Désentrelacement et classification des émetteurs RADARs basés sur l'utilisation des distances de Transport Optimal

Manon MOTTIER
Université Paris-Saclay, CNRS, CentraleSupélec
Laboratoire des signaux et systèmes
91190, Gif-sur-Yvette, France
manon.mottier@centralesupelec.fr

Gilles CHARDON
Université Paris-Saclay, CNRS, CentraleSupélec
Laboratoire des signaux et systèmes
91190, Gif-sur-Yvette, France
gilles.chardon@centralesupelec.fr

Frédéric PASCAL
Université Paris-Saclay, CNRS, CentraleSupélec
Laboratoire des signaux et systèmes
91190, Gif-sur-Yvette, France
frederic.pascal@centralesupelec.fr

***Abstract*—Cet article introduit une nouvelle méthode pour désentrelacer et classifier les émetteurs d'un signal RADAR utilisant des distances de transport optimal. Une stratégie en deux étapes est développée et appliquée pour désentrelacer un signal RADAR : tout d'abord, un algorithme de clustering est utilisé pour séparer les données. Ensuite, les résultats (classes) du clustering initial sont incorporées dans un clustering agglomératif hiérarchique combiné à des distances de transport optimal utiliser pour fusionner (regrouper) les classes. Enfin, le processus d'identification est appliqué sur ces résultats du désentrelacement; la méthodologie est basée sur l'élaboration d'une distance entre un groupe d'impulsions et une description des caractéristiques d'un émetteur RADAR à partir d'une base de données de référence. La méthodologie est construite et évaluée sur des données simulées étiquetées.**

***Index Terms*—Transport Optimal, Désentrelacement, Classification, Processus de Reconnaissance RADAR, Guerre électronique, Machine Learning**

I. Introduction

Les évolutions technologiques de ces dernières années ont permis de nombreuses innovations dans le domaine de la guerre électronique : les équipements deviennent de plus en plus complexes et sophistiqués, entraînant une transformation de la chaîne de traitement du signal RADAR et plus particulièrement du processus de reconnaissance RADAR (amélioration des récepteurs et des émetteurs, étalement des émissions sur le spectre, modifications des formes d'onde, systèmes d'armement autonomes, etc.). Tous ces changements représentent un défi continu afin de proposer de nouvelles techniques plus précises pour classer et identifier les émetteurs RADAR à partir d'un signal reçu.

Le processus de reconnaissance RADAR peut se définir en deux étapes : la première étape consiste à désentrelacer un signal, c'est-à-dire à séparer et regrouper les impulsions mélangées d'un nombre inconnu d'émetteurs. Les premiers travaux de désentrelacement conventionnels sont principalement basés sur l'utilisation de 3 caractéristiques : la direction d'arrivée (DOA, *Direction-Of-Arrival*), la fréquence (F, *frequency*) et le temps d'arrivée (TOA, *Time-Of-Arrival*). La TOA/DOA et la F ont été utilisées dans un premier temps pour filtrer les données, puis les histogrammes du temps d'arrivée différencié (dTOA) pour identifier les motifs de répétition des impulsions [1]. Ces travaux ont servi de fondation et inspiré de nombreux auteurs afin de proposer de nouvelles méthodes. On retrouve notamment l'utilisation des histogrammes de dTOA cumulés [2] ou encore les histogrammes de différences séquentielles [3]. Ces méthodes fonctionnent très bien lorsque les RADARs présentent des caractéristiques simples comme émettre sur une bande de fréquence unique et de façon continue. Lorsque ces caractéristiques sont plus complexes, les méthodes sont le plus souvent basées sur des modèles de Deep Learning (voir *e.g/,* [4]) qui nécessitent une quantité considérable de données et qui sont difficiles à paramétrer. Toutes ces méthodes sont la plupart du temps construites dans un cadre supervisé.

La deuxième phase du processus de reconnaissance RADAR concerne l'identification des émetteurs. L'identification d'émetteurs RADAR à partir d'impulsions reçues représente un enjeu majeur pour le renseignement car cette étape permet de donner un avantage stratégique dans la prise de décision et l'engagement militaire. Pour les RADARs simples, l'analyse temps-fréquence est suffisante pour identifier l'émetteur avec une grande certitude [5]. Néanmoins, les émetteurs RADAR peuvent avoir des caractéristiques plus complexes; dès lors, des classifieurs supervisés issus des modèles de Deep Learning sont très largement utilisés pour faire face à cette problématique [6]–[10]. Ces méthodes nécessitent un volume important de données pour l'étape d'entraînement, ainsi qu'un réentraînement du classifieur lors de l'ajout d'une nouvelle classe et prennent en compte un petit nombre de classes à identifier.

L'acquisition de données réelles représente un véritable défi dans le traitement des données RADARs, particulièrement les données réelles étiquetées. Les algorithmes et les méthodologies sont souvent développés à l'aide de petits ensembles de données et/ou de données simulées. La plupart des méthodes précédentes sont basées sur des données

simulées et leurs performances dépendent fortement de la précision du simulateur. De plus, les développements technologiques ont modifié les profils des RADARs : les caractéristiques des émetteurs sont devenus plus complexes, enrichissant le panorama existant de nouveaux types d'émetteurs aux motifs plus variés. De nouvelles méthodes ont été développées afin de comparer les caractéristiques du groupe à une base de données connue, permettant également de détecter de nouveaux émetteurs [11], [12].

Pour pallier ces contraintes, nous avons construit un nouveau processus simple de reconnaissance RADAR non supervisé basé sur l'utilisation de distances de transport optimal. L'article est organisé comme suit : la première partie présente les données, suivie des sections 2 et 3 qui introduisent les méthodologies de désentrelacement des impulsions et de classification des émetteurs RADARs. La section 4 présente un cas d'application avant de conclure dans la section 5, et de proposer de nouvelles perspectives de recherches quant à la poursuite de ces travaux.

II. Description des données

Cette section présente les données utilisées à partir desquelles nous avons développé notre méthodologie ainsi que la procédure expérimentale mise en place pour obtenir des données étiquetées non confidentielles. Les labels sont ici utilisés uniquement pour évaluer les performances de la méthode développée, qui est par construction totalement non-supervisée. Nous avons eu recours à un simulateur de données construit et certifié par des experts du domaine. La méthodologie a également été validée sur des données réelles que nous avons à notre disposition. Ces données étant confidentielles, les résultats ne sont pas présentés dans cet article. Bien que les théâtres de la guerre électronique aillent de pair avec l'usage de contremesures, cet article se positionne dans un cadre ELINT et ne prend pas en compte la présente de mesures adverses introduisant des signaux perturbateurs. Les écarts de fréquence dûs à l'effet Doppler étant petits par rapport aux séparations fréquentielles entre les différents émetteurs, l'effet Doppler sera negligé dans notre analyse.

La diversité/complexité des signaux simulés se traduit par l'acquisition de signaux mono- ou multi-capteurs étiquetés, des signaux à modulation de fréquence et/ou temporelle, des signaux comportant des erreurs de mesure, des outliers, des données manquantes ou du bruit non-gaussien. Notre base de données contient des signaux hétérogènes de différentes tailles (allant de 20 à plus de 100000 impulsions) pouvant contenir plus de 10 émetteurs et avec des niveaux de bruit pouvant varier fortement. Les récepteurs sont supposés statiques et ont un seuil de détection connu, mais nous n'avons pas d'autres informations sur les caractéristiques des émetteurs et récepteurs. Les données simulées sont anonymisées mais étiquetées. Les récepteurs sont omnidirectionnels, rendant la direction d'arrivée inutilisable. Les impulsions acquises par le système de RADAR passif sont segmentées, analysées puis décrites par un ensemble limité de caractéristiques. Dans ce travail, les N impulsions acquises sont décrites uniquement à l'aide des quatre caractéristiques suivantes :

- t_n : Temps d'Arrivée (TOA, μs)
- d_n : Durée d'Impulsion (DI, ns),
- f_n : Fréquence (F, MHz),
- g_n : Niveau (LEVEL, dBm).

Nous ne prenons pas en compte d'autres informations sur les impulsions (comme par exemple la forme d'onde, la modulation de fréquence, etc.).

La Figure 1 présente un exemple de données simulées regroupant les impulsions de 4 émetteurs identifiables par leurs couleurs. La fréquence, la durée d'impulsion et le niveau sont représentés comme des fonctions du temps sur les 3 premiers graphiques et chaque point caractérise une impulsion. Les valeurs des caractéristiques ont été anonymisées de façon aléatoire. Les RADARs peuvent émettre de façon continue et régulière sur différentes fréquences comme souligné par l'émetteur 1 qui transmet sur des fréquences basses de façon constante. A l'inverse, l'émetteur 2 est présent sur des fréquences plus hautes et de façon plus ponctuelle. Les bandes de fréquences peuvent être partagées par des émetteurs ayant des caractéristiques proches comme les émetteurs 2 et 3, compliquant leur séparabilité. On retrouve fréquemment cette problématique lorsque le périmètre d'écoute concerne par exemple un port ou un aéroport où l'on retrouve des RADARs avec des caractéristiques très similaires. La simultanéité d'émission des RADARs peut conduire à une superposition et un mélange des lobes comme illustré dans le plan (g_n, t_n) pour les émetteurs 0, 1 et 3 qui sont actifs simultanément autour de 47 μs. Enfin, le plan (f_n, d_n) présenté par le dernier graphique montre qu'un RADAR n'est pas forcément représenté par un paquet d'impulsion unique mais au contraire, peut être scindé en plusieurs paquets d'impulsions comme les émetteurs 2 et 3.

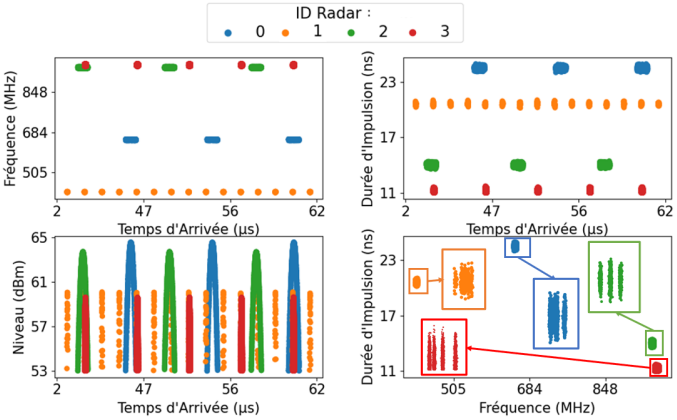

Fig. 1. Exemple d'un signal simulé regroupant les impulsions de 4 RADARs. Chaque couleur identifie un émetteur distinct.

III. Stratégie de désentrelacement d'un signal RADAR

Nous avons construit une stratégie de désentrelacement d'un signal RADAR en 2 étapes; premièrement un algorithme de

clustering préliminaire est utilisé pour séparer les impulsions d'un nombre inconnu de RADARs dans le plan (f_n, d_n). Puis, à partir de ces résultats, un algorithme de clustering agglomératif hiérarchique combiné à des distances de transport optimal est utilisé afin de regrouper les clusters appartenant à un même RADAR dans le plan (t_n, g_n).

A. Pré-traitement

Les dimensions physiques des données en fréquence (MHz) et en durée d'impulsion (ns) sont incohérentes. Une renormalisation est nécessaire pour calculer les distances entre les caractéristiques. Plusieurs méthodes ont été testées comme la normalisation par quantiles ou un pré-clustering via l'algorithme des GMM (Modèle de Mélange Gaussien) [13]. La méthode des quantiles consiste à utiliser les intervalles interquantiles pour mettre à l'échelle les données tandis que celle des GMM estime les variances intra-cluster par rapport aux caractéristiques considérées. Nous avons finalement choisi d'utiliser la méthode des quantiles car elle préserve au mieux la distribution des données et est très facilement implémentable.

B. Clustering dans le plan (f_n, d_n)

Les signaux reçus contiennent les impulsions mélangées d'un nombre inconnu d'émetteurs. Il est donc nécessaire de séparer ces impulsions et de les regrouper en clusters. Tout d'abord, nous appliquons un algorithme de clustering pour séparer les impulsions uniquement à partir de deux caractéristiques très discriminantes et fiables : la fréquence et la durée d'impulsion. Des algorithmes de clustering classiques tels que les K-Means [14] ou les GMM [13] sont fréquemment utilisés. Ces méthodes sont limitées car elles font des hypothèses sur le nombre de clusters à identifier ou sur leurs formes.

Nous avons choisi d'utiliser HDBSCAN [15] qui est particulièrement adapté à notre problématique. HDBSCAN est un algorithme de clustering non supervisé, basé sur une version hiérarchique de l'algorithme DBSCAN [16] capable de détecter des clusters de densités et de formes différentes. Le fonctionnement de l'algorithme est présenté plus en détail dans un précédent article [17]. L'idée principale est de regrouper des points qui vivent dans une même zone dense. Les résultats du clustering dépendent fortement de l'hyperparamètre déterminant la taille minimale d'un groupe pouvant être considéré comme un cluster. HDBSCAN a été paramétré de façon à surestimer le nombre de clusters retournés afin de ne pas mélanger les impulsions de plusieurs émetteurs dans un même cluster. Il est par exemple possible d'avoir des émetteurs avec seulement quelques impulsions ou émettant très peu dans le temps. Des comparaisons ont été effectuées afin de trouver un compromis pour fixer ce seuil.

Les résultats du clustering effectué par HDBSCAN dans le plan (f_n, d_n) sont affichés sur la Figure 2. L'algorithme identifie 9 clusters ainsi qu'une classe d'outliers (-1). Les clusters ont des tailles hétérogènes allant de 180 à plus de 2000 impulsions. Le plan (f_n, d_n) de droite montre que les résultats du clustering sont cohérents et que les impulsions des émetteurs 0 et 1 ont correctement été regroupées dans des clusters uniques. À l'inverse, HDBSCAN a scindé les impulsions des RADARs émettant autour au dessus de 948 MHz en plusieurs clusters. Le plan (g_n, t_n) de gauche met en évidence la distribution de ces clusters le long des lobes.

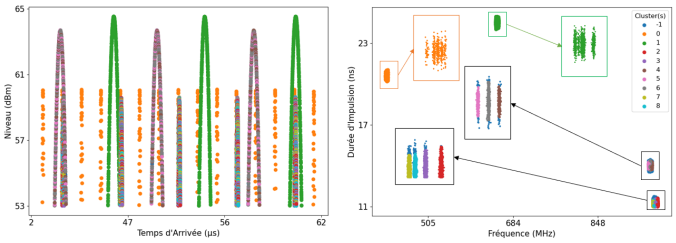

Fig. 2. Résultats du clustering d'HDBSCAN effectué à partir de la fréquence et la durée d'impulsion. Chaque couleur représentent les clusters ainsi qu'une classe d'outliers (-1).

C. Agglomération des clusters

Comme précédemment montré sur la Figure 1, il est possible qu'un RADAR soit représenté par plusieurs clusters; il est donc nécessaire de regrouper ces clusters. À partir des clusters d'HDBSCAN, un clustering agglomératif hiérarchique dans le plan (t_n, g_n) a été construit [18] en utilisant des distances de transport optimal comme illustré dans l'algorithme 1. Notre méthodologie est construite à partir d'une distance [19], [20] permettant de mesurer la similarité et la dissimilarité entre les clusters. La suite de l'article montrera que des distances de transport optimal sont particulièrement bien adaptées à notre problématique.

Algorithm 1 Clustering Agglomératif Hiérarchique combiné aux distances de transport optimal

- Représentation des clusters par une mesure à partir du temps d'arrivée et du niveau : τ.
- Calcul de la distance entre chaque cluster en utilisant le transport optimal : $d(\tau_i, \tau_j)$.
- Aggrégation des deux clusters les plus proches
- Mise à jour des distances.
- Réitération des étapes précédentes jusqu'à l'obtention d'un cluster unique.

Dans la suite de l'article, les distances de transport optimal seront calculées entre des distributions discrètes [21]. Considérons le cas de deux distributions de probabilité discrètes $\nu = \sum_{n=1}^{N} a_n \delta_{x_n}$ et $\mu = \sum_{m=1}^{M} b_m \delta_{y_m}$, avec $\mathbf{a} = (a_1, \dots a_N) \in \mathbf{R}_+^N$, $\sum_{n=1}^{N} a_n = 1$, et $\mathbf{b} = (b_1, \dots b_M) \in \mathbf{R}_+^M$, $\sum_{m=1}^{M} b_m = 1$. Un plan de transport $\mathbf{P}$ est défini par ses coefficients P_{nm}, qui modélisent la masse prise en x_n transportée en y_m. Ce transport à un coût par unité de masse $C_{nm} = c(x_n, y_m)$, où $c(\cdot, \cdot)$ est fonction de coût. Le coût total $C(\mathbf{P})$ d'un plan de transport est donc

$$C(\mathbf{P}) = \sum_{n=1}^{N} \sum_{m=1}^{M} C_{nm} P_{nm} \,. \tag{1}$$

La cohérence du plan de transport $\mathbf{P}$ avec les distributions de départ et d'arrivée est garantie par les conditions $\mathbf{P}\mathbf{1}_M = \mathbf{a}, \mathbf{P}^T\mathbf{1}_N = \mathbf{b}^T$. Le plan de transport optimal $\mathbf{P}^\star$ est obtenu par la résolution du problème d'optimisation linéaire suivant :

$$\mathbf{P}^\star = \underset{\mathbf{P} \in \mathbf{R}_+^{N \times M}}{\operatorname{argmin}} \; C(\mathbf{P}) \text{ tel que } \mathbf{P}\mathbf{1}_M = \mathbf{a}, \mathbf{P}^T\mathbf{1}_N = \mathbf{b}^T \,, \tag{2}$$

et la distance de transport optimal entre ν et μ est donnée par $d(\nu, \mu) = C(\mathbf{P}^\star)$. La détermination et la résolution du plan de transport optimal sont détaillés dans la toolbox POT [21], accessible sous python.

Les impulsions des clusters sont représentées par une mesure de probabilité décrivant la distribution de leur temps d'arrivée et de leur niveau :

$$\tau = \frac{1}{P} \sum_{p=1}^{P} \delta_{g_n, t_n} \,, \tag{3}$$

avec P le nombre d'impulsions du cluster. Afin de diminuer la complexité de calcul des distances, la mesure de probabilité utilisée en pratique est obtenue à partir d'un histogramme des données. Nous faisons l'hypothèse que les clusters appartenant à un même émetteur sont actifs simultanément et répartis le long des lobes comme précédemment illustré sur la Figure 2. Dans la stratégie d'agrégation, pour quantifier cette similarité ou cette dissemblance, nous avons utilisé une distance de transport optimal qui est défini comme le coût pour déplacer les points d'un endroit à un autre. Les clusters obtenus dans le plan appartenant au même émetteur auront entre leurs distributions une distance minime. Les deux clusters ayant la plus petite distance de transport optimal dans le plan (t_n, g_n) sont agrégés :

$$(i, j)^\star = \operatorname{argmin} d\left(\tau_i, \tau_j\right) . \tag{4}$$

Après la fusion, la distance entre les clusters fusionnés et les autres clusters est mise à jour. Puis, le processus est réitéré jusqu'à ce que les clusters soient entièrement agrégés. La figure 3 présente une itération du processus de fusion : le coût de transport entre les clusters 4 et 6 (orange et vert) sera très faible (les histogrammes de ces clusters étant très similaires) tandis que le coût entre les clusters 4 et 0 (orange et rouge) sera très élevé.

Le dendrogramme présente de façon simplifiée les agrégations des clusters réalisées à chaque étape du clustering agglomératif hiérarchique combiné aux distances de transport optimal. On peut voir sur la Figure 4 que les regroupements des clusters 2, 3, 7 et 8, représentés dans l'encadré vert, ont des coûts de transport faibles car tous ces clusters appartiennent au même RADAR. En revanche, le regroupement du cluster 0 avec les clusters 2, 3, 7, 8, représenté dans l'encadré rouge, aura un coût de transport très élevé, ce qui indique que le cluster 0 n'appartient pas au même émetteur que les clusters 2, 3, 7 et 8. Plusieurs métriques non supervisées (Silhouette score [22], Davies-Bouldin score [23], Calinski-Harabasz Score [24]) ainsi que la connaissance de l'environnement RADAR sont utilisées pour créer un critère d'arrêt afin de stopper les fusions. La ligne rouge indiquée sur la Figure 4 permet d'identifier les regroupements finaux : ici quatre ensembles d'impulsions sont retenus (cluster 0, cluster 1, regroupement des clusters 4,5,6 et regroupement des clusters 2, 3, 7, 8).

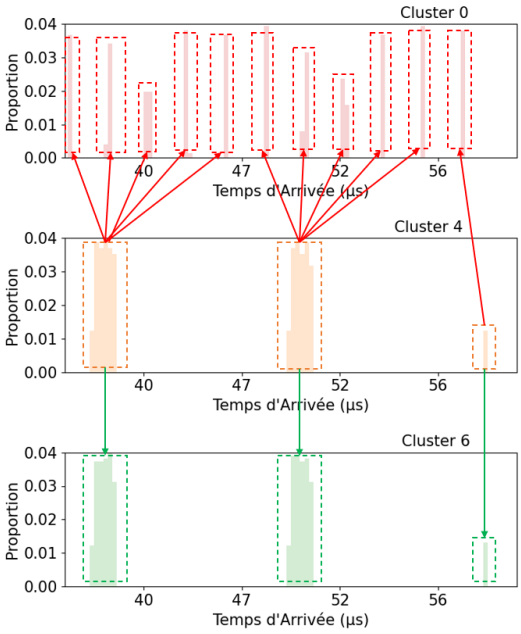

Fig. 3. Distribution des temps d'arrivée des clusters 0, 4 et 6 sous forme d'histogramme.

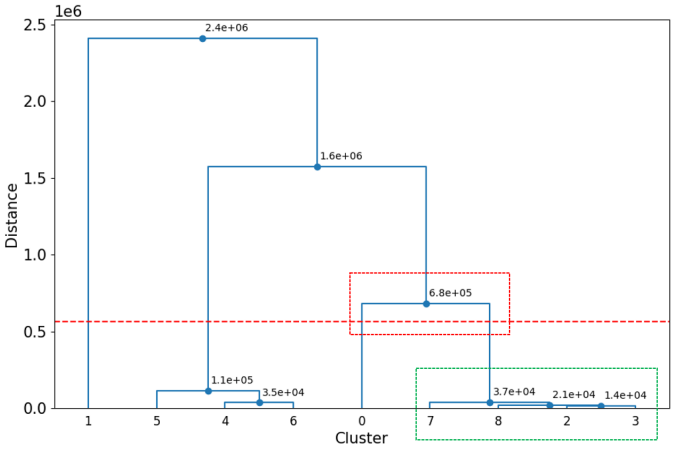

Fig. 4. Dendrogramme représentant les aggrégations à chaque étape du clustering agglomératif utilisant des distances de transport optimal.

Enfin, la Figure 5 illustre le résultat de l'agglomération des clusters. Ici le clustering a parfaitement fonctionné et identifie correctement les 4 émetteurs présents dans le signal. En particulier, les lobes présents autour de 47 μs ont parfaitement été reconstruits et le transport optimal a permis de regrouper correctement les différents clusters répartis sur ces lobes.

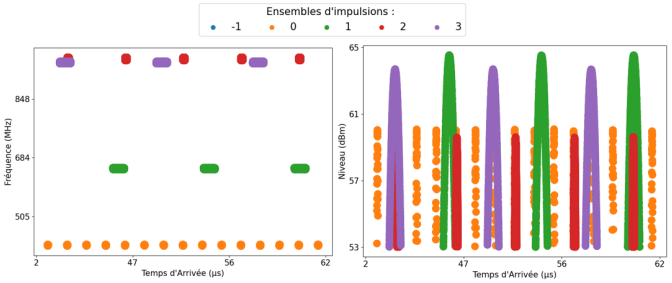

Fig. 5. Regroupements finaux. Les couleurs représentent les ensembles d'impulsions obtenus ainsi qu'une classe d'outliers (-1).

Afin d'évaluer la qualité des regroupements, nous avons utilisé 3 mesures :

- Homogénéité [25] : métrique permettant d'évaluer si toutes les impulsions d'un cluster appartiennent bien au même émetteur.
- Complétude [25] : métrique évaluant le niveau d'étalement des impulsions à travers les clusters.
- Taux d'outliers : mesure évaluant la proportion d'impulsions perdues durant le processus.

Les valeurs d'homogénéité et de complétude sont égales à 1, ce qui signifie que les impulsions sont correctement réparties dans les différents ensembles d'impulsions, et qu'il n'y a pas de mélange. Le taux d'outliers indique que 1.6% des impulsions sont perdues durant le processus de désentrelacement; ces impulsions ont été exclues par HDBSCAN à l'étape 1.

IV. Classification des émetteurs RADARs

Lorsque la phase de désentrelacement est terminée, la classification peut démarrer en utilisant les ensembles d'impulsions précédemment désentrelacés. La méthodologie est décrite dans l'algorithme 2. Elle est basée sur le développement d'une distance entre la description des émetteurs RADARs d'une base de données de référence et les ensembles d'impulsions désentrelacés. Dans notre cas, les ensembles d'impulsions et les classes d'émetteurs RADARs sont représentés sous forme de distributions de probabilités. La classification est ensuite effectuée en identifiant la classe d'émetteur RADAR la plus proche de celle de l'ensemble d'impulsions au sens d'une distance bien choisie. Lorsque l'émetteur RADAR possède des caractéristiques simples, comme par exemple une seule fréquence, des distances classiques telles que la distance euclidienne entre la moyenne des caractéristiques des impulsions reçues et les caractéristiques de l'émetteur de la base RADAR peuvent être utilisées. Néanmoins, dès lors que l'émetteur possède des caractéristiques plus complexes, il n'est pas possible de décrire la distribution des caractéristiques par de simples moyennes. Plusieurs méthodes existent afin de calculer une distance entre des distributions de probabilités comme la distance de variation totale ou la divergence de Kullback-Leibler. Dans notre cas, nos données sont représentées par des distributions discrètes ayant généralement des supports disjoints ce qui ne nous permet pas d'avoir recours à ces distances classiques. Les parties suivantes mettent en évidence l'intérêt de l'utilisation des distances de transport optimal pour répondre à notre problématique ainsi que leur robustesse au bruit.

Algorithm 2 Classification avec les distances de Transport Optimal.

- Construction d'une distribution de probabilité à partir des ensembles d'impulsions : ν
- Construction d'un mesure discrète à partir des classes d'émetteurs RADARs appartenant à une base de données de référence : μ_j
- Calcul du coût de transport entre la distribution de probabilité de l'ensemble d'impulsion et chaque classe d'émetteur : $d_j = d(\nu, \mu_j)$
- Assignation de la classe d'émetteur ayant le plus petit coût de transport entre sa distribution de probabilité et la distribution de l'ensemble d'impulsion : $j^\star$

A. Représentation des Classes d'émetteurs

La base de données RADAR utilisée pour faire la classification comprend les références de plus de 60 classes d'émetteurs distinctes. Certaines classes ont des caractéristiques très similaires tandis que d'autres sont très facilement distinguables. Nous avons choisi de travailler uniquement à partir de la fréquence et de la durée d'impulsion car ce sont des caractéristiques très discriminantes et fiables. À partir de cette base de données, nous avons construit une mesure décrivant chaque classe d'émetteur :

$$\mu_j = \sum_{n=1}^{N} \alpha_n \delta_{f_n, d_n} \,, \tag{5}$$

avec N le nombre de couples de fréquences et de durées d'impulsion sur lesquelles le RADAR émet, α la proportion d'impulsions dans chaque couple, et δ, la masse de Dirac (avec $\sum_{n=1}^{N} \alpha_n = 1$).

La Figure 6 illustre un exemple de classes d'émetteurs simulés. Chaque tige représente une fréquence utilisée par un émetteur et sa hauteur la proportion d'apparition de cette fréquence. Certains RADARs émettent sur des bandes de fréquences très différentes comme les émetteurs D et I tandis que d'autres ont des caractéristiques très proches comme les émetteurs E et F. Les RADARs peuvent être mono fréquence comme l'émetteur K qui émet également sur une fréquence commune à celle de l'émetteur E; cette similarité fréquentielle ne permet pas de les différencier. A l'inverse, tout comme l'émetteur A, les RADARs peuvent être multifréquences.

La Figure 7 montre les précédents émetteurs dans le plan (f_n, d_n). Les émetteurs K et E sont clairement séparés dans ce plan, ce qui nous indique que l'ajout de caractéristiques supplémentaires améliore la séparabilité des émetteurs.

La matrice de coût représentée sur la Figure 8 est construite à partir de la fréquence et de la durée d'impulsion. Elle présente de façon simplifiée les coûts de transport

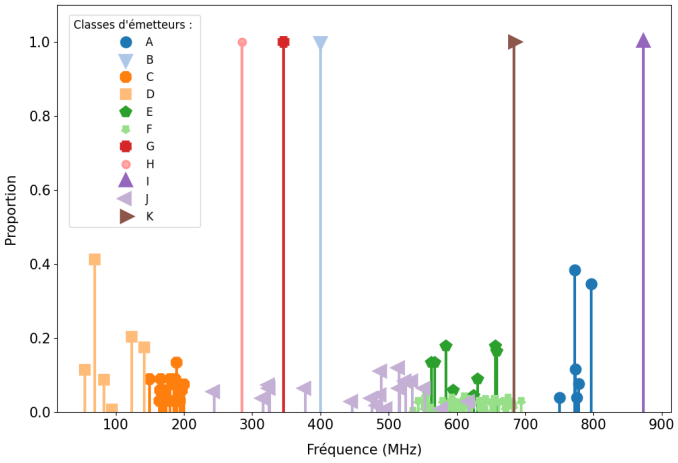

Fig. 6. Représentation de classes d'émetteurs simulées. Les émetteurs sont représentés en une dimension dans le plan fréquentiel. Chaque couleur représente une classe.

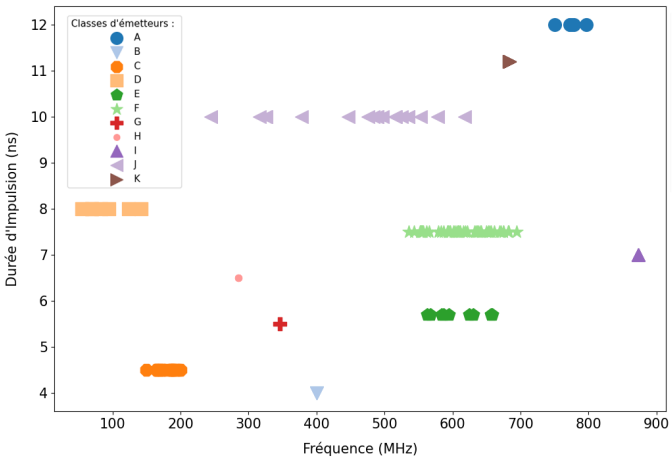

Fig. 7. Représentation des classes d'émetteurs simulées précédentes. Les émetteurs sont représentés en deux dimensions dans le plan F-DI. Chaque couleur représente une classe.

entre les émetteurs simulés précédents. Les couleurs caractérisent le niveau de proximité entre les émetteurs. Comme précédemment expliqué, certaines classes peuvent avoir des caractéristiques similaires. À titre d'exemple, les classes E et F sont séparées par une petite distance à l'inverse des classes A et C pour lesquelles elle est plus grande. Une attention particulière est portée sur les classes ayant des petites distances; on s'attend à ce qu'elles soient fréquemment confondues car elles ont des caractéristiques observables similaires.

La Figure 9 représente le plan de transport de l'émetteur F vers les émetteurs A, E, G, J et K. Le coût de transport des points de l'émetteur F vers l'émetteur E est faible car tous les points sont déplacés vers un emplacement très proche : les deux émetteurs émettent sur des bandes de fréquences très similaires et ont une durée d'impulsion proche. À l'inverse, le coût de transport de l'émetteur F vers l'émetteur G est 20 fois plus élevé car tous les points sont déplacés vers une bande de

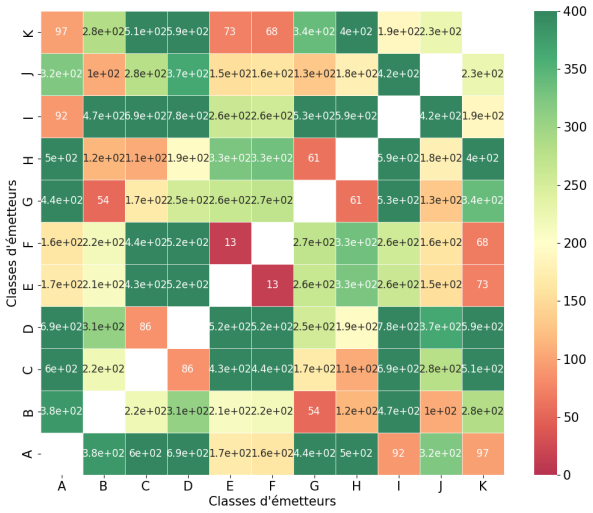

Fig. 8. Matrice de coûts entre les classes d'émetteurs simulées construite à partir de la fréquence et de la durée d'impulsion. Une couleur verte indique un coût de transport faible tandis que le rouge représente un coût élevé.

fréquence unique et très différente; les points doivent effectuer des déplacements plus importants. La distance de transport optimal tient compte de la proportion d'impulsions pour un couple donné de fréquence et de durée d'impulsion.

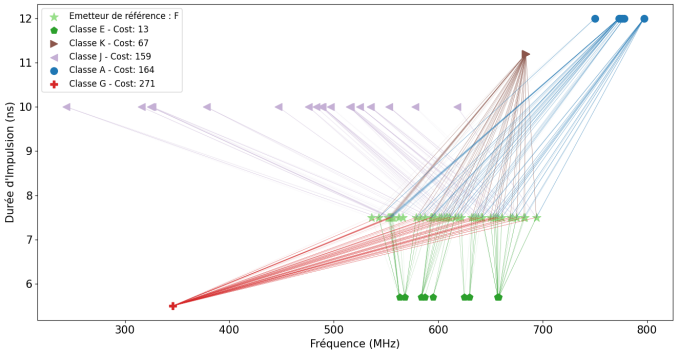

Fig. 9. Plans de transport des points de l'émetteur F vers les émetteurs A, E, G, J, K avec le transport optimal en deux dimensions.

B. Modèle de classification

Les ensembles d'impulsions sont modélisés sous forme de distributions de probabilité de la façon suivante :

$$\nu = \frac{1}{M} \sum_{m=1}^{M} \delta_{f_m, d_m} \,, \tag{6}$$

avec M le nombre d'impulsions de l'ensemble. Les ensembles d'impulsions sont également regroupés en intervalles de fréquences et de durée d'impulsion. L'algorithme identifie la vraie classe d'émetteur $j^\star$:

$$j^\star = \operatorname{argmin} d\left(\mu_j, \nu\right). \tag{7}$$

La classification est faite en attribuant à l'ensemble d'impulsion la classe d'émetteur RADAR la plus proche en termes de distance au sens du transport optimal.

C. Résultats

Le résultat du classifieur appliqué à l'ensemble d'impulsions 1 est illustré sur la Figure 10. Les impulsions et les trois classes d'émetteurs les plus proches sont superposées sur le graphique de gauche. Les points de la classe la plus proche se superposent parfaitement à ceux des données. Le classifieur identifie correctement l'émetteur présent dans cet ensemble; ce résultat est confirmé par le fait que la distance vers la deuxième classe la plus proche est 20 fois plus élevée. Les plans de transports entre la distribution des données et les trois classes les plus proches [21] sont affichés sur le graphique de droite. La deuxième classe représente un RADAR transmettant sur six fréquences différentes tandis que les données sont caractérisées par seulement 5 fréquences : les points des données sont envoyés sur les différentes fréquences de classe 2 en respectant les proportions; c'est pourquoi les impulsions autour d'une fréquence donnée ne sont pas toutes envoyées au même point. Enfin, bien que la troisième classe semble plus proche dans le plan, ces fréquences sont très différentes de celles de l'ensemble d'impulsion; les points doivent donc effectuer de plus grands déplacements, ce qui augmente considérablement le coût de transport.

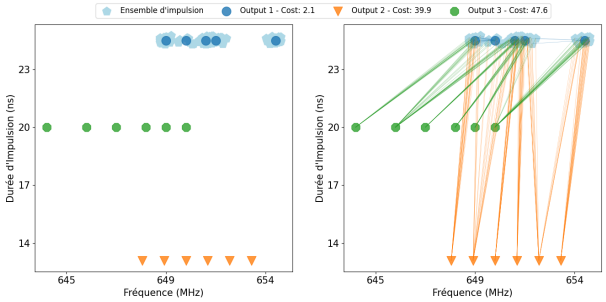

Fig. 10. Résultats de la classification pour l'ensemble d'impulsion 1.

La Figure 11 montre de façon analogue le résultat de notre méthodologie de classification appliquée à l'ensemble d'impulsions 3. Le graphique de gauche superpose les impulsions et les trois classes d'émetteurs les plus proches. Les points bleus se superposent également très bien à ceux des données. Le classifieur identifie correctement l'émetteur présent. Cependant, l'écart de distance entre l'ensemble d'impulsion et les sorties 1 et 2 est plus faible que celui de l'exemple précédent car les classes d'émetteurs 1 et 2 émettent sur des durées d'impulsion proches.

V. Cas d'application

Cette section présente un nouveau signal simulé à partir duquel nous avons appliqué le processus de reconnaissance RADAR. Ce signal contient environ 8000 impulsions et est présenté sur la Figure 12. La représentation fréquentielle met en évidence la simultanéité temporelle de plusieurs émetteurs autour de 27 μs. Certains émetteurs transmettent de façon continue et régulière comme celui à 1020 MHz à l'inverse du RADAR présent après 27 μs qui lui semble n'apparaître qu'une fois sur notre simulation et être multifréquences. Le

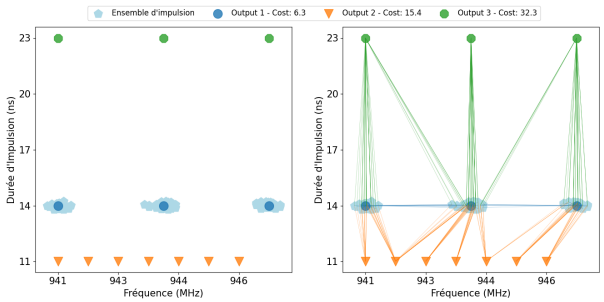

Fig. 11. Résultats de la classification pour l'ensemble d'impulsion 3.

plan (d_n, t_n) est caractérisé par un large étalement de la durée d'impulsion compliquant l'analyse et ne permettant pas de distinguer les émetteurs. De plus, la superposition des lobes sur le plan (g_n, t_n) rend la séparation des émetteurs difficile.

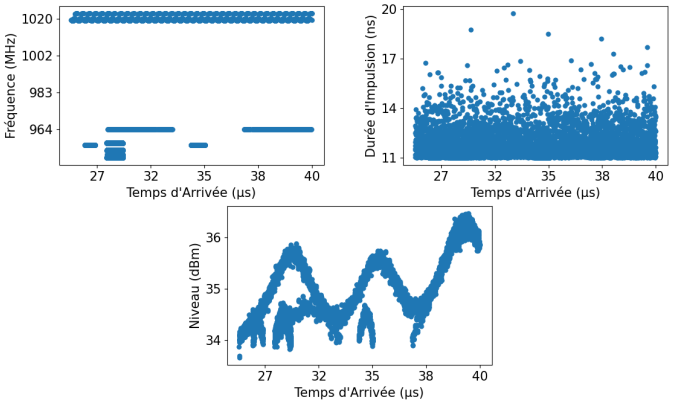

Fig. 12. Signal simulé regroupant les impulsions d'un nombre inconnu d'émetteurs RADAR.

A. Désentrelacement des impulsions

La première étape du désentrelacement consiste à séparer les impulsions à partir de la fréquence et de la durée d'impulsion grâce à l'algorithme de clustering HDBSCAN. La Figure 13 affiche les résultats retournés par HDBSCAN. L'algorithme identifie 7 clusters ainsi qu'une classe d'outliers (-1). Les impulsions des émetteurs 0 et 1 sont correctement séparés dans 2 clusters distincts. En revanche, les impulsions de l'émetteur à 11 ns sont scindées en plusieurs clusters. Il est donc nécessaire de regrouper ces clusters. Le plan (f_n, d_n) illustre la distribution de ces clusters le long du lobe autour après 27 μs.

La Figure 14 affiche les regroupements finaux des clusters effectués à partir du dendrogramme. Le modèle décisionnel identifie quatre ensembles d'impulsions distincts. Le plan (f_n, d_n) montre la présence de deux émetteurs différents transmettant à 11 ns. Notre méthode d'agglomération a été capable de distinguer ces deux émetteurs et stopper les fusions pour ne pas regrouper leurs impulsions.

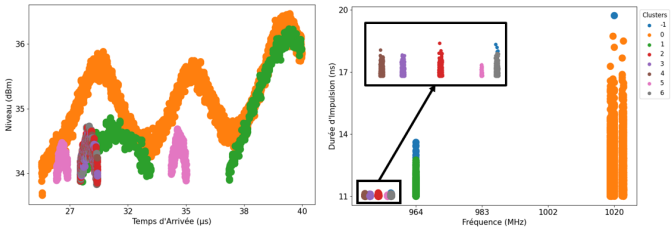

Fig. 13. Résultats du clustering HDBSCAN effectué dans le plan F-DI. Chaque couleur représente un cluster ainsi qu'une classe d'outliers (-1).

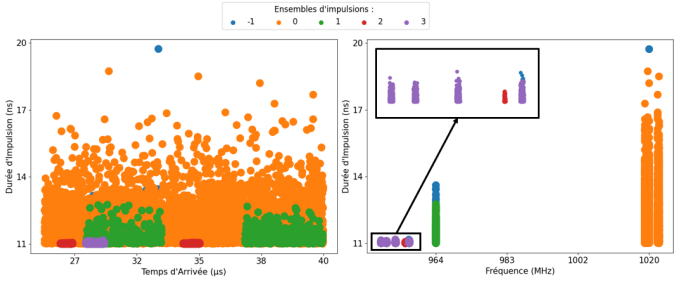

Fig. 14. Regroupements finaux. Les couleurs représentent les ensembles d'impulsions obtenus ainsi qu'une classe d'outliers (-1).

B. Classification de l'ensemble d'impulsion 3

Les résultats de la classification appliquée à l'ensemble d'impulsion 3 sont illustrés par la Figure 15. Le plan (f_n, d_n) de gauche superpose les données et les trois premières classes d'émetteurs identifiées par l'algorithme. Les données de l'output 1 correspondent bien au niveau fréquentiel. En revanche, on aperçoit un décalage sur la durée d'impulsion qui s'explique par une erreur de mesure provenant des capteurs. Dans le cas de notre ensemble d'impulsions, cette erreur est négligeable, d'autant plus que les données sont fréquentiellement variées. La très faible valeur du coût de transport vers la premère classe (< 1) confirme les résultats de l'identification; sur le plan de transport de droite, les données de l'ensemble d'impulsion effectuent très peu de mouvements pour coïncider avec celles de la première classe tandis que ce coût est 19 fois plus élevé pour la deuxième classe la plus proche.

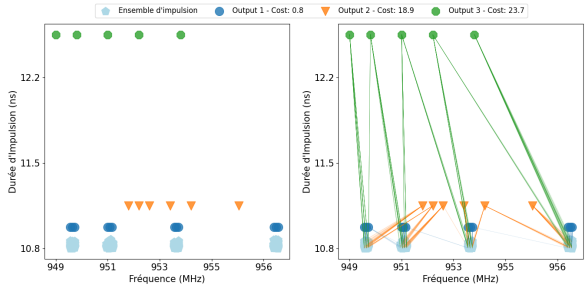

Fig. 15. Résultats de la classification pour l'ensemble d'impulsions 3.

VI. Conclusion et Perspectives

A. Conclusion

Dans cet article nous avons développé un processus de reconnaissance RADAR en deux étapes basé sur le développement et l'utilisation de distances de transport optimal à partir des paramètres primaires des RADARs. L'algorithme HDBSCAN a été appliqué sur de la fréquence et la durée d'impulsion pour séparer les impulsions en différents clusters. Puis, les clusters appartenant à un même émetteur ont été regroupés grâce à l'intégration des distances de transport optimal dans un clustering agglomératif hiérarchique construit à partir du temps d'arrivée et du niveau. Ensuite, en considérant qu'un ensemble d'impulsions correspondait à un seul émetteur, la méthode proposée a permis de classifier l'émetteur grâce à la théorie du transport optimal en deux dimensions à partir de la fréquence et de la durée d'impulsion. Les résultats obtenus sur les données simulées sont très encourageants et permettent d'identifier avec confiance la classe de l'émetteur tout en considérant un grand nombre de classes.

B. Perspectives

Dans la continuité de la méthodologie proposée, nous travaillons actuellement sur l'amélioration de plusieurs axes : tout d'abord, lors de la phase de désentrelacement, nous supposons que les clusters retournés par HDBSCAN contiennent les observations d'un seul émetteur mais il est possible que les RADARs puissent avoir des caractéristiques semblables et être confondus en fréquence et en durée d'impulsion. C'est le cas dans les ports ou les aéroports où plusieurs modèles similaires sont utilisés. Nous travaillons actuellement sur l'amélioration de la séparabilité de ces émetteurs en incorporant d'autres caractéristiques dans le clustering.

Concernant la classification, plusieurs perspectives sont en cours d'investigation pour mieux discriminer les RADARs : premièrement, l'ajout d'une troisième dimension dans le calcul des distances de transport optimal, tout particulièrement la période de répétition des impulsions (PRI), qui est défini comme la différence de temps d'arrivée entre des impulsions successives. Cette caractéristique n'est pas directement utilisable car elle nécessite un prétraitement pour être exploitée : les impulsions manquantes impliquent des distorsions importantes de la densité de probabilité des PRI. Ensuite, comme mentionné dans ce travail et illustré sur la Figure 15, les erreurs provenant de la durée d'impulsion réduisent les performances de la méthode. La robustesse de cette méthode vis-à-vis de telles erreurs doit être améliorée afin de prendre en compte des taux de mitage variables ou encore des erreurs de mesure importantes sur certains paramètres. Enfin, pour proposer une stratégie de classification complète, nous travaillons sur le développement d'une méthode capable de détecter des émetteurs non présents dans la base de données.

Remerciements

ATOS a soutenu ce travail en fournissant les données et son expertise RADAR.

REFERENCES

[1] D. Wilkinson and A. Watson, "Use of metric techniques in ESM data processing," in *IEE Proceedings F (Communications, Radar and Signal Processing)*, vol. 132, no. 4. IET, 1985, pp. 229–232.

[2] H. Mardia, "New techniques for the deinterleaving of repetitive sequences," in *IEE Proceedings F (Radar and Signal Processing)*, vol. 136, no. 4. IET, 1989, pp. 149–154.

[3] D. Milojević and B. Popović, "Improved algorithm for the deinterleaving of radar pulses," in *IEE Proceedings F (Radar and Signal Processing)*, vol. 139, no. 1. IET, 1992, pp. 98–104.

[4] Z. Zhou, G. Huang, H. Chen, and J. Gao, "Automatic radar waveform recognition based on deep convolutional denoising auto-encoders," *Circuits, Systems, and Signal Processing*, vol. 37, no. 9, pp. 4034–4048, 2018.

[5] A. Kawalec and R. Owczarek, "Radar emitter recognition using intrapulse data," in *15th International Conference on Microwaves, Radar and Wireless Communications (IEEE Cat. No. 04EX824)*, vol. 2. IEEE, 2004, pp. 435–438.

[6] J. Lunden and V. Koivunen, "Automatic radar waveform recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 1, pp. 124–136, 2007.

[7] Z.-M. Liu and S. Y. Philip, "Classification, denoising, and deinterleaving of pulse streams with recurrent neural networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 4, pp. 1624–1639, 2018.

[8] Z. Geng, H. Yan, J. Zhang, and D. Zhu, "Deep-learning for radar: A survey," *IEEE Access*, vol. 9, pp. 141 800–141 818, 2021.

[9] L. Ding, S. Wang, F. Wang, and W. Zhang, "Specific emitter identification via convolutional neural networks," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2591–2594, 2018.

[10] M. A. Nuhoglu, Y. K. Alp, and F. C. Akyon, "Deep learning for radar signal detection in electronic warfare systems," in *2020 IEEE Radar Conference (RadarConf20)*. IEEE, 2020, pp. 1–6.

[11] J. Liu, J. P. Lee, L. Li, Z.-Q. Luo, and K. M. Wong, "Online clustering algorithms for radar emitter classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1185–1196, 2005.

[12] S. Apfeld and A. Charlish, "Recognition of unknown radar emitters with machine learning," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 6, pp. 4433–4447, 2021.

[13] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*. M. Dekker New York, 1988, vol. 38.

[14] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[15] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.

[16] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[17] M. Mottier, G. Chardon, and F. Pascal, "Deinterleaving and clustering unknown RADAR pulses," in *2021 IEEE Radar Conference (RadarConf21)*. IEEE, 2021, pp. 1–6.

[18] S. Chakraborty, D. Paul, and S. Das, "Hierarchical clustering with optimal transport," *Statistics & Probability Letters*, p. 108781, 2020.

[19] C. Villani, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[20] N. Bonneel, M. Van De Panne, S. Paris, and W. Heidrich, "Displacement interpolation using lagrangian mass transport," in *Proceedings of the 2011 SIGGRAPH Asia conference*, 2011, pp. 1–12.

[21] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, "POT: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-451.html

[22] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[23] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[24] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[25] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007, pp. 410–420.