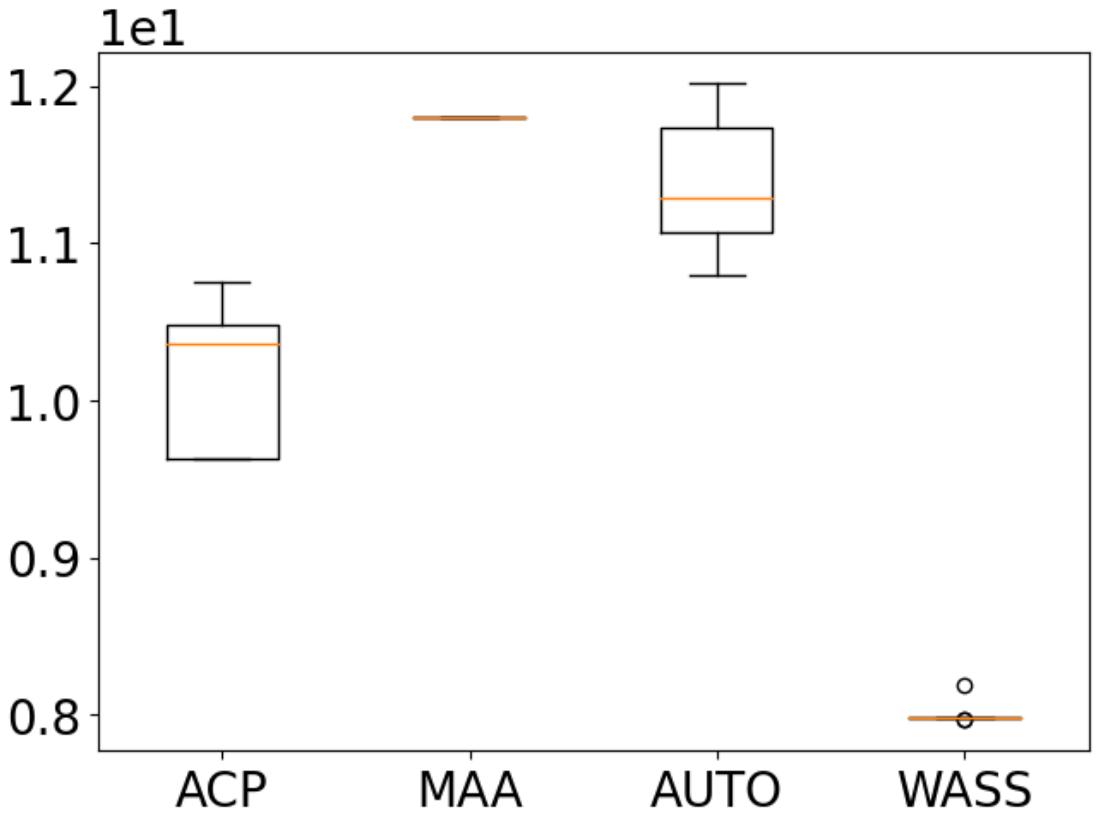


Figure – Nombre de clusters fixé à 7

Indices de Davies Bouldin

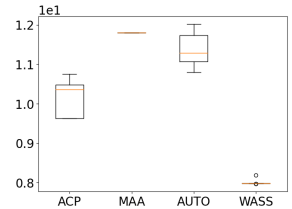


Figure – Nombre de clusters fixé à 7

Pour les comparaisons à l'aide des indices, on a effectué 10 catégorisations par nombre de clusters fixé variant entre 6 et 10 et par méthode.

On obtient les mêmes conclusions en observant ces deux figures. La première est que la méthode distance de Wasserstein est de loin la plus homogène et séparée au vu des indices utilisés. Tandis que les autres méthodes produisent des catégorisations de moins bonnes qualités selon ces indices.

Concernant le nombre de clusters permettant le plus d'homogénéité et de séparation pour la méthode de la distance de Wasserstein, il est de 8 pour l'indice de Silhouette et de 7 pour l'indice de Davies-Bouldin.

Indices de Silhouette

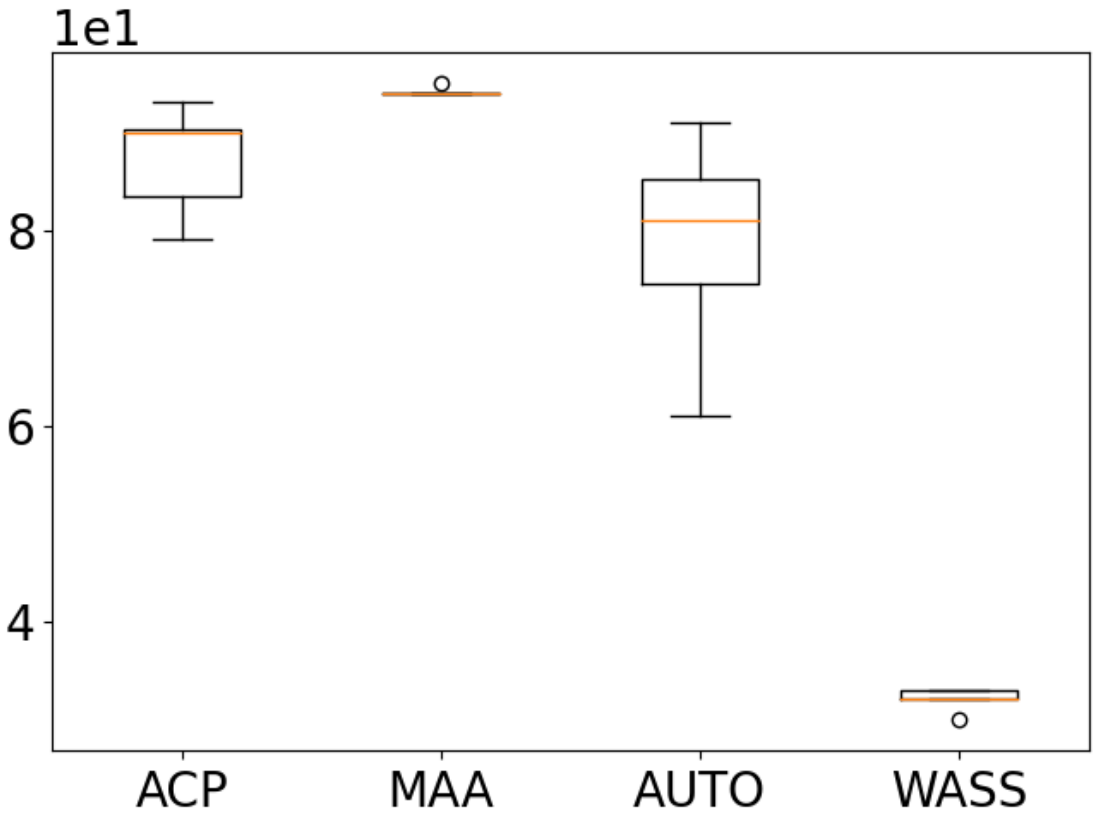


Figure – Nombre de clusters fixé à 8

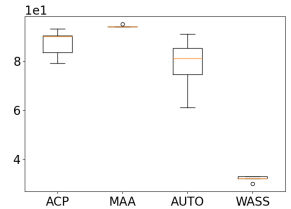


Figure – Nombre de clusters fixé à 8

Il faut cependant noter que ces indices présentent un biais important du fait que l'on utilise la distance de Wasserstein pour calculer les barycentres et les indices. Le score de la méthode distance de Wasserstein était largement prévisible. Cependant, du fait que c'est cette distance qui permet de comparer au mieux les simulations et de représenter les sous-groupes via les barycentres de Wasserstein on peut atténuer la conclusion précédente. En effet, le but de la catégorisation dans notre cas est de compacter l'information du groupe de simulations et donc il convient alors de choisir la méthode performant au mieux vis-à-vis des indices démontrant la meilleure catégorisation possible.

Indices de Rand ajusté

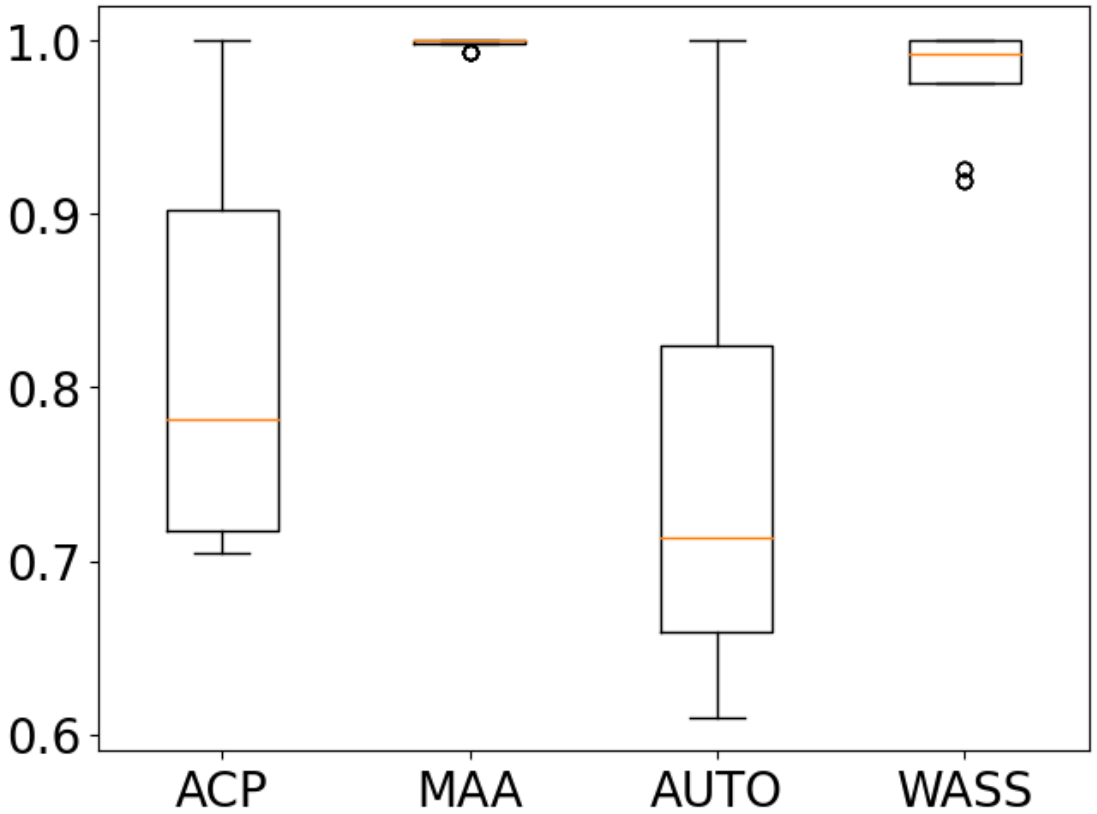


Figure – Nombre de clusters fixé à 8

Indices de Rand ajusté

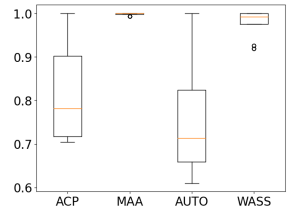


Figure – Nombre de clusters fixé à 8

On obtient les mêmes conclusions en observant ces deux figures. La première est que la méthode MAA est de loin la plus stable, même si la méthode distance de Wasserstein produit de bons résultats. Tandis que les méthodes ACP et auto-encodeur produisent des catégorisations variant beaucoup au vu des scores obtenus.

Indice d'information mutuelle ajustée

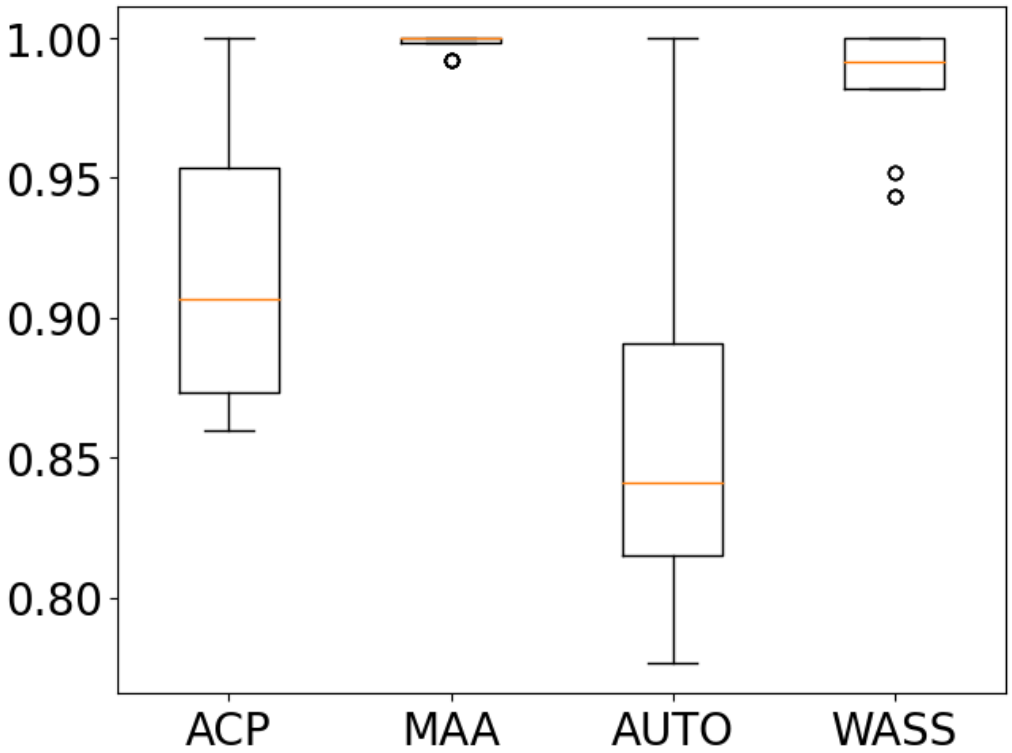


Figure – Nombre de clusters fixé à 8

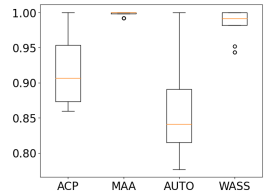
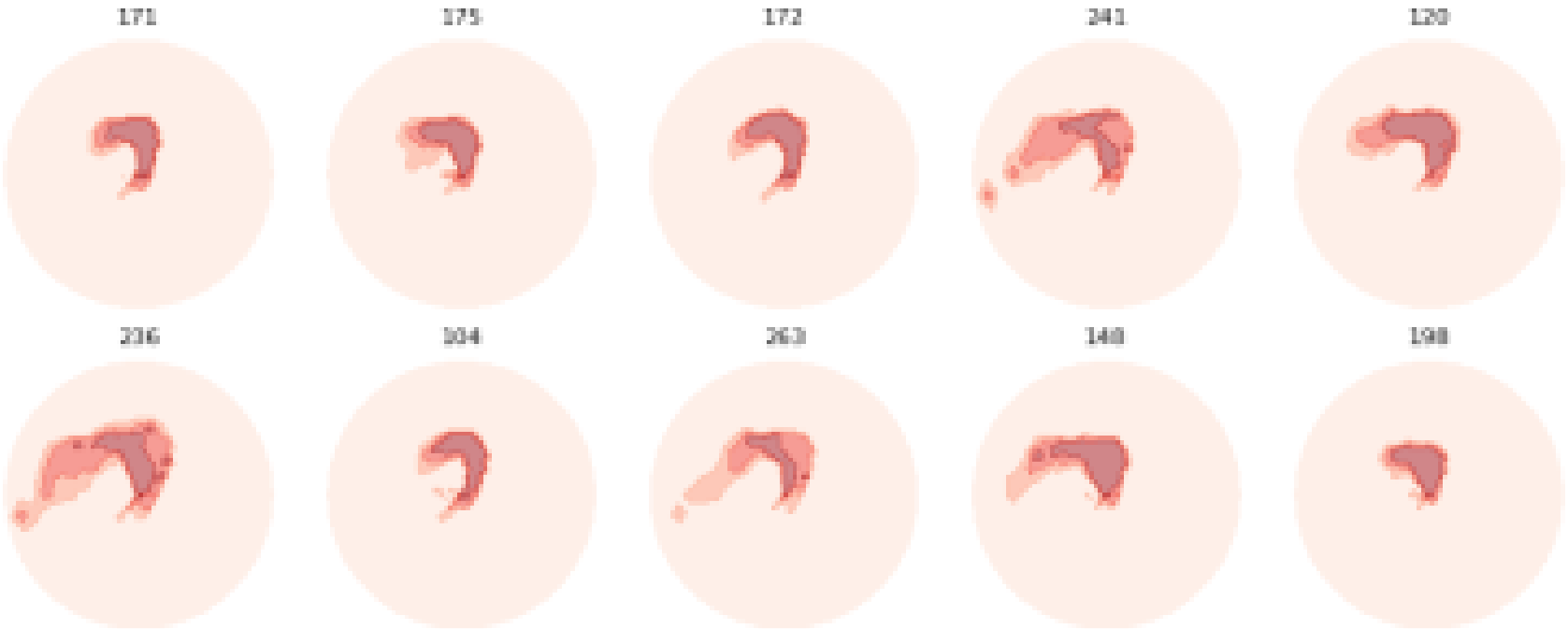


Figure – Nombre de clusters fixé à 8

Concernant le nombre de clusters permettant le plus de stabilité, on remarque que pour la méthode MAA, la stabilité quasi parfaite s'obtient pour un nombre de clusters supérieur strictement à 6 et pour la méthode distance de Wasserstein, la meilleure stabilité s'obtient pour un nombre de clusters fixé à 8. Pour les deux autres méthodes, on ne peut dégager un nombre de clusters idéal étant donné leur instabilité sur l'ensemble des valeurs testées.

Clustering



Application

- Accident à la centrale
- Evaluation du scénario
- Calculs de multiples simulations
- Clustering
- Evaluation des dégâts sur les barycentres
- Prises de décisions

Bilan

- Méthode de tris
- Synthèse de l'incertitude
- Pistes :
  - → Distance learning
  - → Segmentation non-binaire

Bilan

- Méthode de tris
- Synthèse de l'incertitude
- Pistes :
  - → Distance learning
  - → Segmentation non-binaire

Ce stage à Phimeca fut une expérience formidable. J'ai pu y découvrir la vie d'entreprise et le travail d'ingénieur mathématicien au côté d'une équipe qui m'a chaleureusement accueilli. Tout au long du stage, j'ai notamment pu appliquer les savoirs et les qualités développés au cours de ma scolarité tout en enrichissant ma culture mathématique sur de nombreux sujets.

Sur le plan des mathématiques, je suis très satisfait des résultats obtenus même si j'aurais aimé pouvoir approfondir certaines pistes que je n'ai pas étudiées par manque de temps. Particulièrement, l'étude de la théorie "Distance learning" afin d'obtenir une distance différente de la distance de Wasserstein notamment vis-à-vis de la méthode MAA où on peut dégager une importance différentes entre les vecteurs de projections calculés ainsi de la "segmentation non-binaire" en ce qui concerne la méthode des auto-encodeur permettant de déterminer non pas une zone en particulier mais de multiples.