Plénière

Jeudi 30 janvier 2020

AVANTIX

Lexique

Vocabulaire	Définition
Base de données radar	Regroupement des caractéristiques des radars
Jeu de données	Regroupement de lots d'impulsions (regroupement de signaux)
Lot d'impulsions	Ensemble d'impulsions pouvant contenir un ou plusieurs émetteurs (1 signal)
Impulsion	Description d'un lot d'impulsions avec les infos TOA, dTOA, Level, Fréquence, Durée d'impulsion

Plan

1 Introduction

Introduction

- **Objectifs de la thèse :**

  - Introduire de l'intelligence artificielle dans le domaine du renseignement radar
  - Améliorer la séparation, la caractérisation et l'identification des radars par de nouvelles approches
  - Moderniser les méthodes de séparation de sources existantes et proposer de nouvelles approches automatisées dans un cadre non supervisé
  - Enrichir un référentiel data existant

Introduction

- **Objectifs des 6 derniers mois** :
  - Se familiariser avec les données
    - Exploration et manipulation des données
    - Introduction au traitement du signal
    - Découverte des méthodes existantes (Etat de l'art)
  - Problématiques :
    - Déterminer le nombre d'émetteur(s) présent(s) dans un signal
    - Classifier chaque impulsion de ce signal par rapport au(x) émetteur(s) présent(s)

2

Etat de l’art

Etat de l’art

- Travaux précurseurs sur l’identification de radars :

  - 1982 : « *Automatic processing for ESM* » - Davies & Hollands
  - 1985 : « *Signal sorting in ESM systems* » - Whittall
  - 1990 : « *Radar signal categorization using a neural network* » - Anderson
  - 1998 : « *A comparison of self-organizing neural networks for fast clustering of radar pulses* » - Granger
  - 2004 : « *Radar emitter recognition using intrapulse data* » - Kawalec & Owczarek

Etat de l’art

- 2006 : « *Robust estimation of radar pulse modulation* » - Lunden & Koivunen
- 2013 : « *Radar emitter signals recognition and classification with feddforward networks* » - Petrov, Jordanov & Roe
- 2017 : « *Automatic intrapulse modulation classification of advanced LPI radar waveforms* » - Kishore & Rao
- 2019 : « *Radar emitters classification and clustering with a scale mixture of normal distributions* » - Révillon
- 2019 : « *A generalized multivariate student-t mixture model for bayesian classification and clustering of radar waveforms* » - Revillon

Etat de l’art

- De nombreuses similitudes avec d’autres domaines:
  - Milieu maritime
    - Ex : Communication des cétacés (Grand dauphin, cachalot etc)
  - Milieu animalier
    - Ex : Communication chez les chauves-souris
  - Milieu acoustique
    - Ex : identification de musique via l’application shazam

3 Avancements

1) Les données

- **Utilisation de plusieurs sources de données**

  - Jeux de données simulées :

    - Données simulées par Avantix
    - Données labelisées regroupant les principales PDW avec des signaux mono et multi-capteurs

  - Jeux de données réelles : les données sont anonymisées

    - Jeu de données 1 : données fournies par Avantix
    - Jeu de données 2 : données provenant de relevés effectués en Norvège

  - Les Formes d'ondes :

    - Jeu de données fournies par la DGA
    - Chaque signal caractérise un cas mono-radar

1) Les données

Dans la suite de cette présentation tous les graphiques seront construits à partir des jeux de données des Formes d'Ondes ou des données réelles

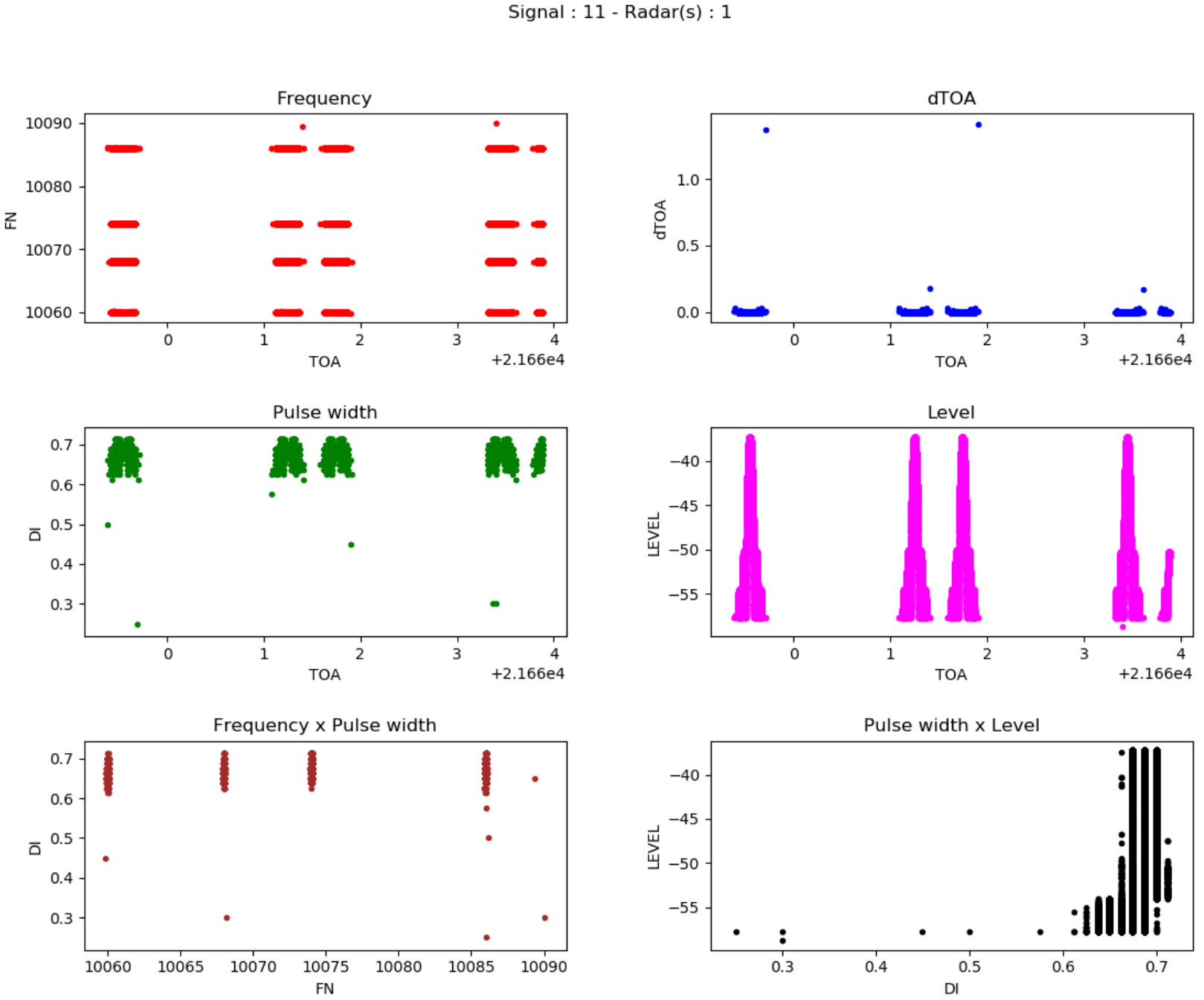
- **Caractéristiques des jeux de données :**

  - Chaque jeu de données regroupe plusieurs caractéristiques sur un signal : Fréquence, Durée d'impulsion, Niveau, TOA / dTOA

  - Selon leur provenance certains jeux de données sont :

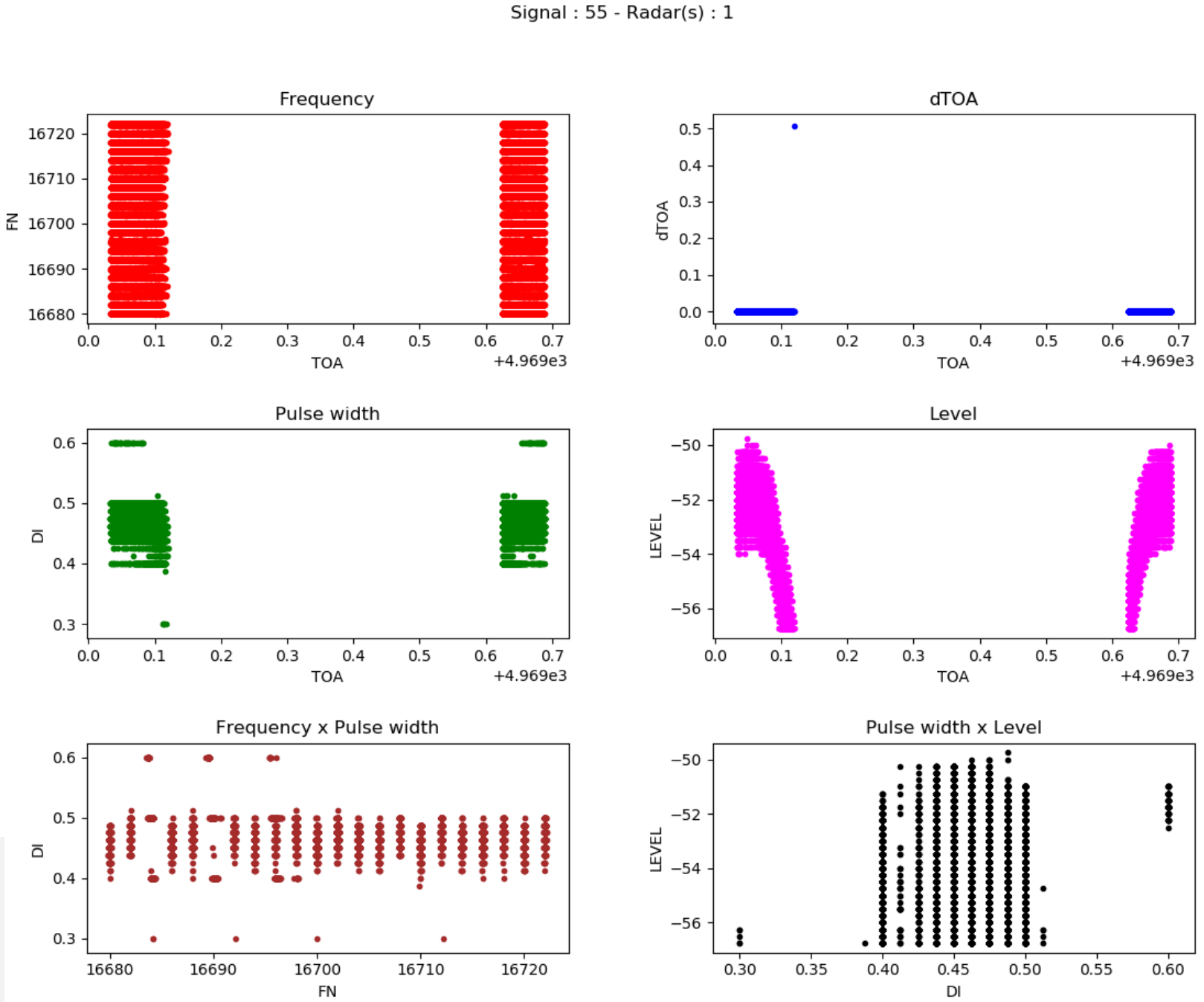
    - Semi-labelisés : jeux de données réelles
    - Entièrement labélisés : jeux de données simulées + FO

1) Les données



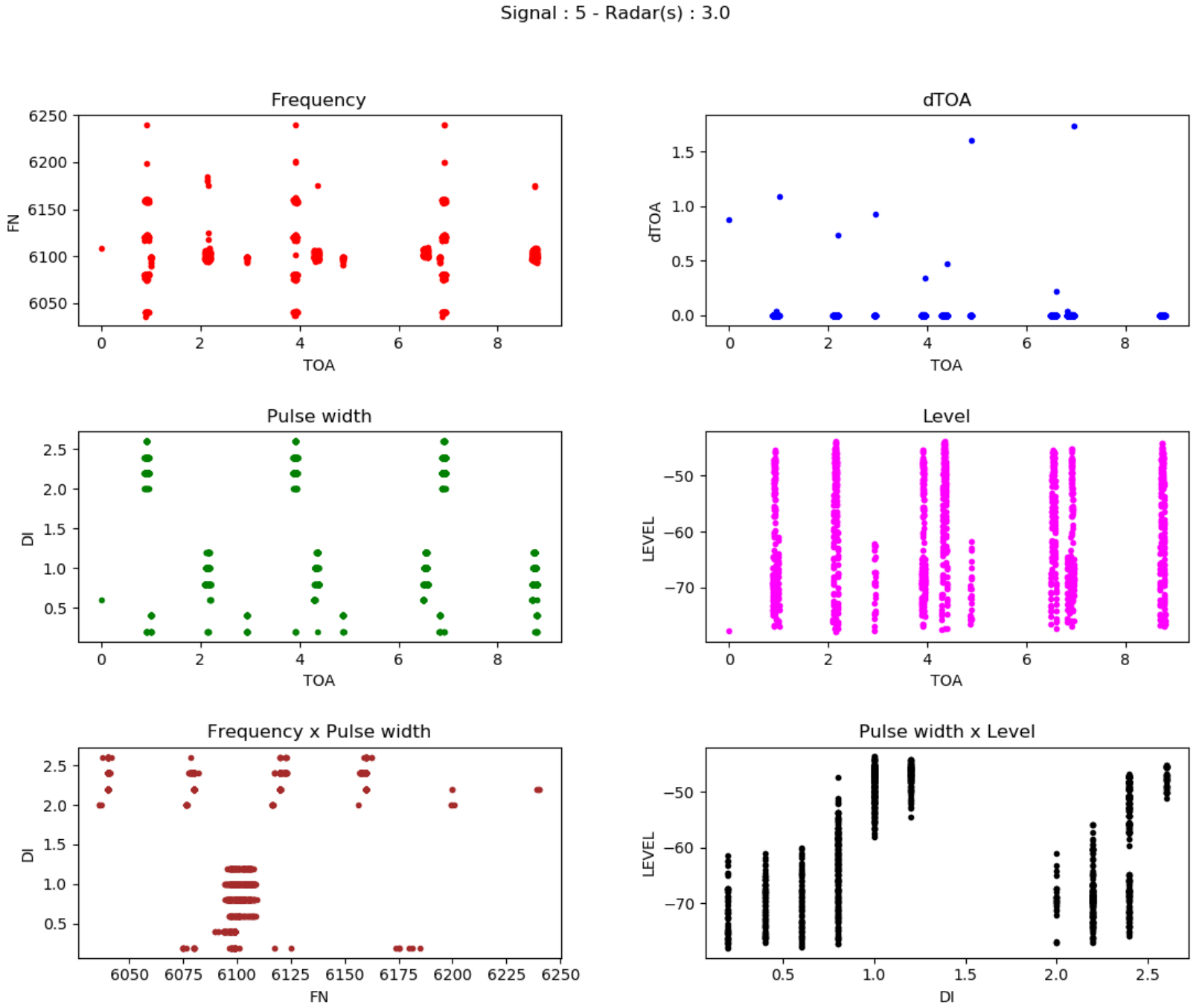
Représentation d'un signal

1) Les données



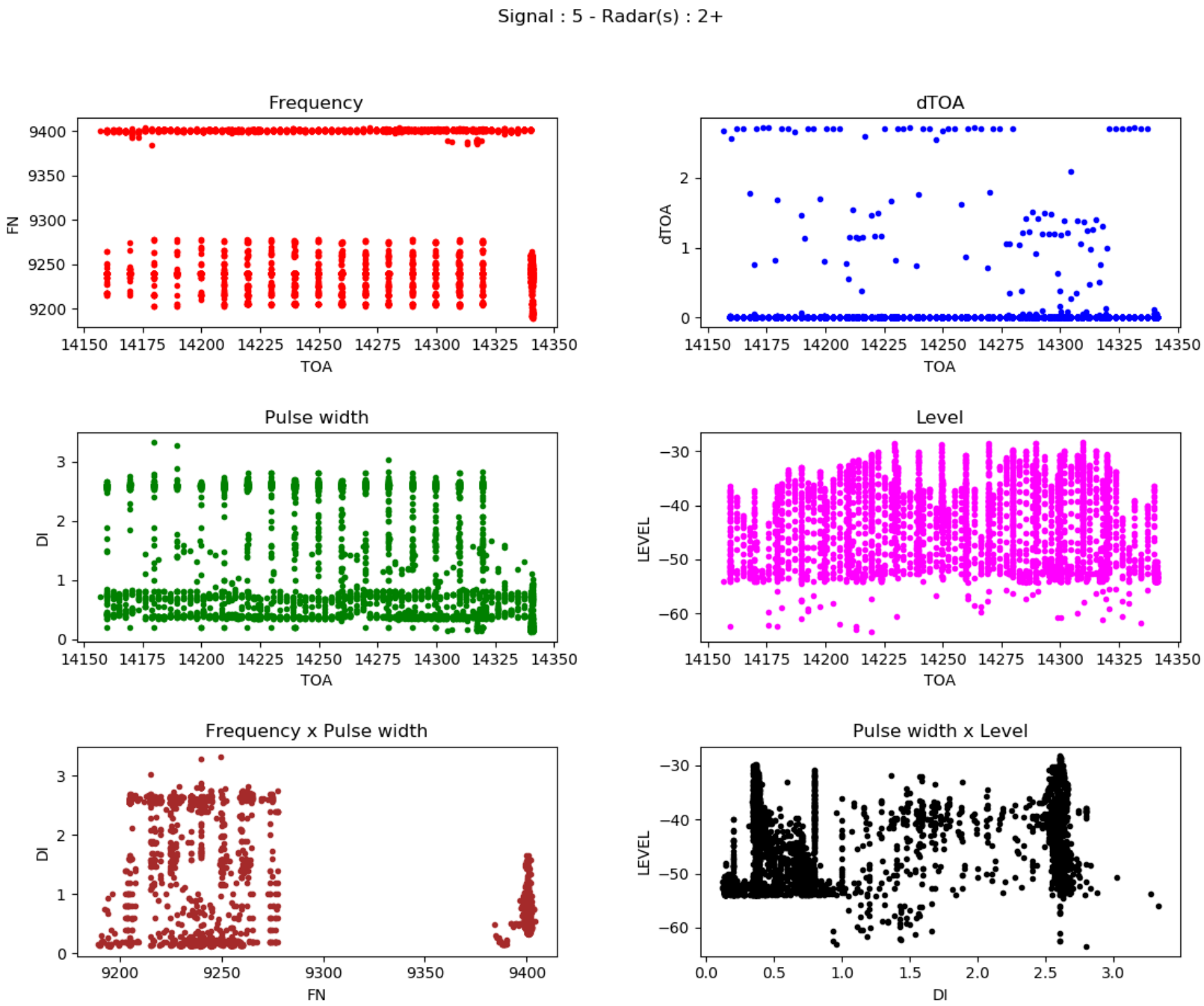
Représentation d'un signal

1) Les données



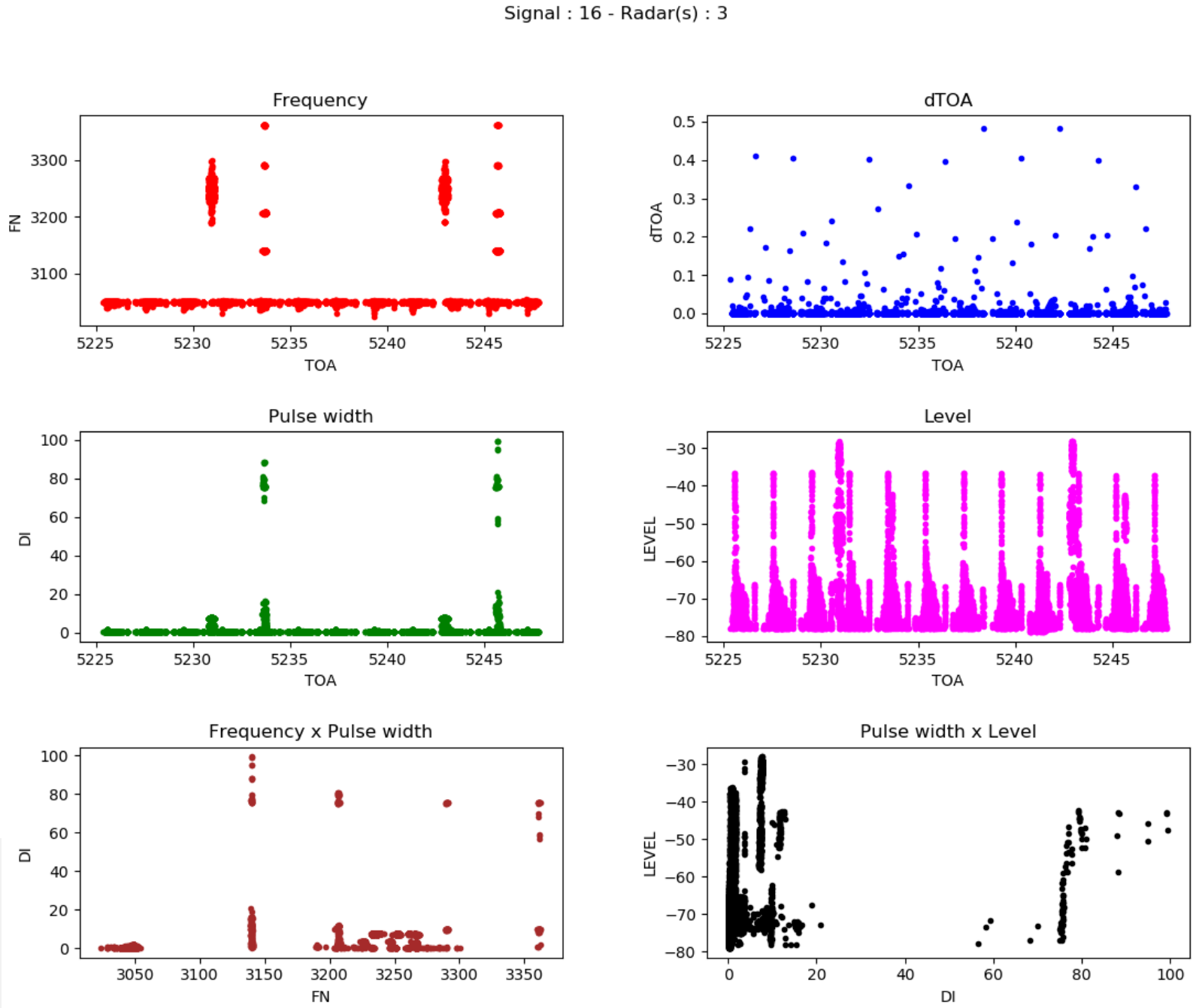
Représentation d'un signal

1) Les données



Représentation d'un signal

1) Les données



Représentation d'un signal

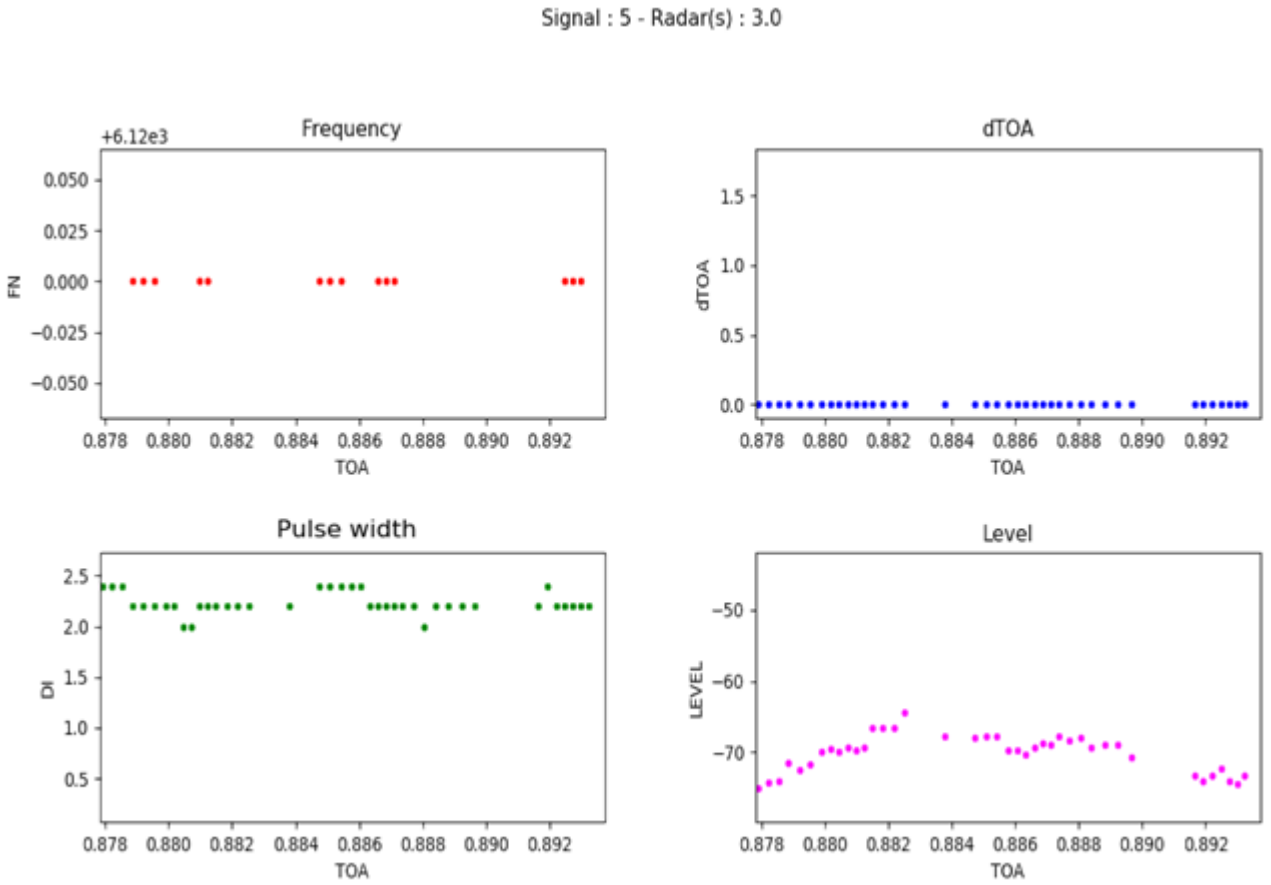
1) Les données

❖ Pré-traitements

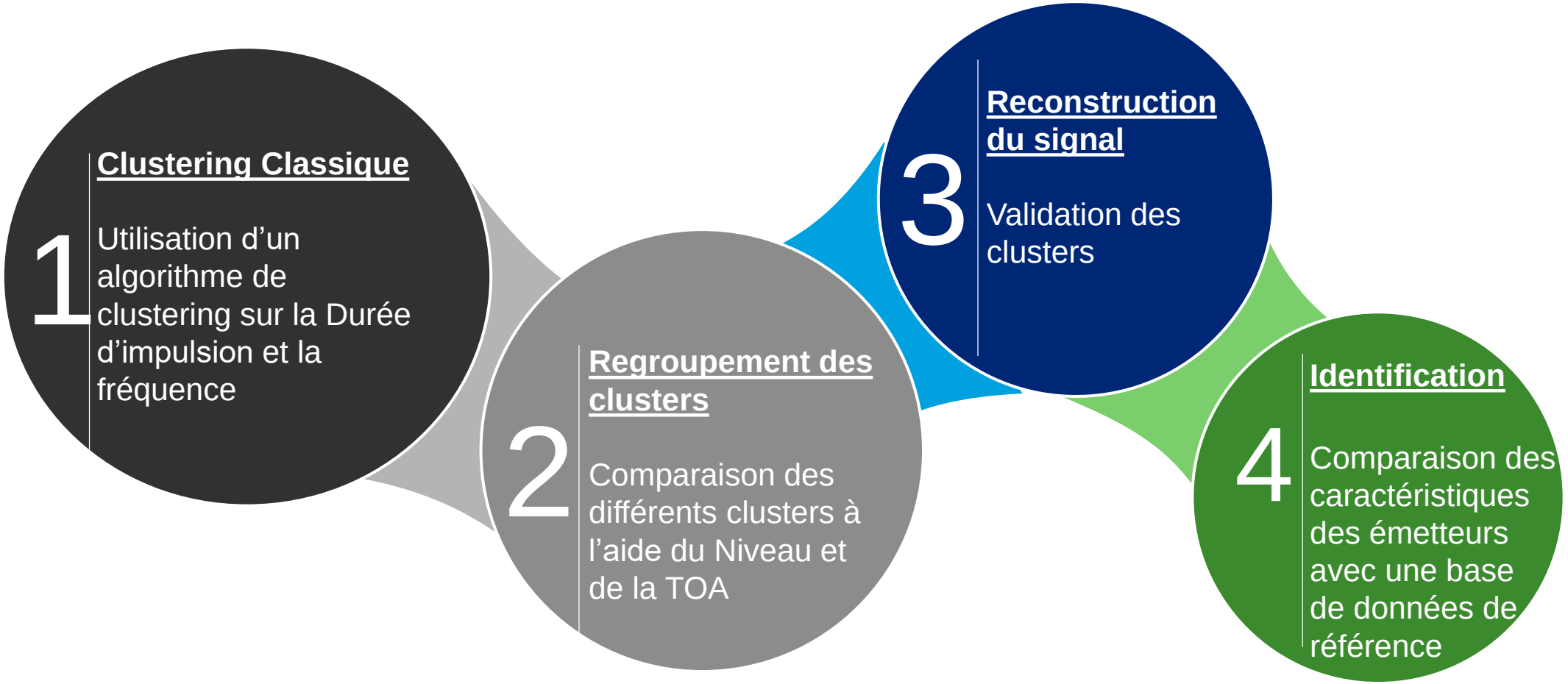
- Modification des valeurs de la Durée d'impulsion et de la fréquence du jeu de données :

  - **Problème :**
    - ➢ Erreur de quantification des données
    - ➢ Les données sont trop espacées

  - **Solution :**
    - ➢ Récupération de la plus petite valeur du pas de quantification pour chaque variable
    - ➢ Création d'un échantillon uniformément réparti sur l'intervalle +/- du pas de quantification
    - ➢ Ajout de ces valeurs à celles de la variable



2) Stratégie de désentrelacement



a) Phase 1 : clustering

- **Méthodes testées : DBSCAN** *(Density-based spatial clustering of application with noise) :*
  - Vocabulaire :
    - **Eps** : Rayon/distance qui forme une région sphérique autour de chaque point (nombre positif)
    - **MinPts** : Nombre de points minimum devant se trouver dans le rayon Eps pour que ce points soit considéré comme un cluster
    - **Eps-voisinage de x** : Ensemble des points dont la distance à x est inférieure à Eps
      - **Core point** : Point dont l'Eps-voisinage contient au moins MinPts points
      - **Border point** : Point dont l'Eps-voisinage cointient moins de MinPts
      - **Noise Point** : Point qui n'est ni un border point ni un core point

a) Phase 1 : clustering

- **Méthodes testées : DBSCAN *(Density-based spatial clustering of application with noise)* :**
  - Fonctionnement :
    - Sélection arbitraire d'un point et construction de l'ensemble des points atteignables par densité puis analyse :
      - Si Eps-voisinage > MinPts alors Dbscan reconsidère chaque point de l'Eps-voisinage jusqu'à ne plus pouvoir agrandir le cluster
      - Sinon ce point est considéré comme du bruit
    - Parcours de l'Eps-voisinage de proche en proche pour trouver l'ensemble des points des clusters
    - Arrêt de l'algorithme lorsque tous les points sont étiquetés

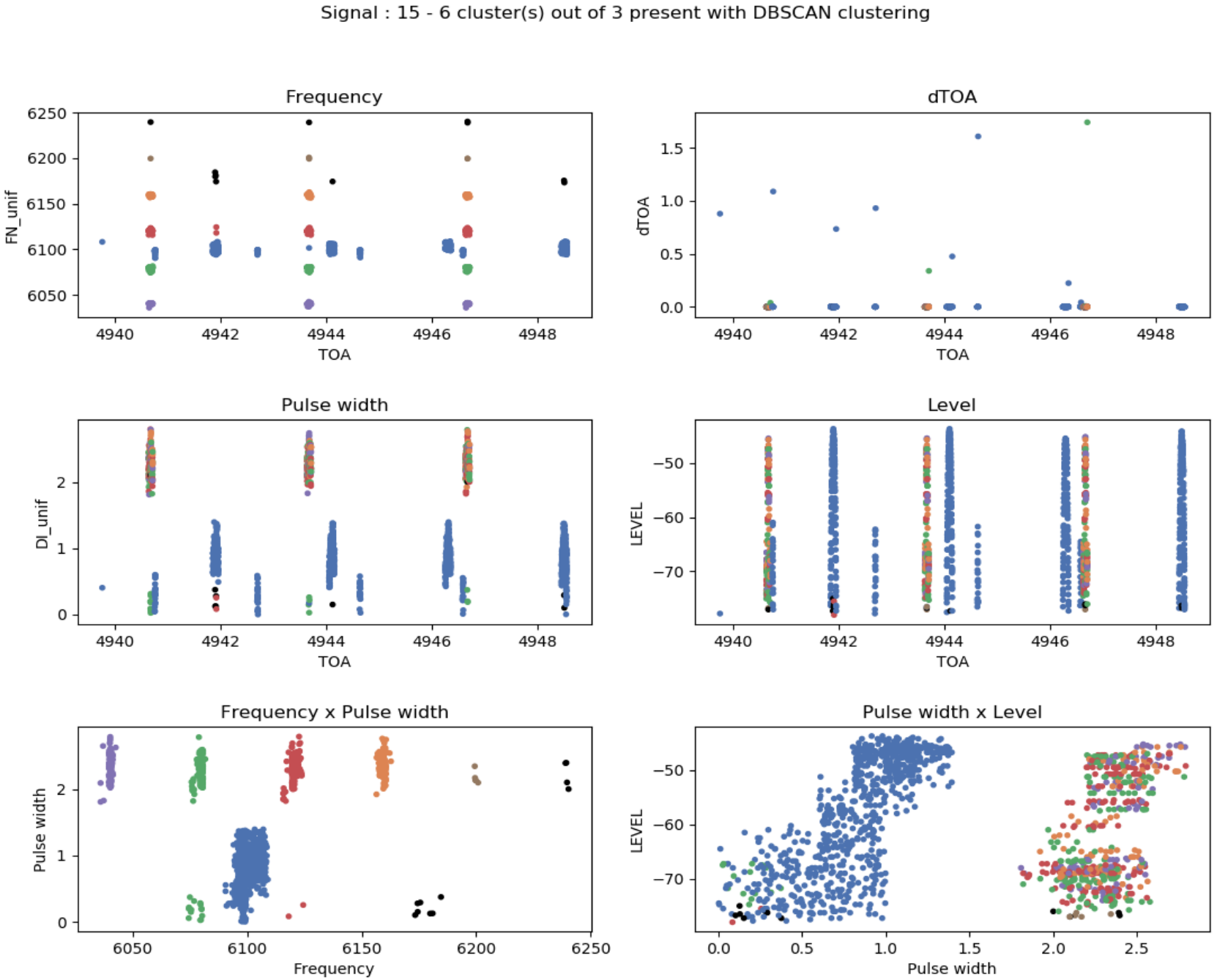
a) Phase 1 : clustering

❖ **Méthodes testées : DBSCAN *(Density-based spatial clustering of application with noise) :***

- **Avantages :**
  - 2 paramètres à régler : Eps & MinPts
  - Méthode de clustering par densité qui permet de détecter n'importe quelle forme de cluster
  - Robuste aux valeurs aberrantes et permet de détecter les outliers
  - Inutile de fixer au préalable le nombre de clusters

- **Inconvénients :**
  - Difficile à utiliser en grande dimension
  - Ne permet pas de détecter des clusters de densités différentes

Dbscan


Signal : 15 - 6 cluster(s) out of 3 present with DBSCAN clustering
Frequency
FN_unif
TOA
dTOA
dTOA
TOA
Pulse width
DI_unif
TOA
Level
LEVEL
TOA
Frequency x Pulse width
Pulse width
Frequency
Pulse width x Level
LEVEL
Pulse width

a) Phase 1 : clustering

❖ **Méthodes testées : OPTICS (*Ordering Points To Identify The Clustering Structure*) :**

- Vocabulaire :
  - **Eps' :** L'Eps minimal afin de faire d'un point x un core point (MinPts fixé)
  - **Core distance**: distance calculée pour un point :
    - Si Eps-voisinage(x)> MinPts alors core-distance(x) = Eps'(x)
    - Sinon core-distance(x) = indéfinie
  - **Reachability-distance**: matrice de distance calculée entre les points :
    - Si Eps-voisinage(x) > MinPts alors reachability-distance(x, y) = max{core-distance(x), d(x, y)}
    - Sinon indéfini

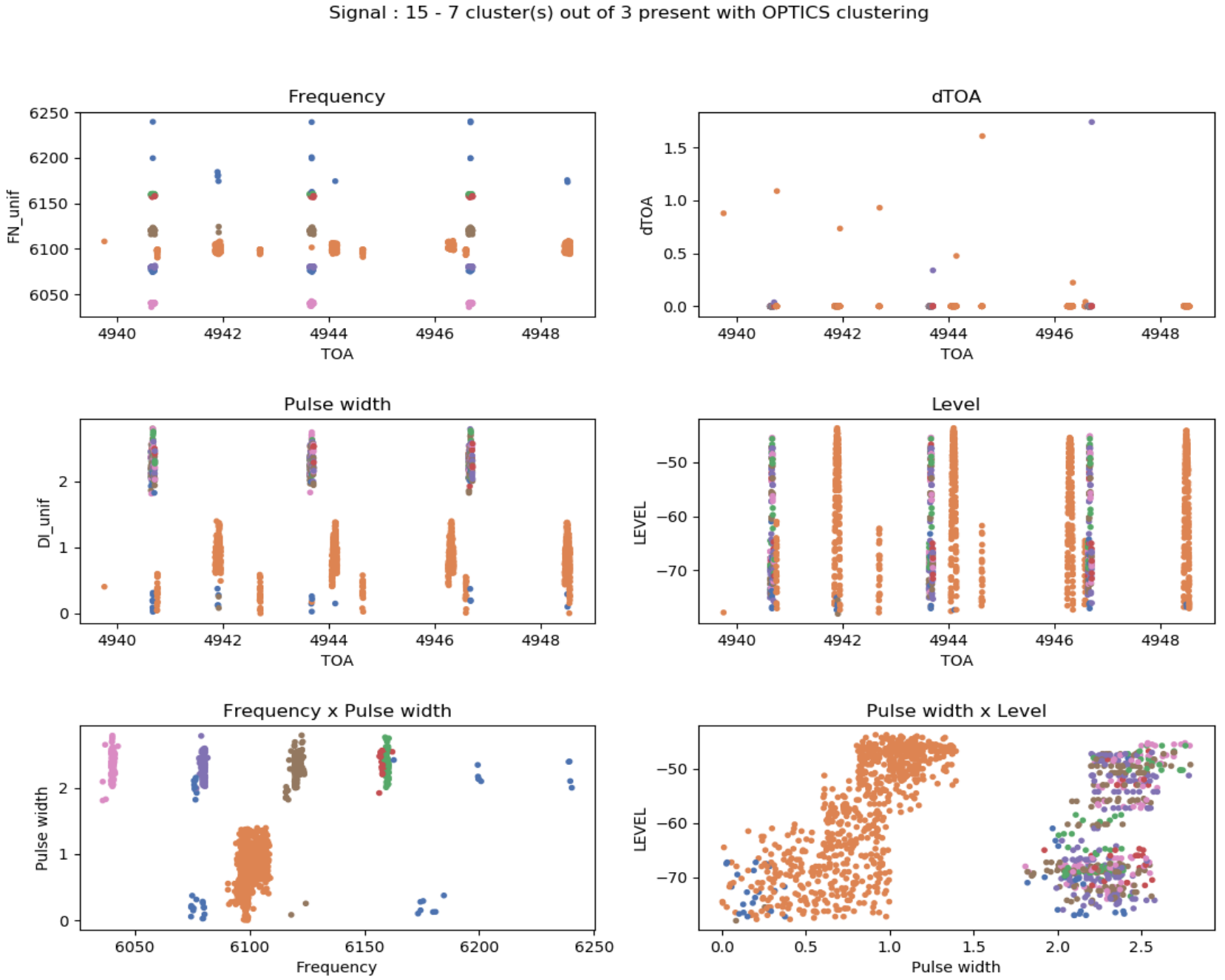
a) Phase 1 : clustering

- **Méthodes testées : OPTICS (*Ordering Points To Identify The Clustering Structure*) :**
  - Basée sur Dbscan mais autorise la recherche de clusters de densités différentes
  - Actuellement utilisé chez Avantix
  - <u>Fonctionnement</u> :
    - Calcul de cores-distances de l'ensemble des points
    - Calcul de la reachability-distance entre les points
    - Parcours du voisinage de proche en proche
    - Création d'un diagramme d'accès ordonné des distances pour détecter les clusters

a) Phase 1 : clustering

- **Méthodes testées : OPTICS** (*Ordering Points To Identify The Clustering Structure*):
  - **Avantages :**
    - 1 seul paramètre à régler : MinPts (Eps est facultatif)
    - Permets de détecter des outliers
  - **Inconvénients :**
    - Long à exécuter
    - Ordre des points important

Optics


Signal : 15 - 7 cluster(s) out of 3 present with OPTICS clustering
Frequency
FN_unif
TOA
dTOA
dTOA
TOA
Pulse width
DI_unif
TOA
Level
LEVEL
TOA
Frequency x Pulse width
Pulse width
Frequency
Pulse width x Level
LEVEL
Pulse width

a) Phase 1 : clustering

- **Méthodes testées : Kmeans :**

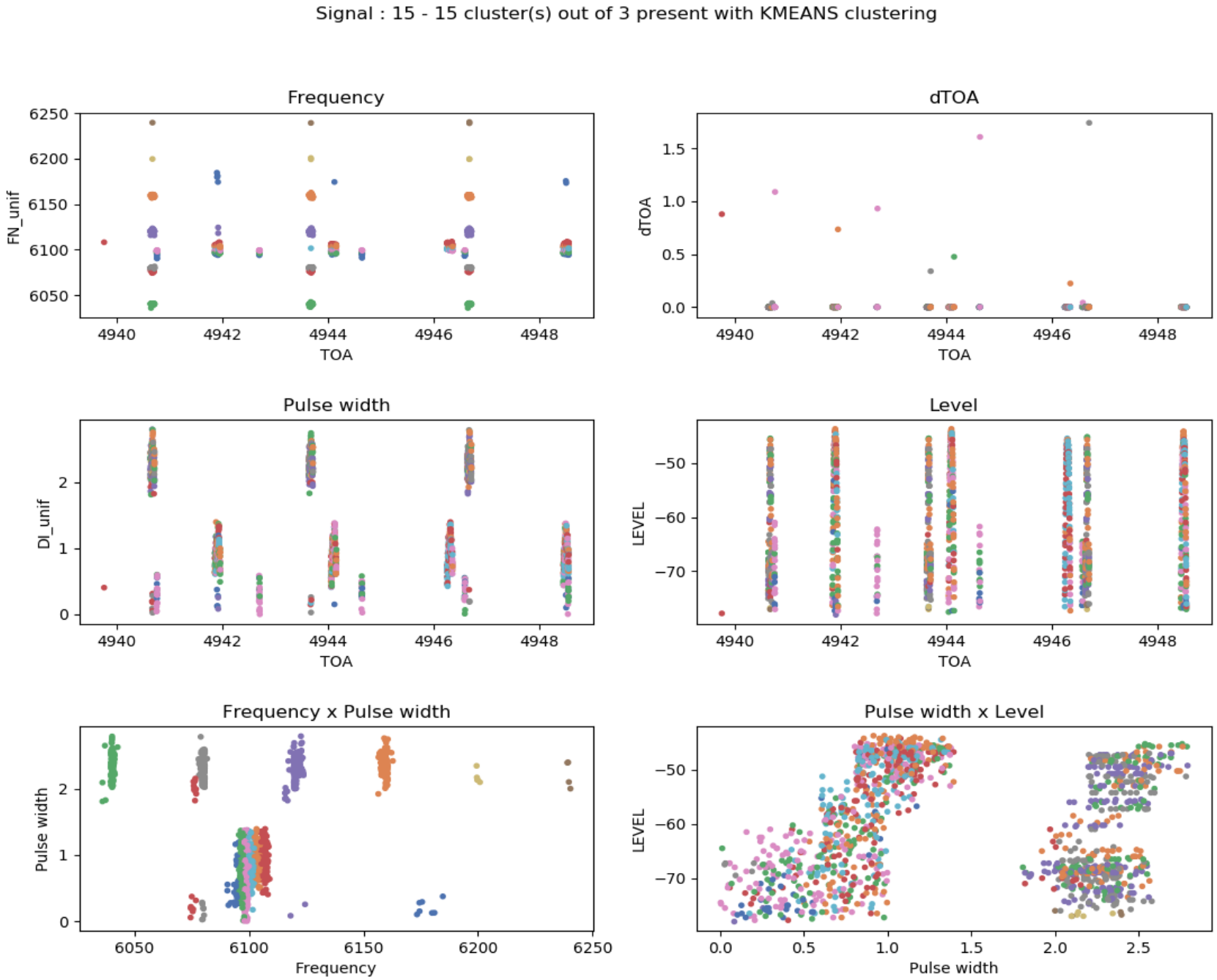
  - <u>Fonctionnement :</u>

    - Sélection aléatoire de K observations à partir des données comme centres de classes
    - Affectation des observations à leurs centres de classes le plus proche afin d'obtenir K clusters
    - Recalcul des centres de classes à partir des observations rattachées
    - Répétition des étapes jusqu'à convergence de l'algorithme

a) Phase 1 : clustering

- **Méthodes testées : Kmeans :**
  - **Avantages :**
    - Efficace sur les gros jeux de données
    - Facile à interpréter et rapide à exécuter
  - **Inconvénients :**
    - Fixation manuelle du nombre de clusters à trouver
    - Sensible aux valeurs extrêmes et à l'initialisation des centres de classes
    - Ne détecte pas n'importe quelle forme de cluster
    - Fonctionne très bien lorsque les clusters sont sphériques et de même taille

Kmeans


Signal : 15 - 15 cluster(s) out of 3 present with KMEANS clustering
Frequency
FN_unif
TOA
dTOA
dTOA
TOA
Pulse width
DI_unif
TOA
Level
LEVEL
TOA
Frequency x Pulse width
Pulse width
Frequency
Pulse width x Level
LEVEL
Pulse width

a) Phase 1 : clustering

❖ **Méthodes testées : GMM (*Gaussian mixture model*) :**

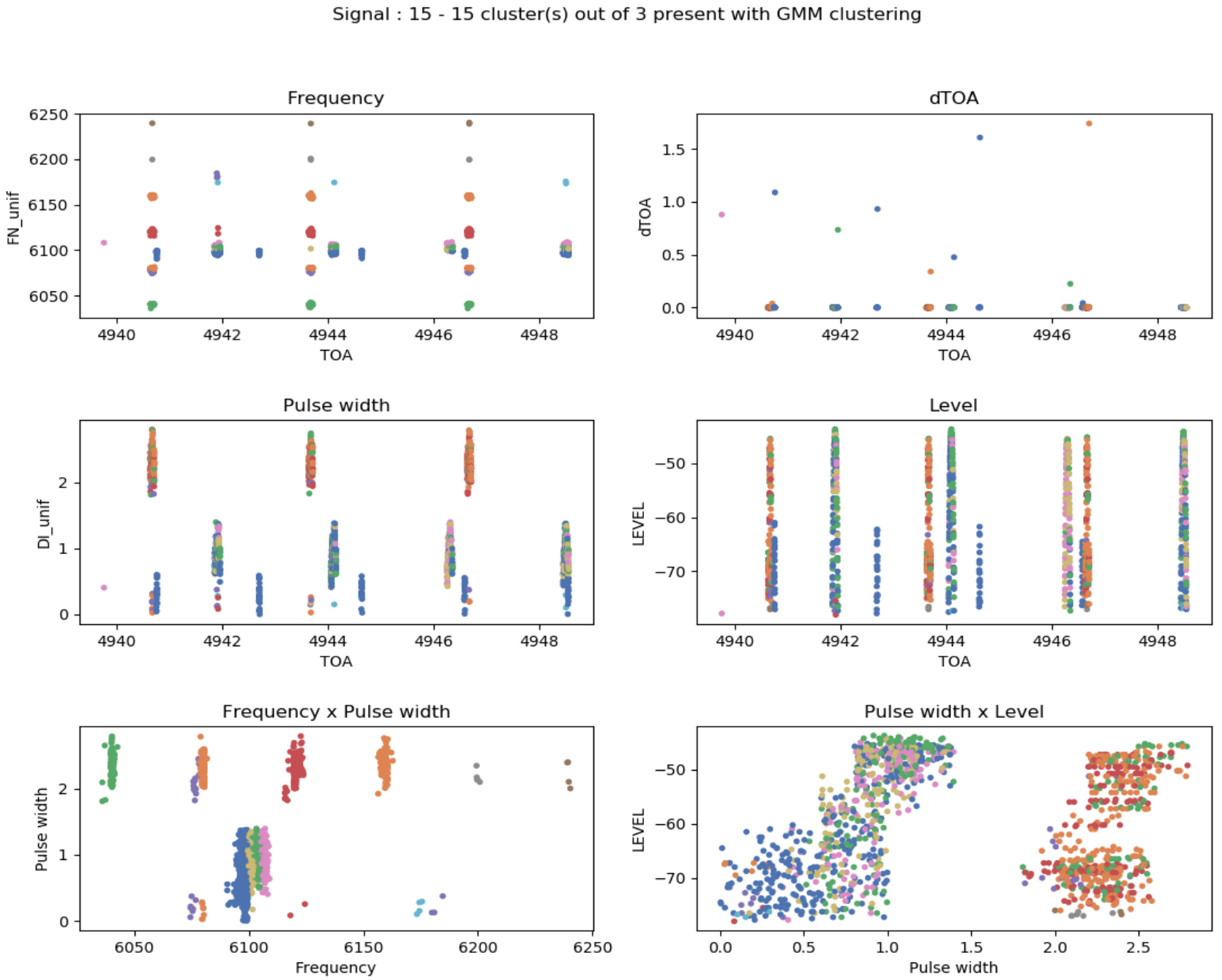
- Similaire aux K-means
- Soft-clustering

- <u>Fonctionnement</u> :
  - Fixation aléatoire d'hypothèses sur les paramètres des gaussiennes
  - Attribution pour chaque observation d'une probabilité/poids d'appartenance à chaque cluster
  - Mise à jour de leurs paramètres jusqu'à convergence

a) Phase 1 : clustering

- **Méthodes testées : GMM (*Gaussian mixture model*)**
  - **Avantages :**
    - Moins restrictif sur la forme des clusters retournés
  - **Inconvénients :**
    - Fixation du nombre de clusters
    - Difficile à interpréter
    - Limitation de la forme des clusters
    - Sensible au bruit

GMM


Signal : 15 - 15 cluster(s) out of 3 present with GMM clustering
Frequency
FN_unif
TOA
dTOA
dTOA
TOA
Pulse width
DI_unif
TOA
Level
LEVEL
TOA
Frequency x Pulse width
Pulse width
Frequency
Pulse width x Level
LEVEL
Pulse width

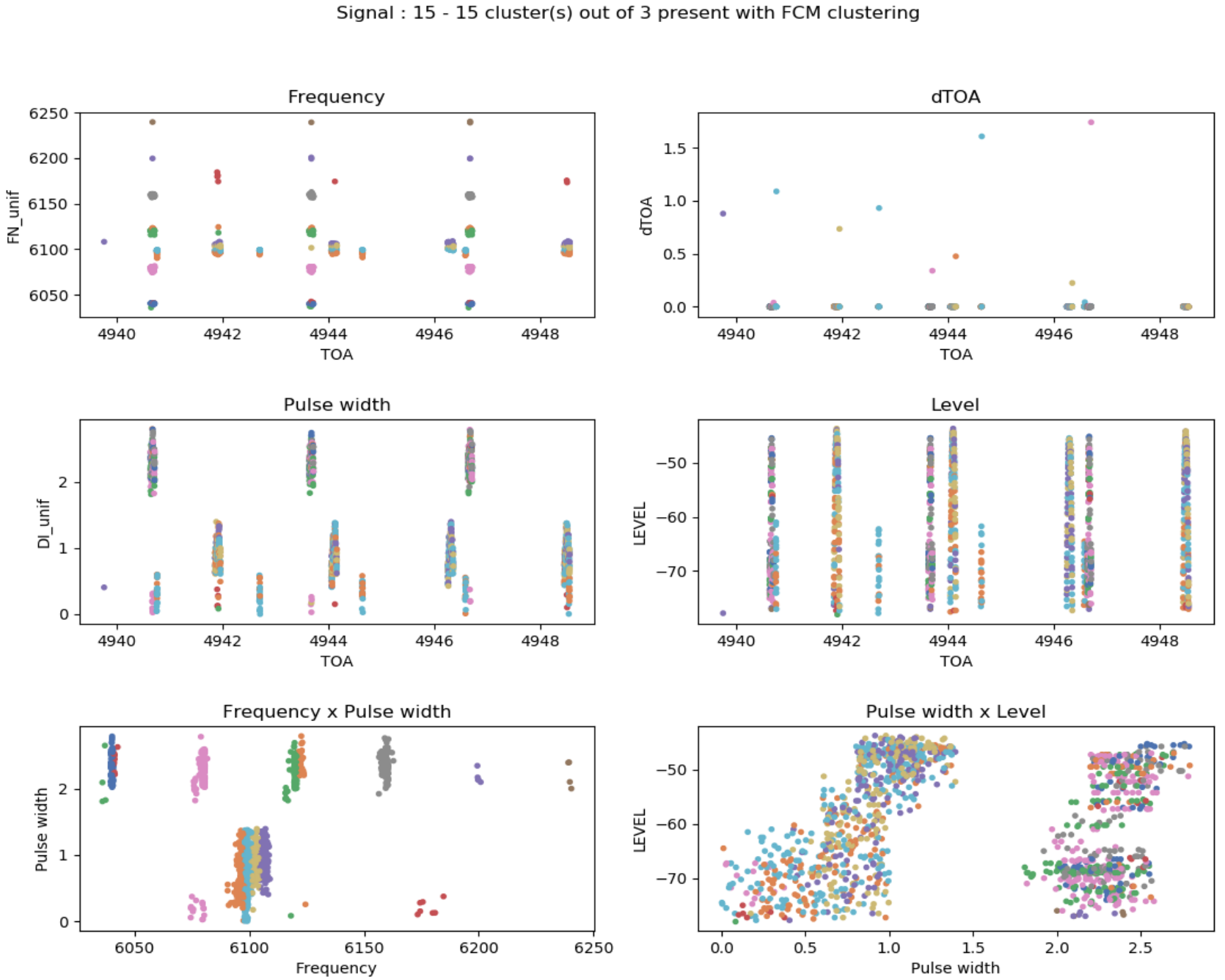
a) Phase 1 : clustering

- **Méthodes testées : FCM (*Fuzzy C-Means*) :**
  - Algorithme similaire à celui des K-means
  - Fonctionnement :
    - Sélection aléatoire de K centres de classes
    - Création d'une matrice de distance entre les observations et les centres de classes
    - Création d'une matrice d'appartenance à partir d'une fuzzy mesure utilisant les distances précédentes
    - Mise à jour des centres de classes et réaffectation des observations
    - Répétition des étapes précédentes jusqu'à convergence

a) Phase 1 : clustering

- **Méthodes testées : FCM (*Fuzzy C-Means*):**
  - **Avantages :**
    - Pas d'exclusivité d'appartenance : une observation peut appartenir à plusieurs clusters avec une certaine probabilité
  - **Inconvénients :**
    - Fixation du nombre de clusters au préalable
    - Temps d’exécution

FCM


Signal : 15 - 15 cluster(s) out of 3 present with FCM clustering
Frequency
FN_unif
TOA
dTOA
dTOA
TOA
Pulse width
DI_unif
TOA
Level
LEVEL
TOA
Frequency x Pulse width
Pulse width
Frequency
Pulse width x Level
LEVEL
Pulse width

a) Phase 1 : clustering

- **Méthodes testées : conclusions**
  - Les algorithmes classiques sont limités :
    - Forme des clusters
    - Densité
    - Fixation du nombre de clusters
    - Temps d'exécution
    - Paramétrage
      - Nécessité d'avoir recours à un algorithme de clustering plus souple : Hdbscan

a) Phase 1 : clustering

❖ **Méthode retenue : Hdbscan (*Hierarchical Dbscan*) :**

- Vocabulaire :
  - **Core distance**: distance avec le MinPts ième point le plus proche; distance minimale afin qu'un point soit considéré comme un core-point
  - **Mutual reachability distance** : Métrique permettant de sélectionner une distance entre les points
    - Mutual-reachability-distance(A,B) = max{core-distance(A), core-distance(B), d(A,B)}
  - **Arbre couvrant minimal**: issu de la théorie des graphes il s'agit de l'arbre reliant tous les points par le chemin le plus court
  - **Algorithme de Prim**: départ d'un arbre initial à un seul sommet et augmentation de la taille de cet arbre en ajoutant à chaque itération une connexion de poids minimale avec son plus proche voisin

a) Phase 1 : clustering

❖ **Méthode retenue : Hdbscan (*Hierarchical Dbscan*) :**

- Fonctionnement :
  - Transformation de l'espace selon les densités
    - ➢ Calcul des cores-distances de chaque point
    - ➢ Calcul de la mutual reachability distance
    - ➢ Construction de la matrice de distance entre les points
  - Transformation des données en un arbre pondéré (arbre couvrant minimal)
    - ➢ Les sommets représentent les points et les arêtes la mutual reachability distance
    - ➢ Récupération de l'arbre minimal via l'algorithme de Prim

a) Phase 1 : clustering

❖ **Méthode retenue : Hdbscan (*Hierarchical Dbscan*) :**

- Construction d'un clustering hiérarchique à partir de cet arbre

  - Triage des arêtes par ordre croissant
  - Affichage des résultats sous forme de dendrogramme

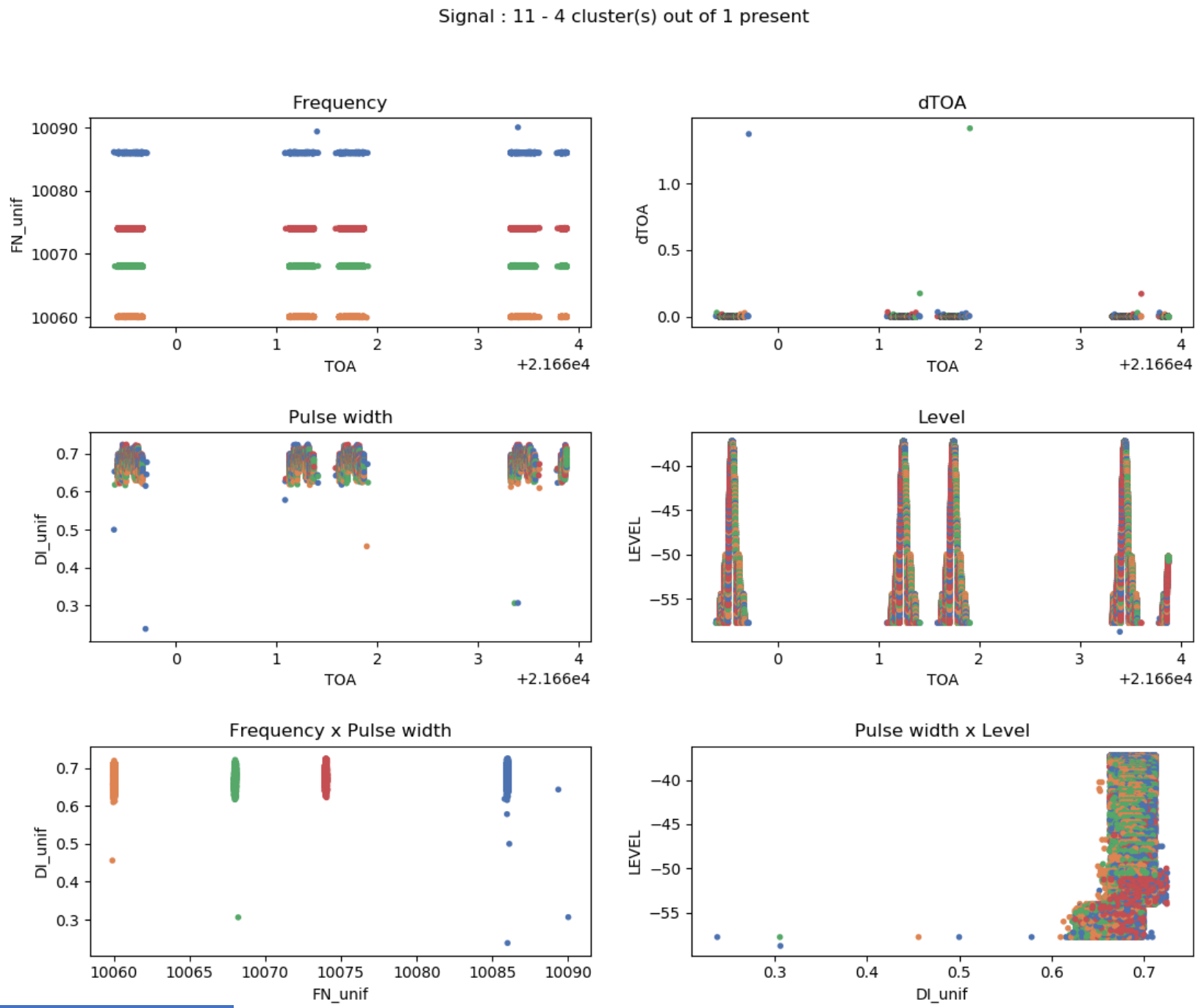
- Extractions des clusters

  - Calcul pour chaque point de la valeur seuil à partir de laquelle le point sort du cluster
  - Calcul d'une mesure de stabilité à partir des valeurs précédentes

a) Phase 1 : clustering

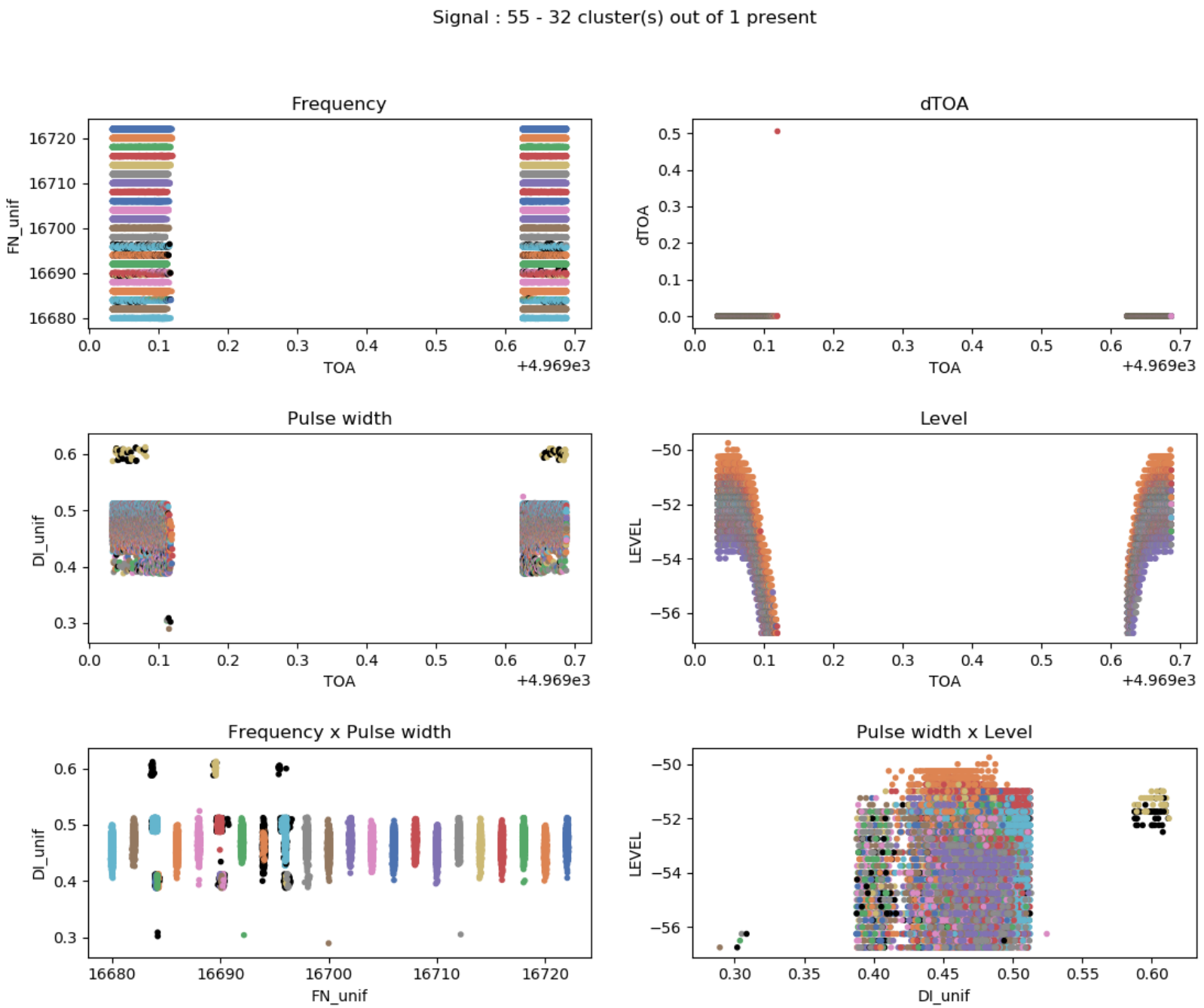
- **Méthode retenue : Hdbscan (*Hierarchical Dbscan*) :**
  - **Avantages :**
    - Exécute DBSCAN sur différentes valeurs d'Eps et intègre le résultat pour trouver une classification qui offre la meilleure stabilité sur Eps
    - Successeur d'OPTICS
    - Permets à DBSCAN de rechercher des clusters de densités variables
    - Moins de paramètres à optimiser

a) Phase 1 : résulats



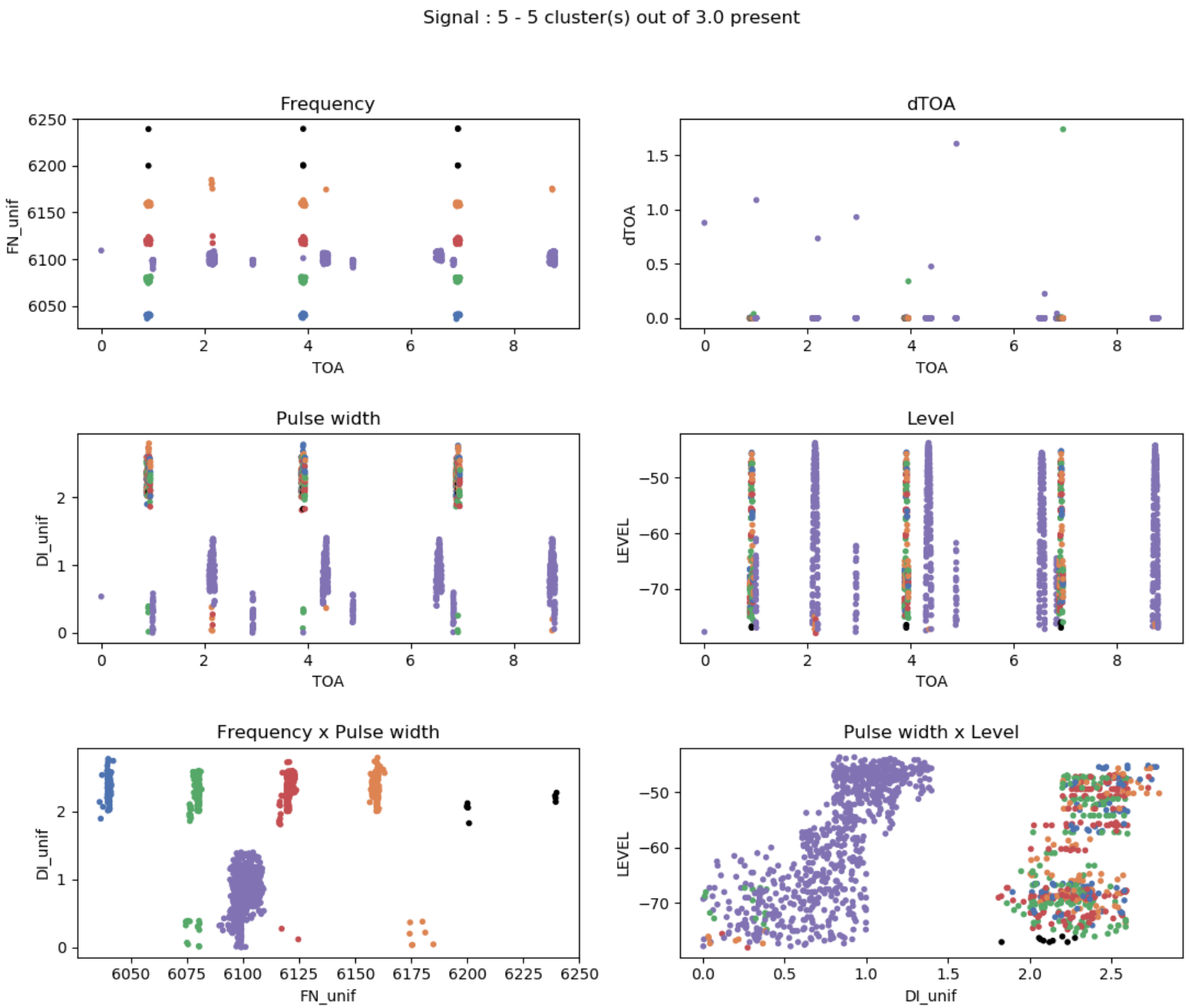
Résultats du clustering via Hdbscan

a) Phase 1 : résulats



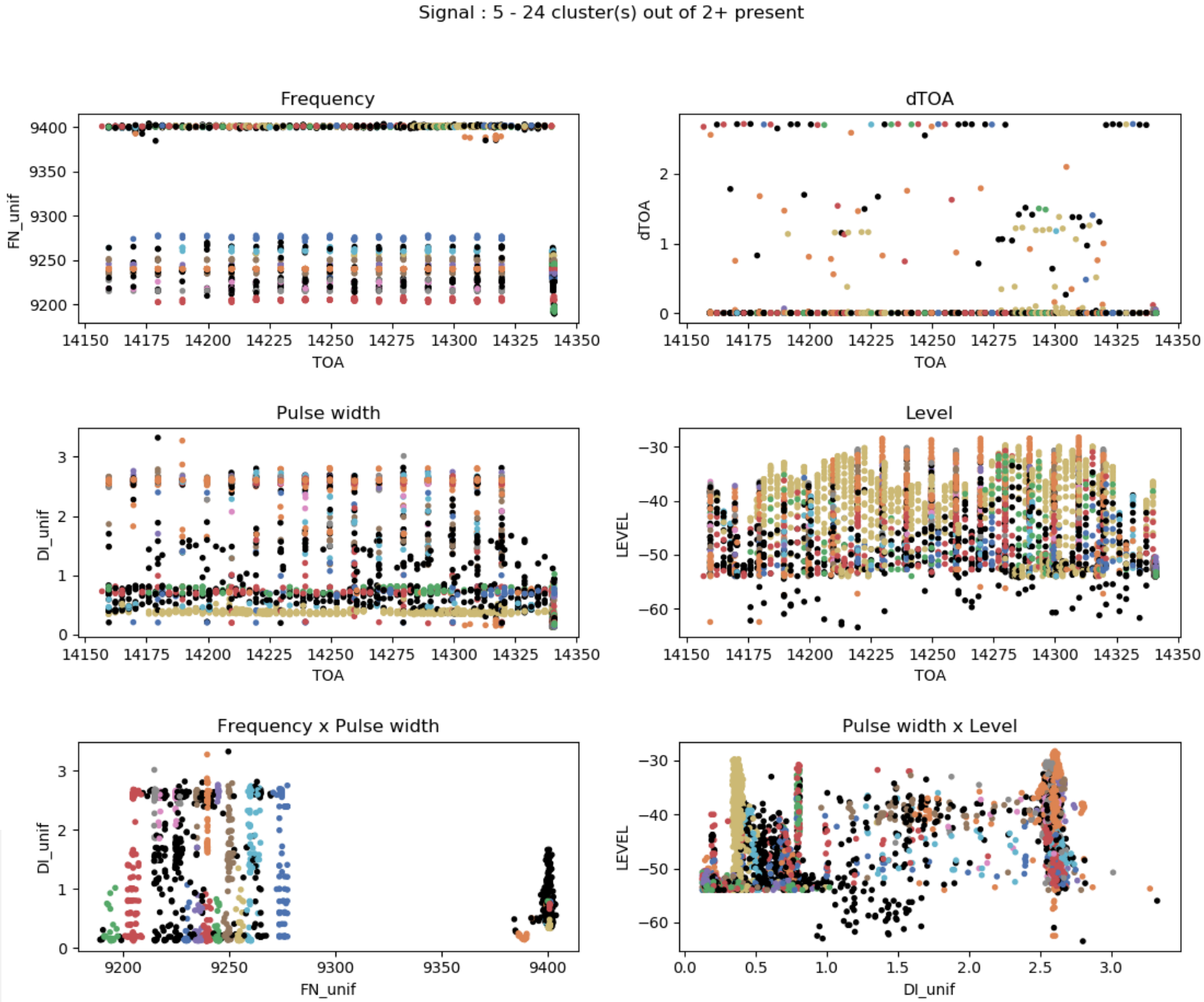
Résultats du clustering via Hdbscan

a) Phase 1 : résulats



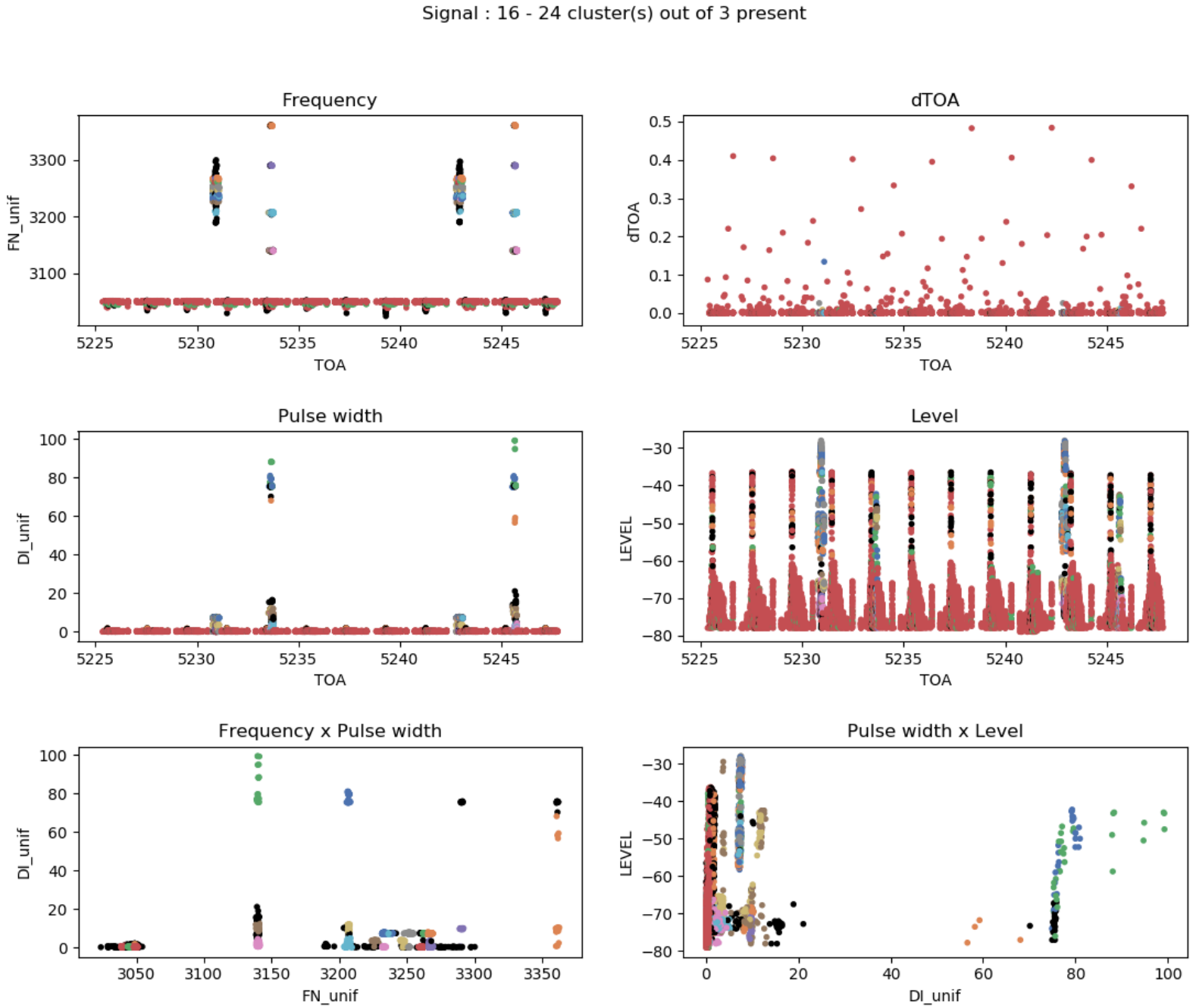
Résultats du clustering via Hdbscan

a) Phase 1 : résultats



Résultats du clustering via Hdbscan

a) Phase 1 : résultats



Résultats du clustering via Hdbscan

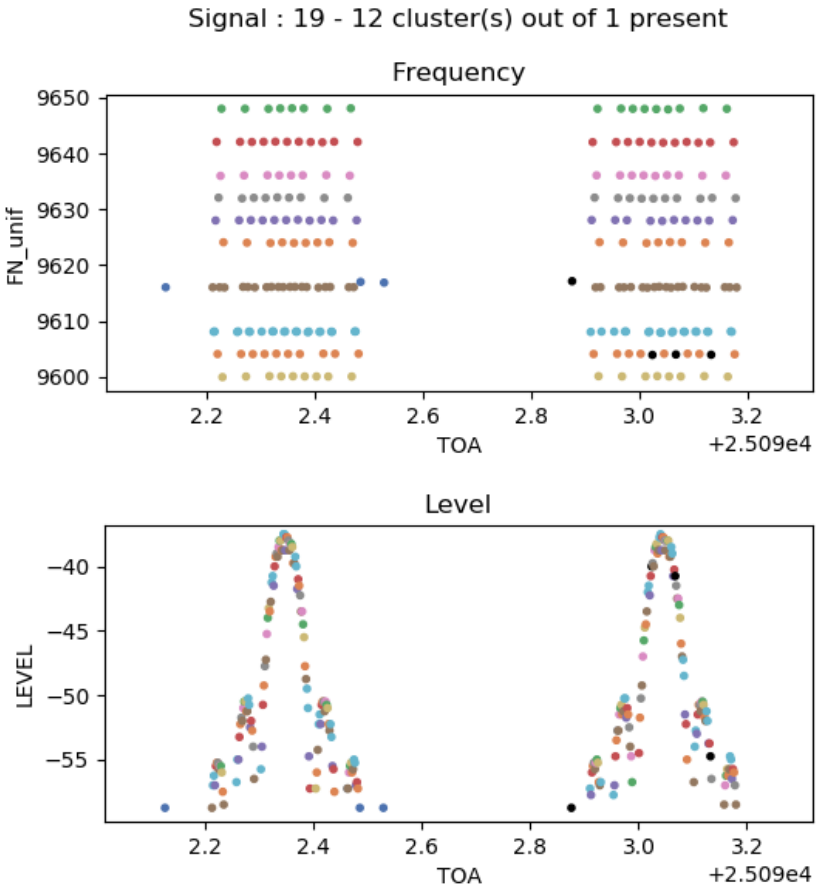
b) Phase 1 : Conclusions

- **Résultats :**
  - Algorithmes testés non adaptés pour la nature des données
  - Hdbscan fournit de meilleurs résultats
  - Surestimation du nombre de clusters dans certains cas

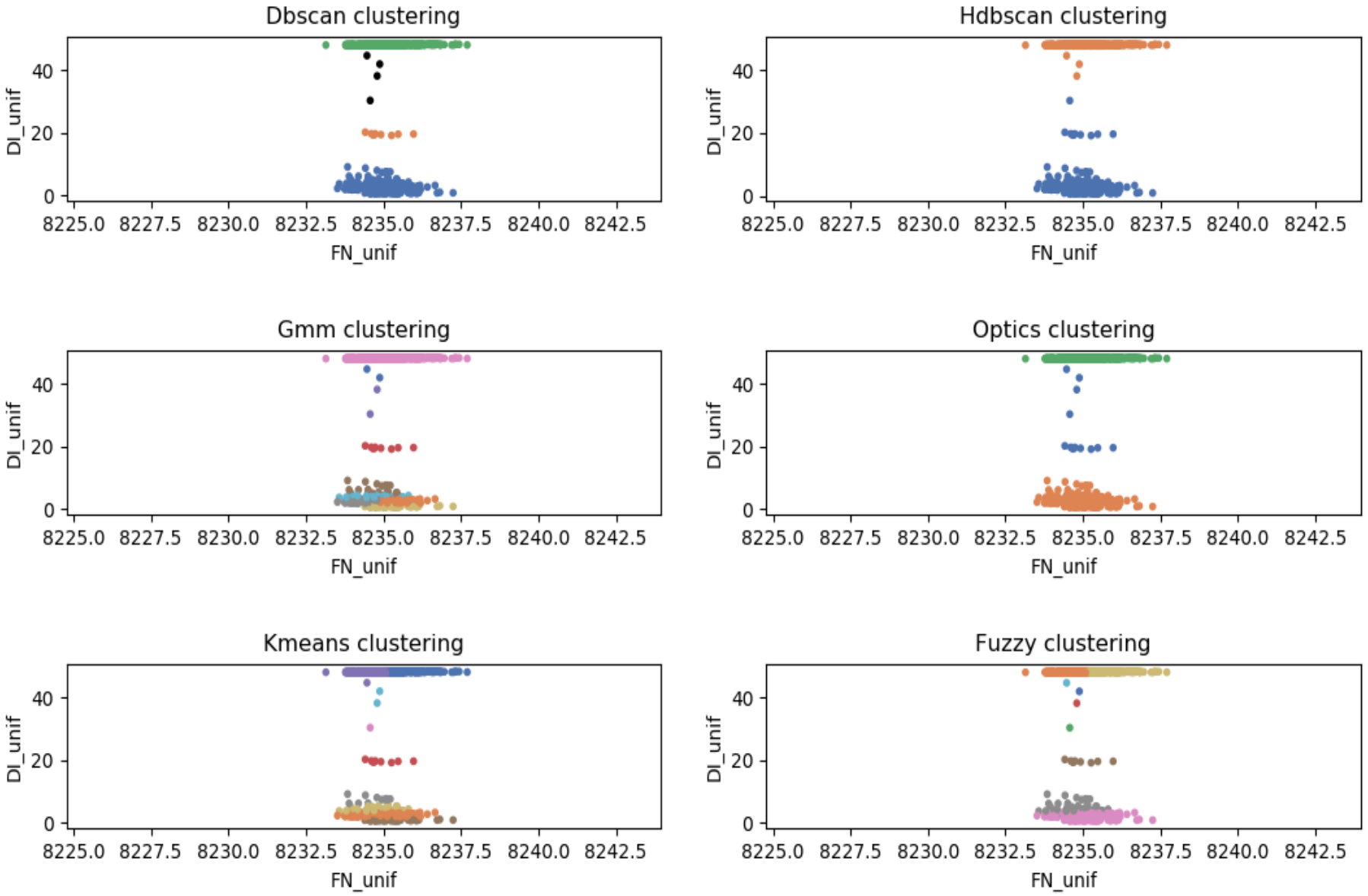
- **Problème :** Plusieurs clusters peuvent caractériser un seul radar

- **Solution :**
  - Effectuer des regroupements entre les clusters en comparant leurs informations relatives à l'aide du niveau
  - Calculer une distance entre les clusters

- **Méthodes testées en 2 étapes :**
  - Interpolation avec HAC classique
  - Transport optimal avec HAC modifiée

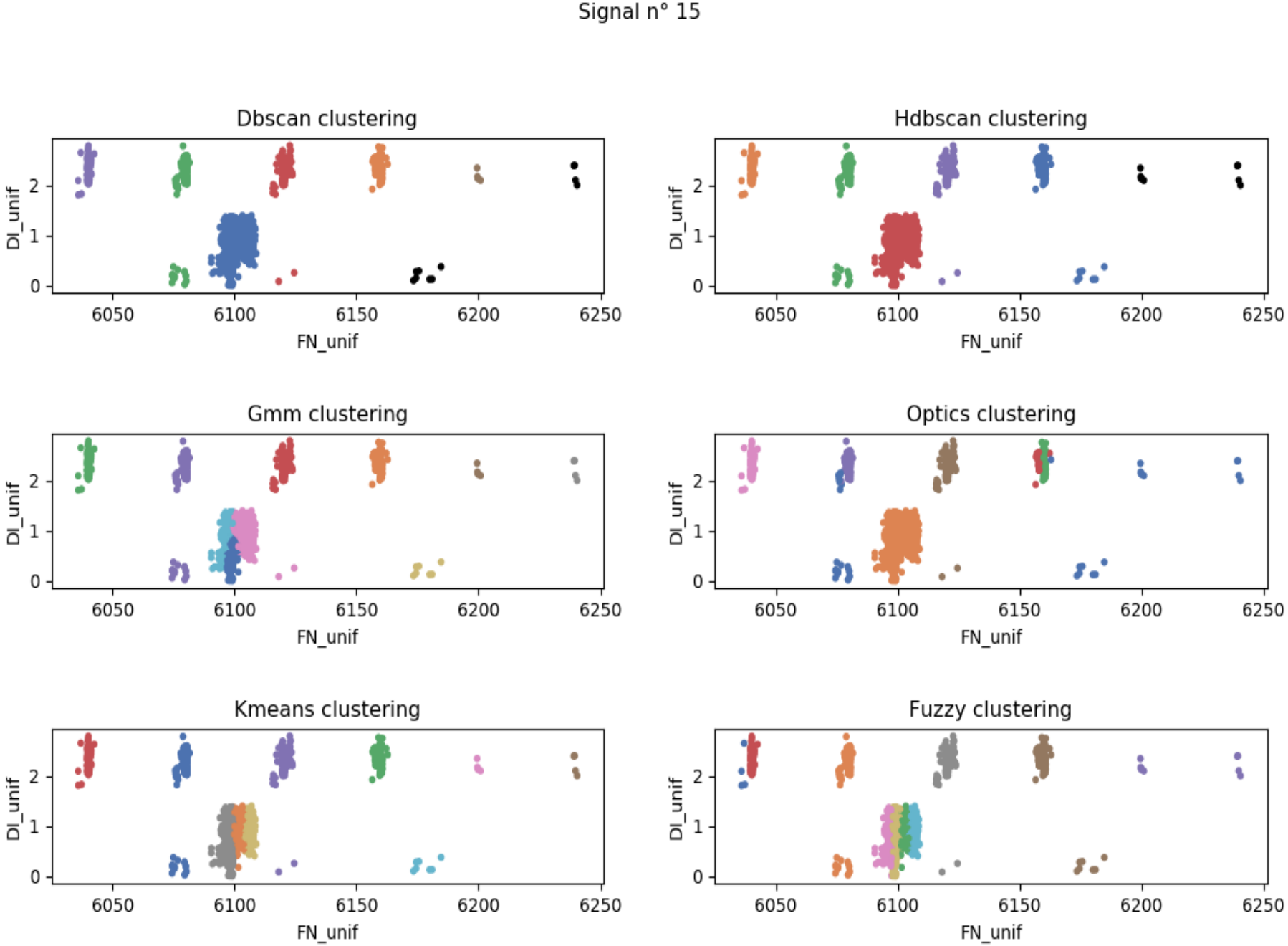


a) Phase 1 : clustering



Comparaison des méthodes testées

a) Phase 1 : clustering



Comparaison des méthodes testées

a) Phase 1 : Conclusions

Algorithme	Remarques	Interprétation	Exécution	Paramétrage	Autorise la recherche de clusters de densités différentes	Autorise la recherche de n'importe quelle forme de clusters	Fixation du nombre de clusters à trouver
DBSCAN	Difficile à utiliser en grande dimension	SIMPLE	RAPIDE	SIMPLE	NON	OUI	NON
OPTICS	Ordre des points importants	SIMPLE	RAPIDE	SIMPLE		OUI	NON
KMEANS	Hypothèse sur la forme et la taille des clusters	SIMPLE	RAPIDE	SIMPLE	NON	NON	OUI
GMM	Moins restrictif sur la forme des clusters retournés que le Kmeans	COMPLIQUEE	LONG	COMPLIQUE	NON	NON	OUI
FCM	Retourne une probabilité d'appartenance aux clusters	COMPLIQUEE	LONG	COMPLIQUE	NON	NON	OUI
HDBSCAN	Amélioration d'OPTICS et DBSCAN	SIMPLE	RAPIDE	SIMPLE	OUI	OUI	NON

b) Phase 2 : regroupement des clusters

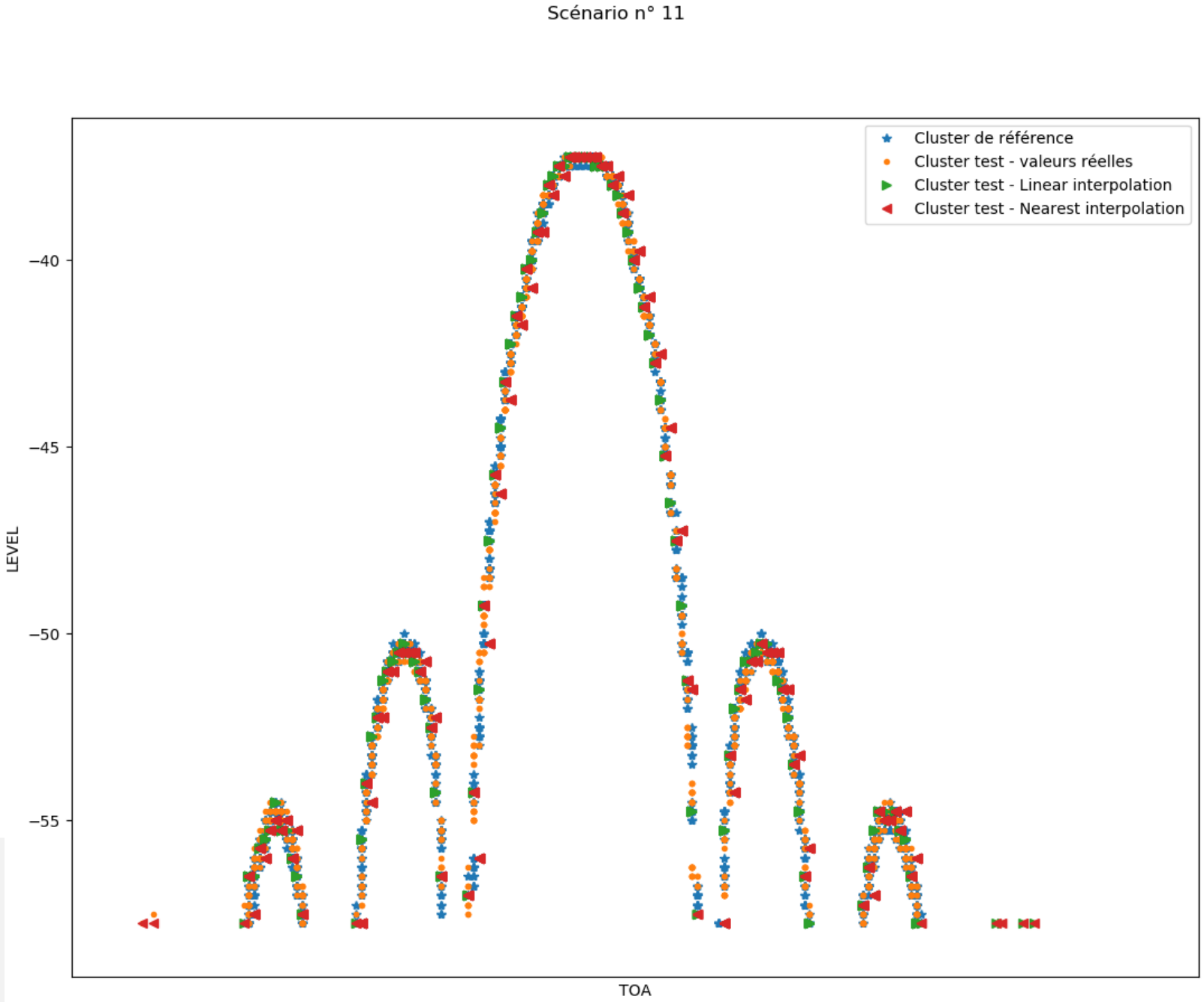
- **Méthode testée : Interpolation avec HAC**
  - Phase 1 : Interpolation à partir du niveau
    - Interpoler les valeurs d'un cluster de référence dans le plan Niveau x TOA afin d'approximer une fonction du type

$$f : y = f(x)$$

avec y : Niveau et x : TOA

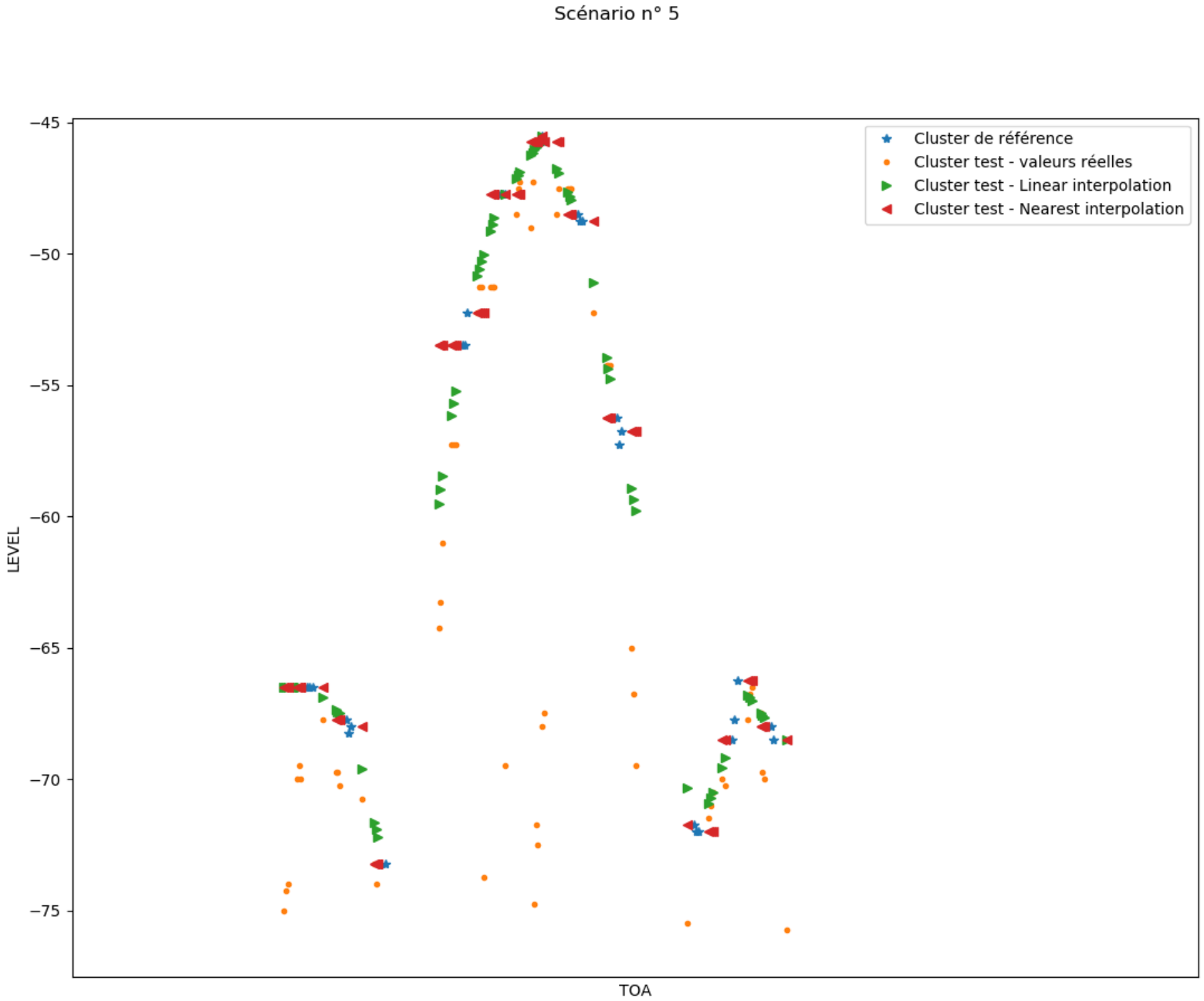
    - Estimer à l'aide de cette fonction les valeurs du niveau des autres clusters
    - Calculer une matrice de distance afin de comparer les valeurs estimées du niveau avec les valeurs réelles à l'aide d'une métrique de distance

b) Phase 2 : regroupement des clusters



Résultats interpolation linéaire

b) Phase 2 : regroupement des clusters

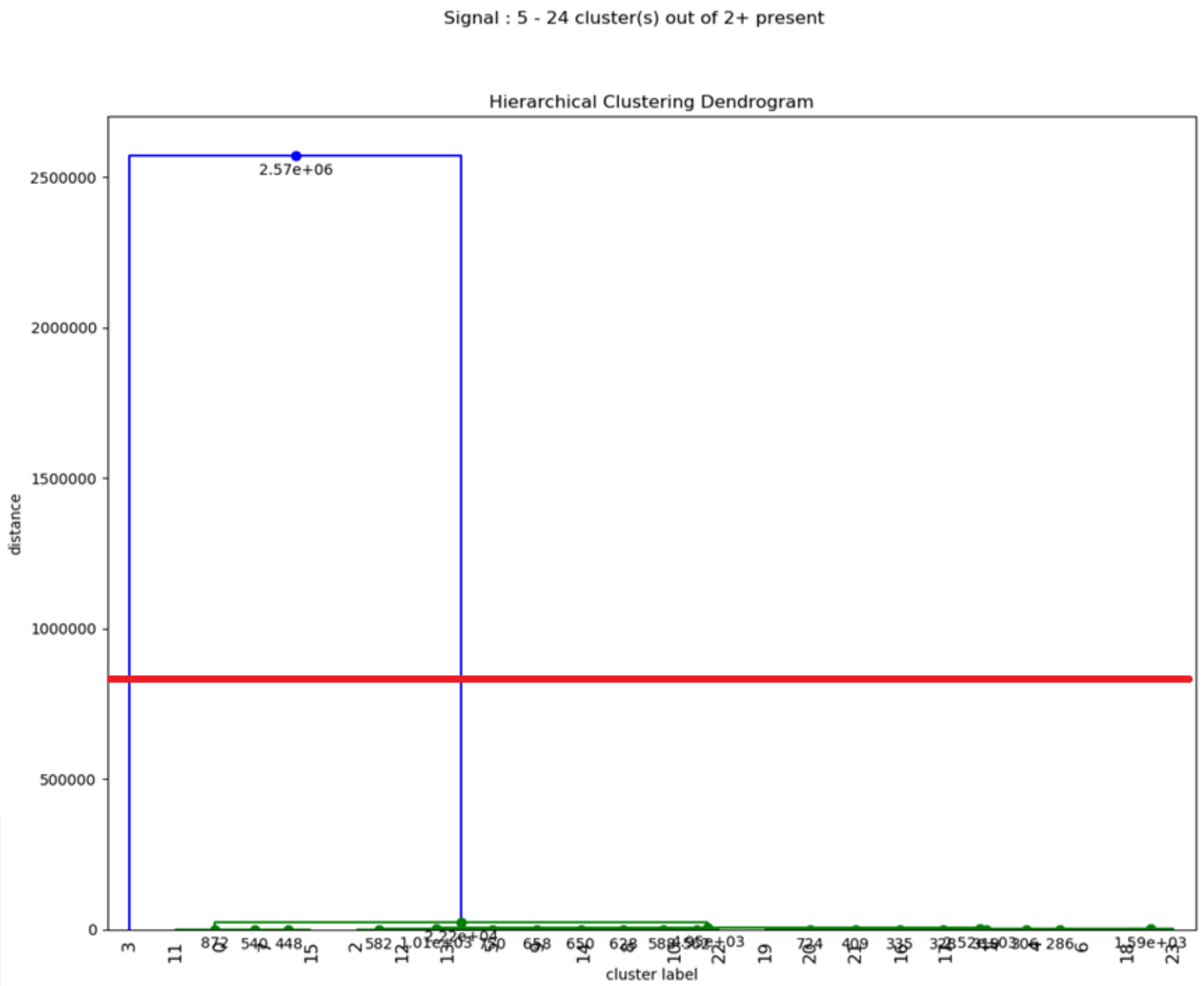


Résultats interpolation linéaire

b) Phase 2 : regroupement des clusters

- Phase 2 : Hierarchical agglomerative clustering
  - Chaque cluster représente une observation
  - A chaque itération les 2 groupes les plus proches sont fusionnés en un seul cluster (ceux ayant la plus petite mesure de dissimilarité)
  - Mise à jour de la matrice de distance
  - Répétition des étapes précédentes jusqu'à obtention d'un cluster unique

b) Phase 2 : regroupement des clusters



Dendrogramme d’une interpolation linéaire avec HAC

b) Phase 2 : regroupement des clusters

- Conclusions :
  - **Avantages :**
    - Pas de nombre de clusters à fixer au préalable
    - Ne fais pas d'hypothèse sur la forme des clusters trouvés
    - Agglomération des clusters de façon successive sans avoir à intervenir

b) Phase 2 : regroupement des clusters

- **Inconvénients :**

  - Interpolation ne fonctionne pas lorsque les valeurs du niveau ne sont pas bien réparties et les clusters pas assez consistants

  - Détermination d'un seuil pour stopper les fusions :

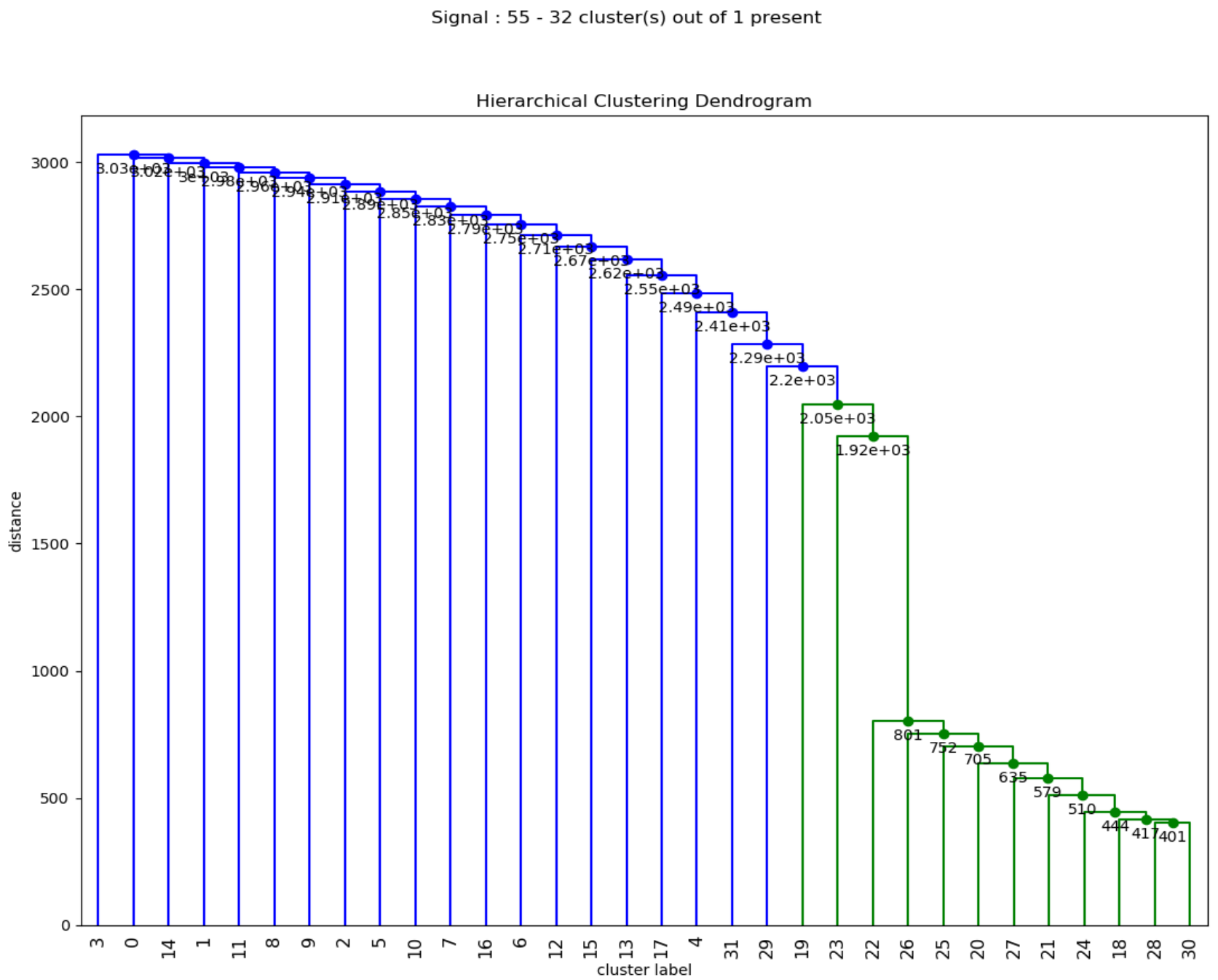
    - ✓ Compliqué à mettre en place car les signaux sont trop différents
    - ✓ Difficulté d'interprétation du dendrogramme

  - Méthode nécessitant un temps de calcul plus ou moins long selon la taille des clusters

  - Spécification d'une mesure de similarité entre les groupes d'observations

    - ✓ Distance entre les clusters = notion compliquée à caractériser
    - ✓ Recours à une autre métrique pour les comparer

b) Phase 2 : regroupement des clusters



Dendrogramme d'une interpolation linéaire avec HAC

a) Phase 1 : clustering

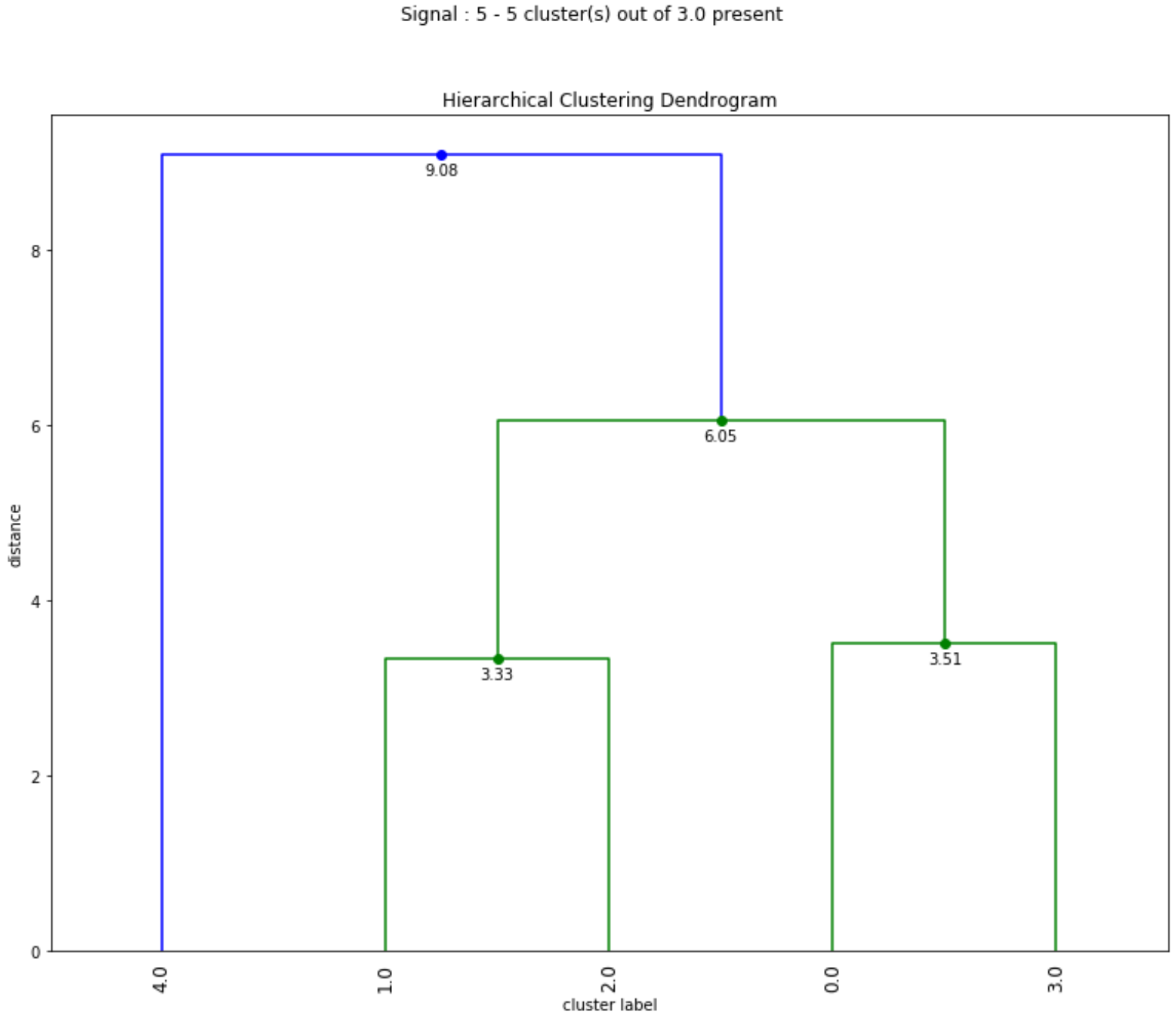
❖ **Méthode retenue : Transport optimal avec HAC modifiée**

- Le Transport optimal :
  - En 1871, Monge posa le problème suivant : pour deux densités de probabilité d et f, sur R, trouver une application T : R → R transformant la première densité en la seconde
  - Objectifs :
    - Répartir les points du cluster 1 qui ont une densité d selon la densité f des points du cluster 2 en minimisant le déplacement moyen
    - Calculer le coût de ce déplacement
    - Comparer la position des clusters grâce à un coût de déplacement et non plus grâce à une distance

b) Phase 2 : regroupement des clusters

- **Méthode retenue : Transport optimal avec HAC modifiée**
  - Fonctionnement :
    - Calcul du transport optimal entre les clusters
      - Sélection des valeurs de la TOA et du niveau du cluster A ainsi que celles du cluster B
      - Détermination du transport optimal pour déplacer les points du cluster A jusqu'à ceux du cluster B
      - Calcul du coût de ce transport
      - Généralisation aux autres clusters
    - Application d'un clustering agglomératif hiérarchique pour regrouper les clusters

b) Phase 2 : regroupement des clusters



Transport optimal avec HAC

b) Phase 2 : conclusions

- Conclusions :

  - **Avantages :**

    - Permet de comparer la position des clusters sans avoir déterminer de distance
    - Possibilité d'utilisation de plusieurs métriques pour calculer le coût

  - **Inconvénients :**

    - Très long a exécuter lorsque le nombre de clusters, la taille du cluster ou encore le nombre de dimension augmente

4 Conclusions & perspectives

1) Conclusions

- Phase 1 : clustering via Hdbscan
  - Fonctionne très bien sur des signaux même complexes
  - Retourne un nombre de clusters raisonnable
  - Facile à implémenter
- Phase 2 : Regroupement des clusters via le Transport optimal et une HAC modifiée
  - Résultats encourageants

2) Besoins

- Volumétrie de données
  - Volumétrie insuffisante pour effectuer des tests robustes
  - Besoin d'un enrichissement de nos jeux de données réelles
- Données plus complexes
  - Les données sont trop simplifiées pour avoir des résultats empiriques
  - Besoin de complexifier les données pour que les algorithmes soient robustes
- Avoir accès à une base de données de données radar afin d'anticiper la phase d'identification

3) Perspectives

- Perspectives
  - Reconstruction de la dTOA pour valider le regroupement des clusters
  - Modèles ensemblistes
  - Machine learning :
    - Analyse de la répétitivité
    - Reconnaissance de formes

Thank you for your attention