MASTER EN SCIENCES INFORMATIQUES À FINALITÉ SPÉCIALISÉE EN DATA SCIENCE

Visualisation et évaluation de projections de données temporelles par temporal dimensional projection

DELVAUX, Alexandre

Award date:
2022

Awarding institution:
Universite de Namur

Université de Namur
Faculté d'informatique
Année académique 2021–2022

Visualisation et évaluation de projections de données temporelles par temporal dimensional projection

Alexandre Delvaux

Mémoire présenté en vue de l'obtention du grade de
Master en Sciences Informatiques.

J'adresse mes premiers remerciements à mon promoteur et maître de stage Benoît Frénay pour l'implication et le temps qu'il m'a consacrés pour ce mémoire. Ses compétences furent pour moi une aide précieuse concernant les expériences qui ont été menées. Ses relectures ont permis de clarifier les sections les plus floues de ce mémoire.

Je souhaite également remercier Marie-Claire Kinet pour l'immense travail qu'elle a fourni concernant la correction orthographique et grammaticale de ce document.

Finalement, un grand merci à Emilie Baijot pour son soutien et sa motivation qui ont représenté une aide psychologique tellement importante !

Table des matières

Glossaire

- Données : Terme générique désignant des éléments non-matériels transportant de l'information.
- Dataset : Ensemble d'instances. Généralement, un dataset regroupe des instances autour d'un même thème. Exemple : Le dataset MNIST est composé d'images de chiffres manuscrits.
- Instance : Occurrence dans un dataset, définissant la nature d'un et un seul objet virtuel ou existant.
- Feature : Attribut quantitatif ou qualitatif composant une instance, elle apporte une information concernant l'instance dont elle fait partie.
- Time step : La notion de time step (étape temporelle) dans un dataset temporel fait référence à l'ensemble des instances récoltées par une ou plusieurs sources à un moment bien précis. Dans un dataset temporel complet, chaque time step est composé du même nombre d'instances. De plus, chaque instance présente dans un time step doit être présente dans tous les autres time steps du dataset. Ces deux contraintes assurent qu'une instance est toujours présente tout au long de la temporalité.
- Temporalité : Une temporalité T est l'ensemble de tous les time steps t définis dans un dataset. L'unité de la temporalité est le time step en lui-même. Pour maintenir une cohérence, tous les time steps consécutifs au sein de la temporalité doivent être présents. Ex : Si un time step représente un jour, et que la temporalité est définie sur un mois entier, alors tous les jours entre le premier du mois et le dernier doivent être présents dans ce dataset. Cette notion a comme conséquence de multiplier le nombre d'occurrences dans un dataset par le nombre de time steps.
- Technique de DP (Dimensional Projection) : Ensemble des techniques de projection dimensionnelle définies dans le domaine du machine learning descriptif.
- TDP (Temporal Dimensional Projection) : Ensemble de techniques de DP spécialisé sur des dataset temporels.
- Basse dimension : Une instance représentable en deux ou trois dimensions est dite en basse dimension, elle est composée de maximum trois features.
- Haute dimension : Une instance non représentable en deux ou trois dimensions est dite en haute dimension, elle est composée d'au moins quatre features.
- Métrique : Score évaluant un critère spécifique d'une technique de projection. La conservation de la dynamique par exemple est un graphique monitorant l'évolution de la vitesse et de l'accélération de chaque instance d'un dataset. Une comparaison entre graphiques permet de distinguer les bonnes des mauvaises techniques de projection.

Abstract

In machine learning, TDP techniques are meant to help visualize time dependent data. For each time step, an individual dimensional projection is generated, and all those projections create a coherent animation based on the evolution of each data points in time. Dynamic t-SNE is a rare case of existing TDP that tends to solve the problem around this kind of data, but techniques are missing to assess the real performance of Dynamic t-SNE and other TDP. Five metrics have been adapted to TDP techniques and are presented in this document. Dynamic and direction shifting conservation, reliability maps, trustworthiness scores and temporal AUClogRNX metrics have been adapted or created. Four new TDP techniques are also described. Among those techniques, TCP and TCP+ overpass the dynamic t-SNE's performances.

Key words: time dependant data, dimensional projection, dynamic t-SNE, assessment, TDP, time step

Résumé

En machine learning, les techniques de TDP représentent des visualisations d'instances de datasets temporellement dépendants. Une série de projections est générée pour chaque time step et permet de créer une animation qui est cohérente avec l'évolution des instances dans le temps. Le Dynamic t-SNE est l'un des rares TDP existants tentant de résoudre les problèmes concernant ce type de données, mais les techniques manquent pour évaluer les performances de cette solution. Cinq métriques adaptées aux techniques de TDP sont présentées dans ce document mesurant les conservations de la dynamique et des changements de direction des instances projetées, construire une reliability map, un tableau de scores de trustworthiness et observer les similitudes dans les voisinages des instances entre time steps. Quatre nouvelles techniques de TDP sont également présentées. Parmi ces nouvelles techniques, le TCP et le TCP+ sont les solutions qui ont fourni des résultats très prometteurs et surpassant les résultats du Dynamic t-SNE.

Mots clés : time dependant data, dimensional projection, dynamic t-SNE, assessment, TDP, time step

1 Introduction

Les open data représentent une source d'information inépuisable, les données qu'elles abritent sont aussi variées les unes que les autres et les exploiter est un défi. Ce défi se complexifie lorsque les données de l'open data ne sont pas statiques dans le temps. Des données, relevées à intervalles de temps régulier provenant de source dynamique (capteurs, données d'étude d'une pandémie, économie d'un pays, etc), vont construire des bases de données temporelles. L'exploitation de données passe souvent par le biais de leur visualisation. Qu'elle soit via un graphique, un diagramme ou encore un tableau de valeurs, la visualisation est à la base de toute exploitation de données.

Dans le monde du machine learning, de nombreuses solutions ont été développées pour faciliter l'exploitation des datasets par des modèles d'exploration. Ces derniers vont recourir à des techniques diverses et variées ayant pour but de réduire la complexité des datasets (réduction dimensionnelle, feature selection etc). Exemple : le t-SNE, le PCA ou encore le MDS.

Malheureusement, peu de solutions existent aujourd'hui pour résoudre le problème de la visualisation de données temporelles. Pire encore, le peu de solutions existantes souffrent d'un manque de métriques les évaluant. La question de recherche de ce mémoire se base sur ce problème : « Quelles nouvelles solutions peuvent être développées afin d'améliorer les techniques existantes de visualisation de données temporelles et de quelles manières est-il possible d'évaluer la qualité de projection de telles techniques ? »

Ce mémoire introduit en chapitre 2 les techniques de DP et explique comment ces algorithmes sont utiles à la visualisation de données hautement dimensionnelles. Ensuite, une partie sur les résultats de recherche dans la littérature concernant les temporal dimensional projection techniques est présentée développant l'état de l'art sur lequel repose la suite du mémoire dans le chapitre 3. Une présentation des métriques existantes évaluant la qualité d'une projection dimensionnelle termine la phase de recherche. Elles sont au nombre de trois, le score de trustworthiness, la Reliability map et le score AUClogRNX.

La première contribution de ce mémoire sera ensuite introduite par la description des nouvelles métriques développées à l'occasion. Elles sont au nombre de trois : la conservation de la dynamique, la conservation des changements de directions et le temporal AUClogRNX. La phase d'expérience sur l'état de l'art permet d'établir une base de résultats utiles à la comparaison des nouvelles solutions. Chaque projection effectuée est évaluée à l'aide des métriques.

Les nouvelles solutions constituent la deuxième contribution de ce mémoire. Quatre nouvelles solutions développées pendant le stage sont décrites et expérimentées dans les chapitres 6 7, 8 et 9. Une phase de vérification des résultats et des conclusions tirées sur des datasets de contrôle suit au chapitre 10. Ces datasets de contrôle sont décrits dans le chapitre 10. Les métriques présentées servent de base de comparaison entre chaque solution. Une analyse finale est réalisée résumant les tenants et aboutissants de chacune des techniques dans le chapitre 11.

2 Projection dimensionnelle

Quand des données sont collectées, nettoyées et stockées, ces dernières deviennent généralement des instances multi-dimensionnelles au sein d'un dataset. Cela signifie que le nombre de features attribués à chaque occurrence de la base de données est supérieur à 1. Une fois les données récoltées, il est intéressant pour l'entreprise ou le centre de recherches de pouvoir en tirer une plus-value à des fin marketing, pour faire avancer la recherche, ou tout simplement pour en tirer des informations relatives au comportement de masse de ces dernières. Un ou plusieurs outils permettant la visualisation de ces données sont nécessaires afin d'en tirer les bonnes interprétations.

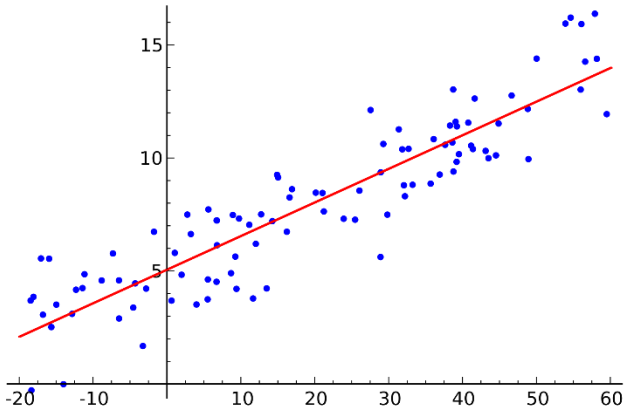


Figure 1 Régression linéaire sur un nuage de points bi-dimensionnel [1]

Lorsque le nombre de features est égal à deux ou trois, le problème ne se pose pas vraiment. En effet, un outil très connu permet de mettre en relation une valeur « x » avec une autre « y » et parfois une 3^{e} « z » … les graphiques (Figure 1). S'ils sont bien réalisés, l'entièreté de l'information contenue dans les données est représentée en une fois sur un seul support. De plus, avec des techniques de visualisation avancées, de nouveaux canaux d'information permettent d'accroître le nombre de features qu'un graphique communique. La couleur des instances, leurs tailles ou encore leurs formes peuvent inférer une information complémentaire quant à la classification de ces dernières.

Un problème persiste… Comment faire avec des données « hautement » dimensionnelles ? Par hautement dimensionnelles, c'est de 10, 100 ou encore 1000 dimensions dont il est question. Aucune solution mentionnée précédemment ne permet de résoudre cette situation où « beaucoup trop » d'informations sont présentes. De nombreuses techniques existent et permettent de répondre à cette problématique. Une première approche est d'effectuer de la feature selection. Cette méthode détermine les deux à trois features les plus importantes et les sélectionne comme étant les features les plus représentatives de la masse. Le problème est qu'il est peu fiable de faire confiance à des résultats provenant de trois features quand une instance est définie par plusieurs centaines d'entre elles. Il faut donc utiliser une technique permettant de « projeter » les données d'une dimension N vers une dimension inférieure afin de retomber dans un univers où il est possible d'exploiter les instances et où la totalité des informations initiale est considérée. Le défi réside dans le fait de minimiser la perte d'informations inévitable lors de la projection. Par analogie, une projection de données est équivalente à interpréter la complexité d'une structure en trois dimensions, mais en observant uniquement l'ombre projetée de l'objet sur un mur. En fonction de la qualité de la projection, certaines informations concernant l'objet en haute dimension pourront être déduites. Une attention toute particulière doit être émise lors de l'interprétation de ces ombres. Comme sur la Figure 2

provenant du jeu vidéo mobile « Shadowmatic », il est facile d'interpréter l'ombre du mur comme étant celle d'un lapin, mais la réalité peut être tout autre.

Figure 2 projection d'une forme complexe sur un mur [2]

Les méthodes utilisant des techniques de projection sont toutes bien différentes dans leur fonctionnement ; leurs avantages et inconvénients varient fortement de ceux venant de la feature selection. L'avantage premier d'une technique de dimensional projection (DP) est sa capacité de rendre toute instance hautement dimensionnelle exploitable dans un espace bi ou tri-dimensionnel. Il est intéressant de constater que la projection de données venant d'un espace tri-dimensionnel dans un espace bi-dimensionnel sera bien plus qualitative qu'une projection de 100 dimensions en 2. Donc, l'inconvénient principal des DP est lié à leur fiabilité. En effet, des clusters peuvent se former en basse dimension alors que ces derniers sont loin d'exister en haute dimension et inversement. Des métriques de fiabilité peuvent toutefois aider à déceler ces problèmes de pertinences. Ce mémoire s'est penché sur la technique DP « t-SNE » (t-Distributed Stochastic Neighbor Embedding) d'écrit dans la section suivante.

2.1 Cas spécifique : le t-SNE

Cette section décrit le t-SNE, son implémentation, son fonctionnement et ses problématiques. Une explication vulgarisée de la technique débute cette section et, afin de mieux introduire cette DP, les techniques ayant menés à sa conception suivent cette vulgarisation.

2.1.1 Explication vulgarisée

Le t-SNE [3](t-Distributed Stochastic Neighbor Embedding) est une méthode de réduction dimensionnelle basée sur la probabilité de voisinage. Comme son nom l'indique, il se base sur une méthode permettant de détecter les voisins proches en haute dimension de chaque instance et de projeter ces instances en basse dimension tout en conservant leurs voisins. Il est considéré que deux points sont voisins en haute dimension si la distance qui les sépare est en dessous d'un certain seuil.

Le nuage de points généré par t-SNE sert de base pour interpréter le comportement des données en haute dimension. Il est important de noter que seul le voisinage proche est conservé, signifiant que deux instances proches, mais ne faisant pas partie du même cluster, ne seront pas nécessairement proches en haute dimension et pourraient en réalité être fortement éloignés. Sur la

Figure 3, les deux instances entourées sont proches, mais n'appartiennent pas au même cluster. Il est probable que les deux n'aient pas beaucoup en commun (en considérant que la projection soit réalisée par t-SNE).

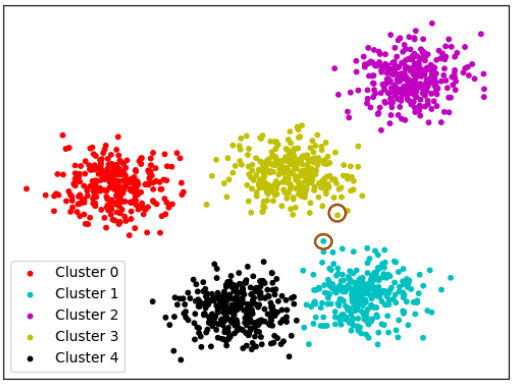


Figure 3 Exemple de clusters [4]

2.1.2 SNE

SNE [3] (Stochastic Neighbor Embedding) est la technique à l'origine du t-SNE, voici comment il s'implémente. L'équation $p_{j|i} = \frac{\exp\left(-\left||x_i - x_j|\right|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} exp(-||x_i - x_k||^2 / 2\sigma_i^2)}$ calcule la similarité entre deux points x_i et x_j et se définit comme étant la probabilité conditionnelle notée $p_{j|i}$ de choisir x_j (instance du dataset) comme voisin de x_i (une autre instance du dataset différente de x_j). En construisant une gaussienne centrée sur x_i, plus x_j est proche, plus cette probabilité conditionnelle va être élevée. Si $x_j = x_i$, la probabilité, bien que maximale, sera définie comme nulle car il n'est pas intéressant de considérer les paires d'instances similaires. C'est ainsi que toutes les instances du dataset peuvent être classées dans l'ordre de leur probabilité conditionnelle par rapport à x_i et considérées comme voisines si elles sont dans les k plus probables (k étant défini par la perplexité du SNE). Le terme σ est le paramètre définissant « l'étendue » de la gaussienne, plus σ est grand, plus les points éloignés vont être rapprochés. Dans les zones denses du nuage de points HD, un plus petit σ est préférable, alors qu'un plus grand est adapté aux zones moins peuplées d'instances. La perplexité est un paramètre déterminé par l'utilisateur. Cette valeur de perplexité induit directement la valeur du sigma. Dans la littérature, la perplexité est définie comme le paramètre déterminant le nombre de voisins moyens qu'aura chaque point dans la projection LD.

Le calcul des similarités s'effectue également en basse dimension avec les probabilités conditionnelles $q_{i|j}$ représentant la probabilité de trouver le point y_j comme étant voisin du point y_i dans la projection (y_i représente l'instance x_i projetée). Cette similarité se calcule à l'aide de l'équation $q_{j|i} = \frac{\exp\left(-\left||y_i - y_j|\right|^2\right)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$.

L'apprentissage d'un SNE est la phase permettant à la technique d'améliorer son modèle de projection en manipulant les positions de chaque instance dans la direction qui réduira le coût du t-SNE. Le coût *C* d'un t-SNE est défini par la formule de la divergence de Kullback-Leibler comparant les probabilités conditionnelles en haute dimension et en basse dimension. À chaque itération d'apprentissage, appelée époque, le SNE cherche à réduire ce coût afin d'obtenir une projection la plus

fiable et représentative possible. La formule de la divergence Kullback-Leibler se définit comme $C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$. Le but est d'obtenir, pour toutes les paires de points, une probabilité conditionnelle en basse dimension qui est équivalente à celle en haute dimension, de telle sorte que la fraction dans le logarithme se neutralise tendant vers 0. Cette neutralisation va réduire le coût global du SNE.

Pour réduire le coût du SNE, un apprentissage par descente de gradient prenant la forme $\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$ doit s'effectuer. Cette équation, lorsqu'elle est appliquée à chaque instance projetée, va calculer un vecteur de déplacement à ajouter à l'ancienne position de l'instance pour la déplacer vers un endroit plus semblable à sa position relative aux autres instances en haute dimension. Ce vecteur va tout de même passer par une autre fonction d'optimisation incluant les notions de momentum « α » et de learning rate « η ».

Le but premier de l'équation $Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$ est d'accélérer l'apprentissage du SNE et surtout d'éviter les minimums locaux. Une autre optimisation utilisée pour éviter les faibles minimums locaux est d'appliquer un bruit gaussien aux instancese projetées à chaque itération d'apprentissage en en réduisant l'intensité au fil de l'apprentissage. Ce bruit va légèrement déplacer les instances et améliorer l'apprentissage du SNE.

2.1.3 t-SNE

Pour passer du SNE au t-SNE, quelques modifications ont été nécéssaires. La simplification de la fonction de coût est possible en modifiant les formules de similarité en utilisant des relations de probabilités combinées. Sans plus de détails, cette optimisation permet de faciliter le calcul de descente de gradiant et augmente légèrement la vitesse de calcul de la technique.

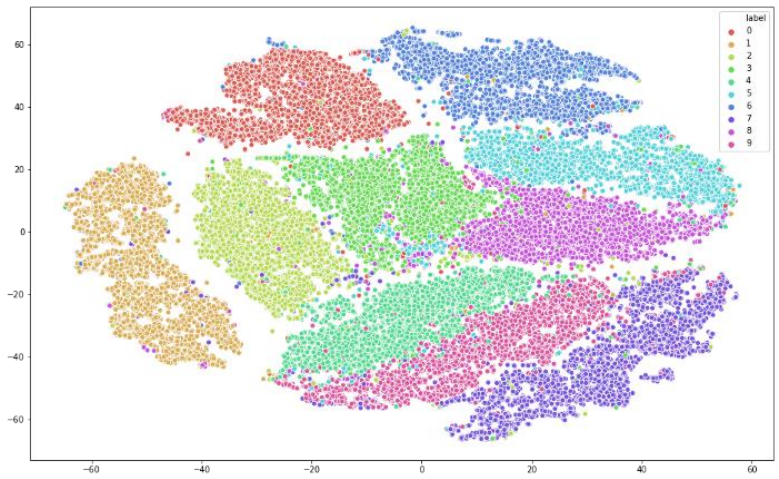


Figure 4 Visualisation du dataset MNIST projeté par t-SNE [5]

Un problème que tente de résoudre t-SNE est le problème très commun en réduction dimensionnelle, le « crowding problem ». Ce problème repose sur le fait que passer d'une dimension à une plus petite pose des soucis d'espace et d'arrangement. Si, dans un espace hautement dimensionnel, une dizaine d'instances sont réparties de manière équidistante, il sera impossible de les représenter parfaitement en deux dimensions.

Le t-Distributed Stochastic Neighbor Embedding (t-SNE) fonctionne selon les principes de base du SNE, mais tente de contourner les problèmes d'optimisation et de crowding précédemment cités. En basse dimension, la fonction gaussienne utilisée pour calculer les similarités a été remplacée par une distribution student-t considérée comme plus appropriée.

2.1.4 Problématique du t-SNE

La problématique lors de l'utilisation du t-SNE intervient dans le contexte où le dataset reçu inclut une nouvelle dimensionnalité bien particulière : la temporalité. Cette dernière peut très difficilement être manipulée comme une dimension « spatiale ». De plus, des méthodes permettant de manipuler des données temporelles manquent terriblement et il est très intéressant de pouvoir visualiser des dataset temporels hautement dimensionnels évoluant dans le temps.

Jusqu'à présent, une approche simple, mais très inefficace est d'effectuer un t-SNE par tranche temporelle (Time step) et d'afficher les différentes projections en séquence pour en faire une animation. Le problème flagrant est le manque de cohérence qui apparaît entre les images (Figure 5).

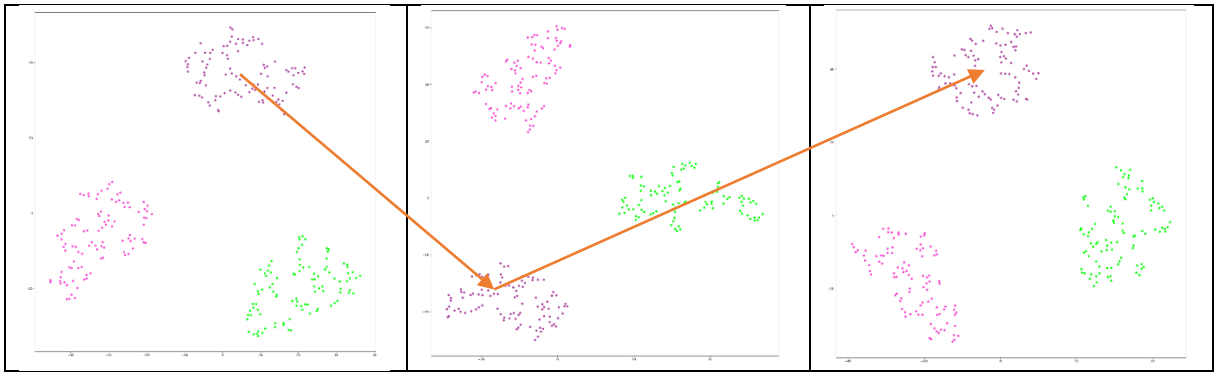


Figure 5 Le t-SNE avec des données temporelles. Les 3 images correspondent à trois time steps consécutifs

Dans l'exemple de la Figure 5, trois clusters sont clairement distincts. Les trois figures ci-dessus représentent l'évolution de ces trois clusters sur trois time steps consécutifs. Néanmoins, il est très complexe de déterminer où chaque instance se positionne d'un time step au suivant. De plus, une information trompeuse est communiquée, l'impression que les instances se déplacent rapidement est fortement présente, les flèches orange témoignant du phénomène. Dans la réalité, chaque instance présente sur les projections ne se déplace que de quelques pas dans une direction linéaire et non aussi brusquement que montré sur les figures. Avec des datasets bien plus peuplés et plus complexes dans leur structure, la solution devient complètement inutilisable et, pire encore, communique des comportements erronés sur le déplacement des données.

2.2 Visualisation de la problématique

À la page suivante se trouve deux tableaux représentants les 12 premières images d'une animation obtenue lors de la création d'un t-SNE sur chaque time step de deux datasets différents et mis en séquence chronologique. Un grand désordre s'observe sur ces deux animations et il est très peu utile et conseillé de travailler avec ce genre d'animation.

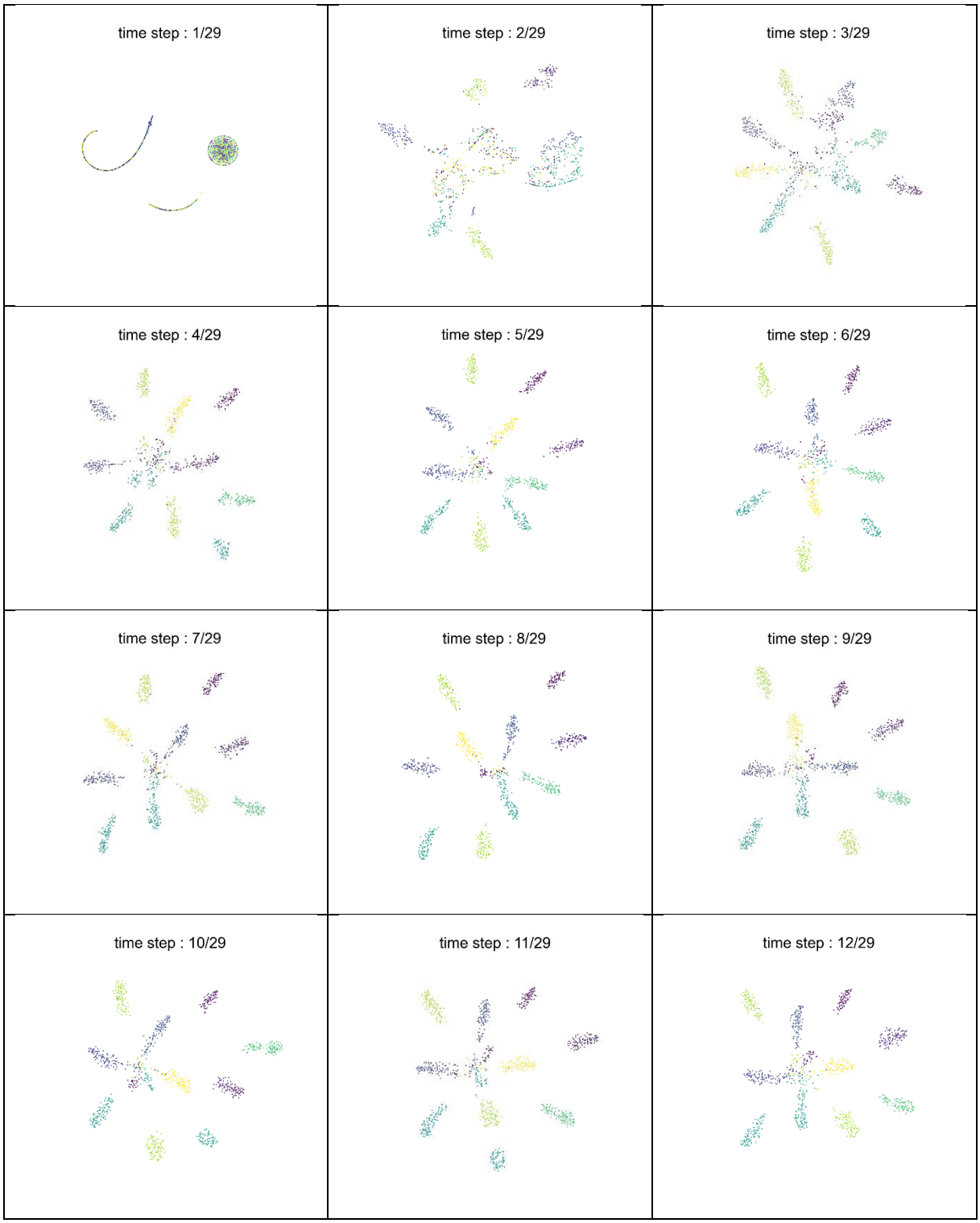


Figure 6 Animation d'une succession de t-SNE sur le dataset SVHN

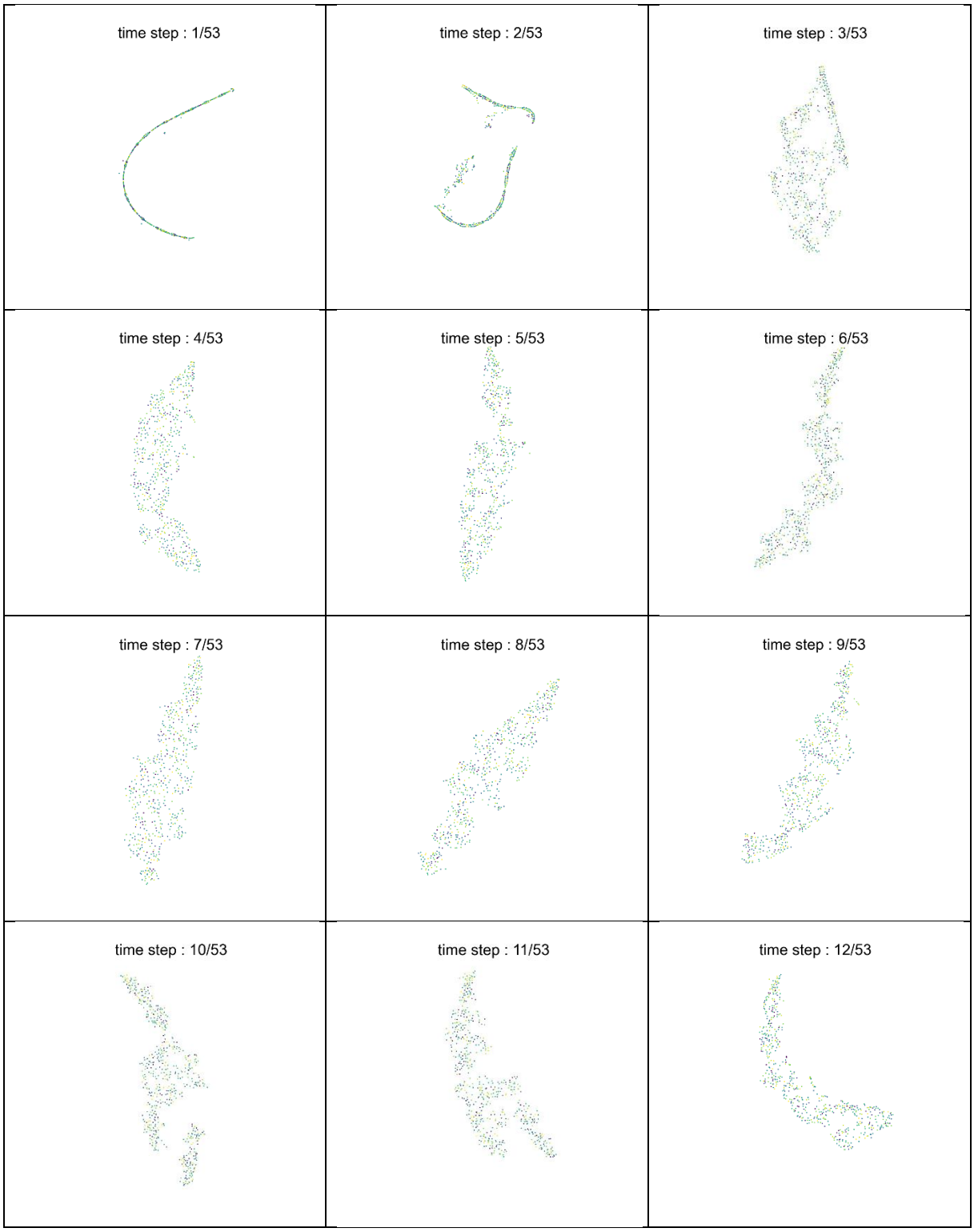


Figure 7 Animation d'une succession de t-SNE sur le dataset du Covid

3 Les techniques « temporal dimension projection » TDP

Les dimensional projections ne sont pas capables de traiter les datasets temporels correctement, comme démontré dans le chapitre 2.1.4. Il est donc nécessaire d'introduire un nouveau domaine de projection dimensionnelle. Ce domaine est orienté sur des datasets possédant des instances temporelles et fournissant des TDP (temporal dimensional projection).

3.1 TDP

Il est important de distinguer les instances temporelles des time series. Une time series est liée à une loi de probabilité ou de statistique. Ce qui implique qu'il est possible de déduire mathématiquement le comportement passé et futur de telles données. La détection de tendance et de pattern périodique est faisable avec une time series. Par exemple, il est possible de prédire la densité du trafic routier d'une autoroute car cette variable est soumise à un facteur de périodicité. Les lundis matin de chaque mois se ressembleront, ainsi que les autres jours de la semaine suivant les variations périodiques des conducteurs de cette route. Une instance temporelle, cependant, n'est aucunement dépendante d'une loi mathématique. Il n'est pas facile, voire impossible de la prédire car aucune tendance ni aucun pattern n'apparaît dans son évolution au sein de sa temporalité.

Les techniques TDP représentent un tout nouveau domaine dans les modèles d'exploration. L'objectif de ces techniques est d'aider à fournir une visualisation permettant d'observer l'évolution des données au fil de leur temporalité tout en essayant de respecter un maximum la structure des données et leur dynamique en haute dimension. Une seule technique existe pour l'instant et se base sur l'adaptation du t-SNE pour traiter les instances temporelles.

3.2 Dynamic t-SNE

L'un des rares TDP dans la littérature est le dynamic t-SNE [6]. Cette technique débute par la réalisation un t-SNE pour chaque time step de la temporalité. En parallèle, afin de tisser un lien de cohérence entre chaque t-SNE consécutif, le dynamic t-SNE va calculer, par différence finie, la position des instances entre chaque time step. Cette expression a pour objectif de contrecarrer l'initialisation aléatoire d'un t-SNE rendant chaque visualisation complètement indépendante l'une de l'autre.

La méthode du dynamic t-SNE est sensible à la modification des paramètres suivants :

- Sa perplexité : paramètre hérité du t-SNE ;
- Le nombre d'itérations pour son entraînement : combien de fois faut-il itérer l'entraînement du modèle pour s'assurer de sa qualité. Une valeur par défaut de 1000 itérations est définie dans l'état de l'art ;
- Son λ : valeur scalaire permettant de manipuler l'intensité du terme de pénalisation dans le calcul de coût final du dynamic t-SNE.

La méthode dynamic t-SNE débute par positionner aléatoirement les instances pour chaque time step. Ces positions vont ensuite être modifiées au fil des époques en fonction du coût des projections. À la fin de l'entraînement, au bout des 1000 itérations, il est considéré que le résultat fourni par le dynamic t-SNE aura convergé vers une projection plus fiable. Ce choix de 1000 itérations est considéré comme suffisant pour la plupart des datasets, mais il reste possible que l'entraînement puisse être plus court ou qu'il ait besoin de se prolonger.

La formule calculant le coût du dynamic t-SNE se définit comme suit :

$$C = \sum_{t=1}^{T} C[t] + \frac{\lambda}{2N} \sum_{i=1}^{N} \sum_{t=1}^{T-1} ||p_i[t] - p_i[t+1]||^2$$

Où « *C* » est la somme de deux termes. $\sum_{t=1}^{T} C[t]$ Représente la somme des coûts respectifs de tous les t-SNE de chaque time step t. La fonction $C[t]$ n'est rien d'autre que la fonction de coût du t-SNE définie dans le chapitre 2.1.3. Dans les faits, cette partie n'est rien d'autre qu'un apprentissage généralisé à toute la temporalité des t-SNE, chaque coût de chaque t-SNE est ajouté au coût total du dynamique t-SNE. Le problème est que si cette partie est considérée comme suffisante, chaque t-SNE sera entraîné indépendamment les uns des autres rendant le positionnement de leurs points indépendant d'un time step à un autre. Ce comportement rendra la visualisation complètement illisible et sera similaire au comportement observé sur la Figure 5. Le deuxième terme de l'équation doit être ajouté, $\frac{\lambda}{2N} \sum_{i=1}^{N} \sum_{t=1}^{T-1} ||p_i[t] - p_i[t+1]||^2$ est modulé par $\frac{\lambda}{2N}$ et se présente comme la somme des vitesses que chaque instance subit entre chaque time step. $p_i[t] - p_i[t+1]$ Calcule la vitesse entre deux time step d'une instance. Ce deuxième terme, quand il reçoit une importance suffisante (grâce au λ) va empêcher les points de prendre trop de vitesse entre chaque t-SNE consécutif et donc, maintenir une structure globale stable entre chaque itération de la temporalité. Un λ trop petit donnerait plus d'importance au premier terme de l'équation et diminuera le contrôle qu'a dynamic t-SNE sur la dynamique des instances. Cependant, s'il est trop important, le coût de la fonction ne dépendra plus que du terme de pénalisation, le gradient ne se calculera plus correctement et l'entraînement ne se terminera pas. Cette régularisation permet de visualiser plus facilement la structure des données tout au long de la temporalité et il est donc très important de bien choisir le λ. Par défaut, dans la littérature, ce λ est fixé à 0.1.

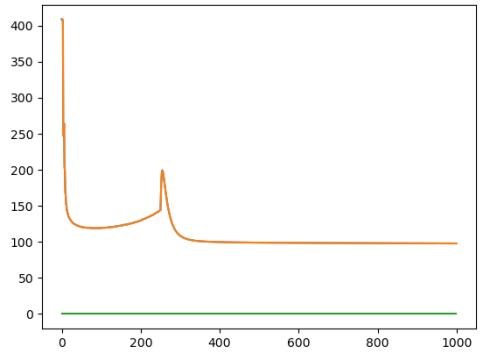


Figure 8 évolution du coût (courbe orange) du dynamic t-SNE lors de l'apprentissage sans pénalisation. La courbe verte représente l'évolution du terme de pénalisation (constant et nul dans cet exemple).

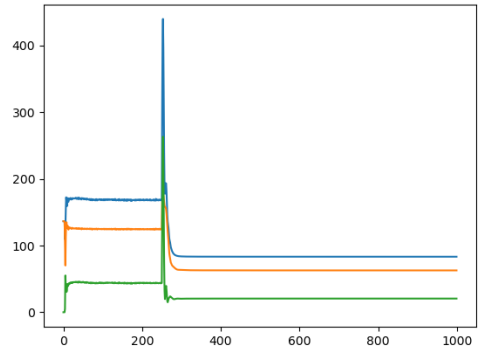


Figure 9 évolution du coût (courbe bleu) du dynamic t-SNE lors de l'apprentissage avec pénalisation. La courbe bleue est la somme du terme de pénalisation (vert) et de la somme des coûts des t-SNE (orange).

Lors de l'entraînement du dynamic t-SNE, les paramètres du learning rate et du momentum ont été configurés avec une valeur de 2400 et de 0.5. L'importance de cette information réside dans le fait que ces valeurs changent dès la 250^{e} itération atteinte. À ce moment, le learning rate passe à 250 et le momentum à 0.8, ce changement permet d'accroître la puissance d'apprentissage et permet d'écraser le score du coût de la fonction objective. Sur les deux figures ci-dessus, l'évolution du coût

n’est que brièvement impacté lorsqu’aucun λ ne pénalise l’apprentissage, mais dès lors qu’une pénalisation intervient, l’apprentissage subit une explosion du coût de la fonction objective lorsque les paramètres d’apprentissage changent. La situation converge ensuite vers un coût bien plus efficace qu’avec les précédents paramètres. L’interprétation et la comparaison de la valeur numérique du coût est à discuter, deux apprentissages n’utilisant pas la même fonction de coût ne sont pas propices à être comparés, donc ce n’est pas sur la valeur directe du coût que la qualité d’une projection sera évaluée, mais sur l’interprétation de plusieurs métriques décrites dans les chapitres 4.2 et 4.3.

Le dynamic t-SNE est une technique améliorable. De nouvelles techniques ont été développées pour contrecarrer les défauts du dynamic t-SNE et fournissent une autre approche s’avérant parfois plus intéressante sur des aspects non négligeables d’un TDP tel qu’une meilleure fiabilité ou une meilleure cohérence de conservation de voisinage. Le chapitre suivant définit le setup expérimental et les métriques qui constituent la base de l’évaluation de chaque technique.

4 Setup expérimental

Cette section aborde les différents préparatifs pour réaliser les expériences : les données utilisées, le programme développé pour visualiser les projections de ces données et les métriques instaurées pour évaluer les expériences.

4.1 Datasets utilisés

Différents datasets ont été sélectionnés ou créés afin de réaliser tous les tests avec la technologie décrite par l'état de l'art et les nouvelles solutions. Il y en a au total 6.

4.1.1 Dataset Covid

La situation sanitaire actuelle a engendré la récolte de beaucoup de données en Belgique. Ses communes ont enregistré, depuis son déploiement, l'évolution de la vaccination de la population. Un dataset a été formé afin de tester les différentes solutions sur ce genre de données. Ce dataset est constitué de plus de 500 occurrences (chaque commune belge) sur une cinquantaine de semaines. Les données sont récupérées sur le site epistat [7] puis traitées à l'aide d'un programme développé en Python. Chaque instance présente les features affichées en vert dans le tableau ci-dessous.

Commune	Time step	% population vaccinée avec la première dose	% population vaccinée avec la deuxième dose	% population vaccinée avec la troisième dose

Les trois features sont basées sur des événements cumulatifs, la vaccination ne va faire qu'augmenter dans le temps et va tendre vers un état final. Les TDP tirés de ces données devraient faire apparaître ce phénomène.

4.1.2 Dataset SVHN

Figure 10 Échantillon de 100 images présentes dans le dataset SVHN [8]

SVHN [8] est un dataset connu. Il est l'acronyme de « street view house number » (numéro de maison de google street view). Il référence plus de 60.000 images de numéros de maisons. Ce type de

dataset est utile pour s'entraîner sur la confection de réseaux de neurones ou autres modèles de classification car sa grande diversité et ses données déjà préparées (labellisé et structuré) permet à ces modèles d'être très efficaces et donc pédagogiques. Le dataset MNIST reprenant plusieurs milliers d'images de chiffres écrits à la main a également été sélectionné pour passer les tests, mais ce dernier, considéré comme trop académique, a laissé sa place à SVHN. Une question reste encore à définir : comment ces données peuvent-elles être temporelles ?

Un CNN (réseau neuronal à convolutions) a été développé en parallèle s'entraînant sur ces images et fournissant un modèle permettant d'identifier efficacement le numéro présent sur chacune d'entre elles. Pour se faire, plusieurs époques d'entraînement ont été nécessaires et à chacune d'entre elles, la valeur de sortie des 128 neurones de l'avant dernière couche du CNN a été récupérée et annexée avec la valeur attendue sur l'image et le numéro de l'époque d'entraînement pour former une occurrence du dataset. Afin de convenir d'une efficacité et d'une optimisation de temps, 1000 des 60000 images ont été gardées afin d'éviter d'avoir un dataset trop volumineux pour les entraînements.

Le but derrière l'utilisation du SVHN est d'essayer d'observer l'évolution de la classification des images d'un CNN au fil de son entraînement et d'évaluer la capacité de toutes les solutions développées dans ce travail.

4.1.3 Les datasets « blob »

Les datasets « blob » sont 4 datasets artificiels ayant chacun pour but de simuler une caractéristique spécifique au niveau de la dynamique des données. Le blob est un nuage de point généré aléatoirement par une fonction appelée « make_blob » définie dans le package « sklearn.datasets » de scikit-learn. Les arguments que prenne cette fonction sont les suivants : une liste contenant le nombre de points de chaque nuage, la dimension de l'espace de ces points, une liste des positions des centroïdes de chaque nuage (un centroïde est l'équivalent du centre de gravité des points, un peu comme le point moyen représentant la foule).

Ces datasets factices interviendront à la fin des expérimentations de toutes les solutions implémentées pour vérifier les hypothèses relevées lors de l'analyse des résultats de chaque métrique.

Les quatre datasets « blob » sont tous composés de trois nuages de points en trois dimensions et chacun introduit une dynamique unique de déplacement de points dans le temps. Voici les caractéristiques de chacun :

- Blob_MRU : Les nuages se déplacent en MRU (mouvement rectiligne uniforme). Une vitesse constante et linéaire permet de cerner la conservation du mouvement au cours de la projection. Les trois nuages intervertissent leur position ;

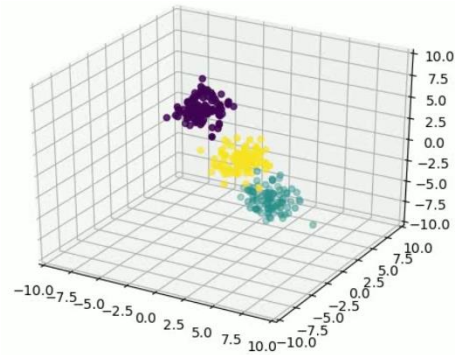


Figure 11 Etat initial des instances de Blob_MRU

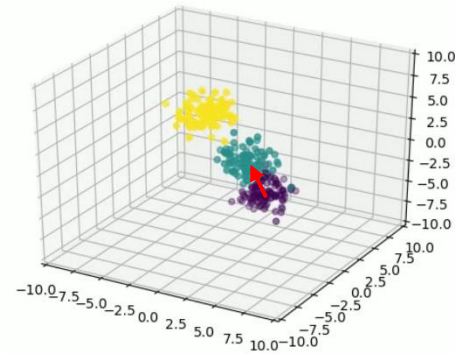


Figure 12 Etat final des instances et trajectoire (nuage vert) de Blob_MRU

- Blov_A : Les nuages se déplacent en MRUA (mouvement rectiligne uniforme accéléré). Une accélération uniforme et unique est appliquée à chaque nuage qui s'éloigne/se rapproche des autres. Ce dataset cerne la conservation de l'accélérations des points lors de la projection ;

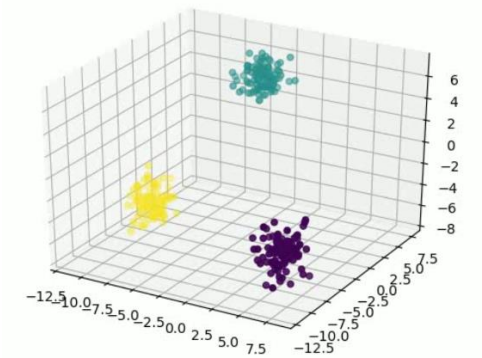


Figure 13 Etat initial des instances de Blob_MRUA

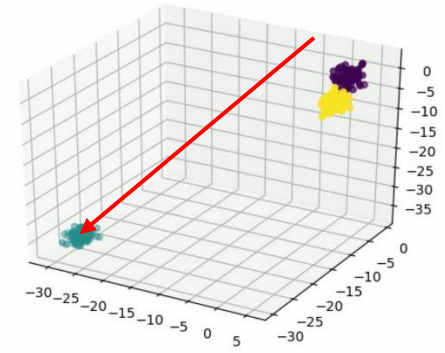


Figure 14 Etat final des instances et trajectoire (nuage vert) de Blob_MRUA

- Blob_MCu : Les nuages se déplacent en MCA (mouvement curviligne accéléré). Chaque nuage de points tourne autour d'un centre de gravité avec une orbite circulaire ou elliptique. Cette méthode permet de cerner la conservation du changement de direction des points lors de la projection ;

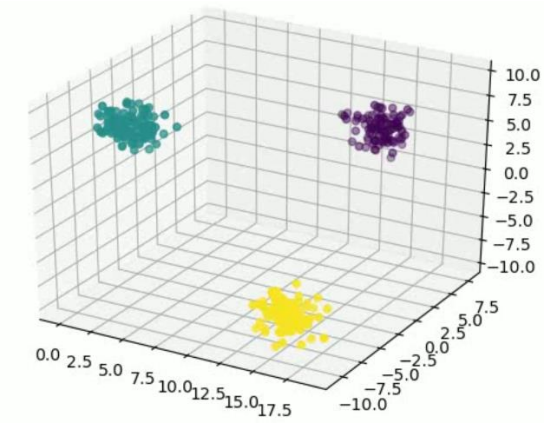


Figure 15 Etat initial des instances de Blob_MCu

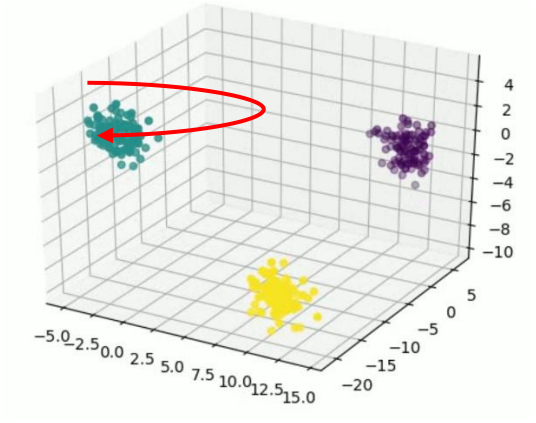


Figure 16 Etat final des instances et trajectoire (nuage vert) de Blob_MCu

- Blob_Rand : Les nuages se déplacent de manière erratique. Il n'y a pas de structure ou de logique entre le déplacement ; ceux-ci sont aléatoires et d'intensité variable. Ce dataset permet de détecter le comportement des différentes solutions face à un comportement chaotique des données.

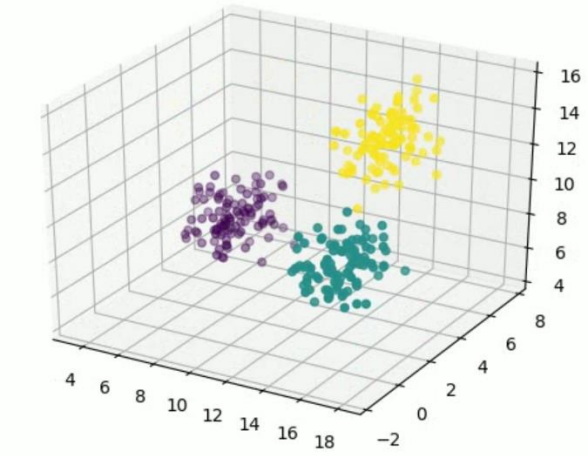


Figure 17 Etat initial des instances de Blob_Rand

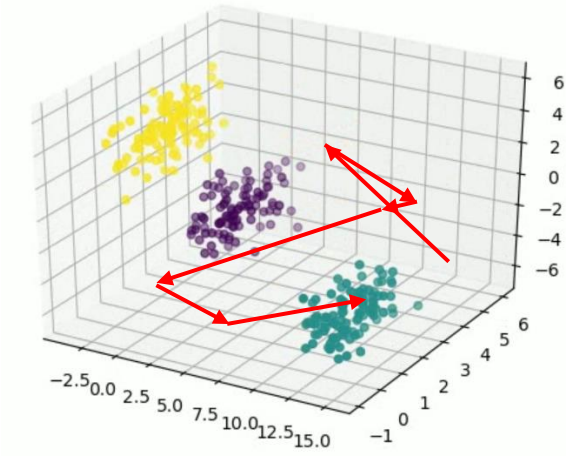


Figure 18 Etat final des instances et trajectoire approximée (nuage vert) de Blob_Rand

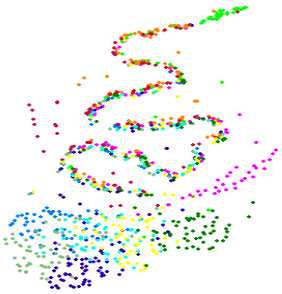
4.2 Métriques existantes évaluant la qualité d'une DP

La littérature présente peu de méthodes d'évaluation d'une DP. Quand une projection doit être évaluée, c'est généralement de méthodes d'évaluation de la qualité des clusters dont il est question [9] [10]. De plus, les techniques de TDP étant un domaine très spécifique des techniques de projection, le travail de recherche devient encore plus maigre et il est important d'inventer ses propres métriques.

Deux catégories de métriques ont été définies. Les métriques initialement prévues pour évaluer les DP existantes dans la littérature et décrites dans les sections suivantes et les métriques développées uniquement pour les TDP décrites dans la section 4.3.

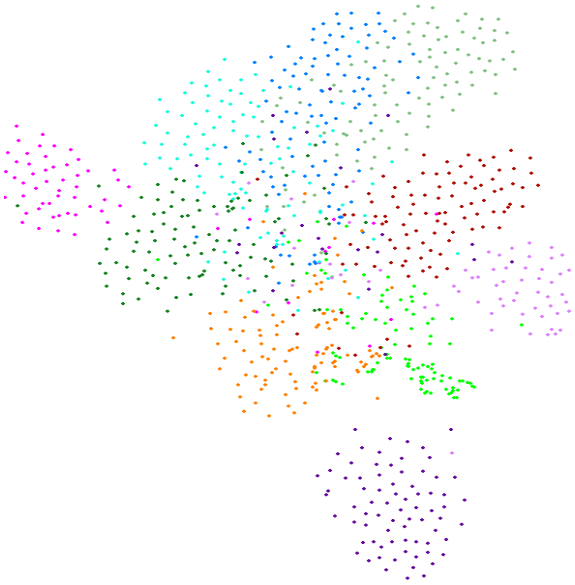
4.2.1 Trustworthiness

Le trustworthiness [11] (fiabilité en français) calcule le score de fiabilité d'une DP. Le score de fiabilité d'une projection est une valeur comprise dans l'intervalle [0,1] et informe l'utilisateur de la fiabilité de retranscription des voisins de chaque instance lors de la projection. Une fiabilité de 1 indique une retranscription parfaite des voisins, ce qui représente une valeur très difficile à atteindre, mais y tendre fortement reste possible. Dans le cas d'instances multidimensionnelle, le calcul de la distance euclidienne détermine la distance entre deux instances. La fonction T(k) calculant le score du trustworthiness est définie comme $T(k) = 1 - \frac{2}{Nk(2N-3k-1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i,j) - k)$, où r(i,j) est une fonction calculant le rang d'un point j selon sa distance avec le point i. Le point j fait partie de l'ensemble $U_k(i)$ des points étant k-voisin (k points les plus proches) de i dans la projection, mais n'étant pas k-voisins en haute dimension. $U_k(i)$ représente donc tous les points n'étant pas censés être aussi proches du point i dans la projection. La valeur k choisie lors des expérimentations a été définie à 20% du nombre d'instances dans le dataset (choix arbitraire). Ci-dessous sont représentées respectivement une visualisation fiable et une beaucoup moins du premier time step du dataset SVHN.



Trustworthiness : 0.8729843602573266

Figure 19 Projection d'instances à bon score de trustworthiness



Trustworthiness : 0.7693157258064516

Figure 20 Projection d'instances à mauvais score de trustworthiness

4.2.2 Reliability map

La métrique suivante s'appelle la reliability map [12]. Elle est similaire à la trustworthiness. Elle sert également à estimer la fiabilité d'une projection, mais utilise comme support la projection pour représenter la fiabilité des voisins de chaque instance. Il s'agit d'une métrique graphique.

Visualisation classique	Visualisation de la reliability map

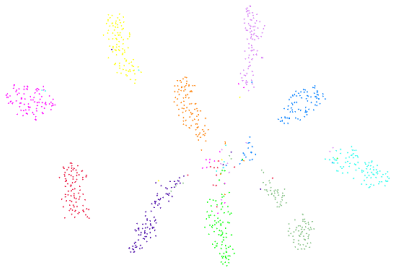
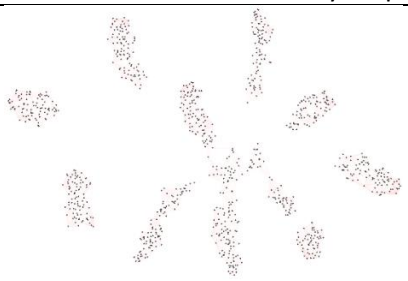

	

Figure 21 Projection sans pénalisation sur le dataset SVHN au time step 10/29	*Figure 22 Reliability map sur une TDP du dataset SVHN sans pénalisation time step 10/29*

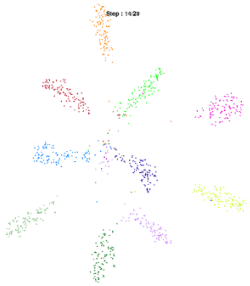
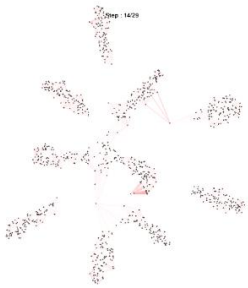
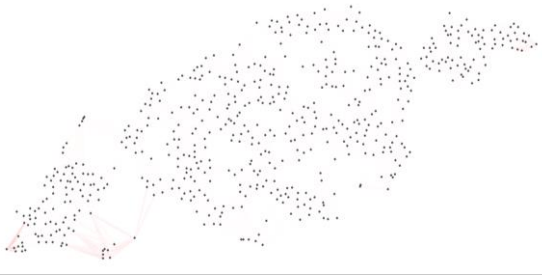
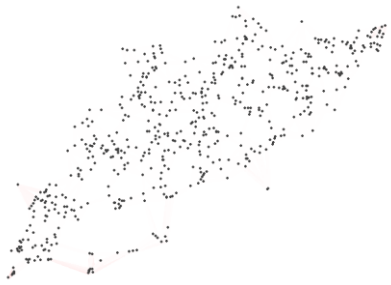

	

Figure 23 Projection avec pénalisation (λ = 0.1) sur le dataset SVHN au time step 14/29	*Figure 24 Reliability map sur une projection avec pénalisation (λ = 0.1) sur le dataset SVHN au time step 14/29*
	
Figure 25 Projection sans pénalisation sur le dataset Covid au time step 20/54	*Figure 26 Reliability map sur une projection sans pénalisation sur le dataset Covid au time step 20/54*
	
Figure 27 Projection avec pénalisation (λ = 0.1) sur le dataset Covid au time step 20/5	*Figure 28 Reliability map sur une projection avec pénalisation (λ = 0.1) sur le dataset Covid au time step 20/54*

Il est possible, pour chaque instance d'une projection, d'évaluer la fiabilité de leur position par rapport à leurs voisines. Si une instance x en basse dimension est voisine avec les instances a, b et c, mais, qu'en haute dimension, elle est voisine avec les instances a, b et d, alors, une incohérence est détectée par la reliability map et un vecteur coloré est tiré entre l'instance x et l'instance c. La couleur

de ce vecteur d'indication est relative à l'échelle d'erreur qui a été faite lors de la projection. Plus le vecteur est rouge, plus l'erreur est à prendre en compte. Inversement, plus une instance est bien positionnée par rapport à ses voisines, plus le vecteur sera blanc, et donc, moins il se verra. La reliability map établit la rougeur de ce vecteur, pour toutes les paires de voisins des points projetés. Pour des raisons de performance, ce calcul se réalise avec les 10 voisins les plus proches de chaque point. Voici des exemples de reliability map appliqués sur différents types de datasets. Pour étudier la fiabilité concernant plus de dix voisins d'une instance, il est plus intéressant de se tourner vers la métrique du trustworthiness qui sera plus efficace et pourra aller plus en profondeur. L'objectif final à atteindre lors de la réalisation d'une projection est d'observer une reliability map neutre en couleur, informant directement l'utilisateur que la projection est fiable et que des interprétations peuvent se faire sans problème.

4.2.3 AUClogRNX

AUClogRNX [13] est une métrique d'évaluation de préservation du voisinage. La différence entre AUClogRNX et le score de trustworthiness vient de la manière dont il est calculé. Pour chaque instance dans un dataset composé de *N* instance, l'intersection est calculée entre les k-voisins des instances en haute dimension et les k-voisins des instances en basse dimension, puis une division s'effectue par le produit entre le nombre d'instances du dataset et le nombre de voisins K. Elle permet d'obtenir un pourcentage correspondant au taux de similarité entre les voisins en haute et en basse dimensions. Cette équation se définit comme $Q_{\mathrm{NX}}(K) = \sum_{i=1}^{N} \left| v_i^K \cap n_i^K \right| / (KN)$. L'objectif ici est de calculer le $Q_{\mathrm{NX}}(K)$ pour tous les K de 1 jusqu'à *N*-2. Ainsi, ce procédé va calculer le taux de similarité entre les voisins en haute et basse dimension des instances d'un nombre incrémental de voisins jusqu'à considérer N-2 instances comme voisines. Une étape supplémentaire transforme $Q_{\mathrm{NX}}(K)$ en $R_{\mathrm{NX}}(K)$ via un procédé de rescaling pour donner plus d'importance aux valeurs $Q_{\mathrm{NX}}(K)$ utilisant un K plus faible et diminuant progressivement la valeur calculée des K plus élevés, car il est moins important de commettre une erreur de projection entre deux instances très éloignées que deux instances étant très proches. Ainsi, $R_{\mathrm{NX}}(K) = \frac{(N-1)*Q_{\mathrm{NX}}(K)-K}{N-1-K}$ est l'équation obtenue. Quand toutes les valeurs $R_{\mathrm{NX}}(K)$ sont calculées, elles définissent un graphique pouvant représenter ces valeurs et la fonction définie comme $\mathrm{AUC}_{\log K}(\mathrm{RNX(K)}) = (\sum_{K=1}^{N-2} R_{NX}(K)/K)/(\sum_{K=1}^{N-2} 1/K)$ calcule ensuite l'aire sous ce graphique. La valeur de l'aire sous la courbe des $R_{\mathrm{NX}}(K)$ représente le score final appelé AUClogRNX.

Cette métrique constitue la base d'une nouvelle métrique créée lors des expériences et propre aux techniques TDP, le temporal AUClogRNX réalisant un graphique des scores de similarité des voisinages par rapport à un time step de référence décrit dans la section 4.3.3.

4.3 Contribution à la création de nouvelles métriques

Une partie de la contribution de ce mémoire dans le domaine des techniques de TDP est la création de trois nouvelles métriques les évaluant et leur étant propre.

4.3.1 Conservation de la dynamique

La première métrique est graphique et mesure la conservation de la dynamique des instances. La dynamique est définie par la norme moyenne des vecteurs de vitesse et des vecteurs d'accélération de chaque instance à chaque time step. Il est à noter que ces normes de vecteurs n'ont pas d'unité.

Bien que rudimentaire, cette métrique permet de distinguer rapidement les effets des variations de paramètres (λ) sur la dynamique des instances. De plus, ce même graphique peut être réalisé sur les instances non projetées afin d'obtenir un élément de comparaison pour déterminer quelle solution représente au mieux la courbe de la dynamique réelle des données. L'objectif à atteindre avec les différentes solutions décrites dans ce document est la correspondance entre la dynamique des instances projetées et la dynamique en haute dimension.

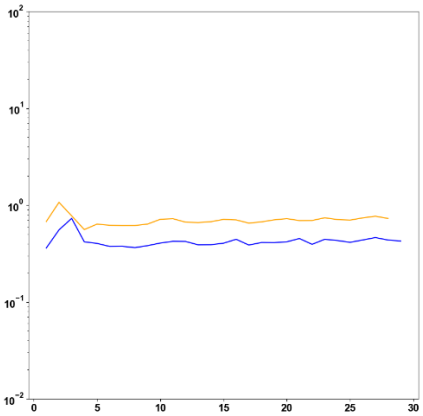


Figure 29 Exemple de graphique représentant la dynamique des instances du dataset SVHN en haute dimension

Les notions de distances, de dynamique ou encore de voisinage sont très différentes en haute dimension. Ce phénomène s'appelle la « *curse of dimensionality* ». Il est donc nécessaire de se pencher uniquement sur la forme des courbes obtenues. En abscisse, les valeurs numériques représentent le time step et n'ont donc pas d'unité générale car à chaque dataset correspondent sa temporalité et ses tranches temporelles relatives.

4.3.2 Conservation de la direction

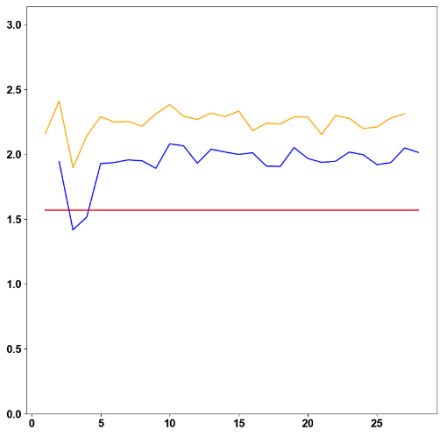


Figure 30 Exemple de graphique des changements de direction des instances du dataset SVHN en haute dimension

Le deuxième outil consiste à déterminer si la direction des données est globalement respectée dans la projection. L'intérêt ici est de déterminer si l'outil ne fournit pas de fausses informations concernant le changement de direction des instances. Le graphique représente l'évolution des changements de direction moyens des vecteurs de vitesse et des vecteurs d'accélération dans le temps

de chaque instance. La barre rouge correspond à un changement de direction de 90°, peu importe la direction. Toute valeur supérieure à cette barre représente un angle supérieur à 90° jusqu'à la limite de 180° représentant un demi-tour.

4.3.3 Temporal AUClogRNX

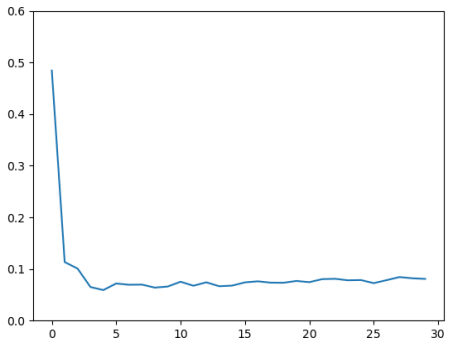


Figure 31 Exemple de temporal AUClogRNX ayant un time step de référence sur le premier time step de la temporalité

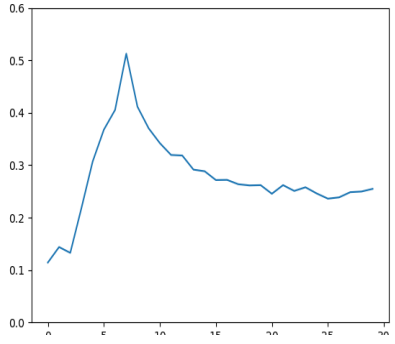


Figure 32 Exemple de temporal AUClogRNX ayant un time step de référence sur le time step au quart de la temporalité

En adaptant la métrique AUCLogRNX pour convenir à une TDP, une troisième métrique apparaît. Le temporal AUClogRNX mesure la cohérence temporelle des voisinages d'un dataset tout au long de sa temporalité. Le temporal AUClogRNX va comparer tous les voisinages de chaque instance d'un time step de référence avec les voisinages respectifs des autres time steps et en déduire la ressemblance. Sur la Figure 31 et la Figure 32, deux exemples montrent, sous deux time steps de référence différents, l'évolution des similarités de voisinage dans le temps d'un même dataset. Pour la Figure 31, le pic indique soit l'endroit où la similarité est la plus grande, soit l'endoit où le time step de référence compare ses voisins aux voisins de ses propres données projetées. Les deux figures sont issues du même dataset et indiquent que le début du dataset est fortement différent d'un point de vue des voisinages par rapport au reste du dataset. En revanche, quand le time step de référence se pose plus loin dans la temporalité (Figure 32), une similarité de « base » apparaît à la suite du time step de référence indiquant un début de stabilité dans les voisinages. Le temporal AUClogRNX permet d'analyser avec précision l'évolution des voisinages le long de la temporalité.

4.4 Outil de visualisation

Des outils de visualisation de données comme matplotlib existent et sont pratiques à utiliser lorsqu'il s'agit de visualiser des données projetées. Cependant, l'animation de données est plus complexe et peu permissive avec cette librairie. Certaines librairies existent et permettent facilement de dessiner des formes à des coordonnées et de les faire bouger facilement tout en donnant un grand pouvoir d'action à l'utilisateur sur l'application… les moteurs de jeu. La librairie "pygame" est la base de l'outil de visualisation développé contribuant à la visualisation des projections du dynamic t-SNE ou toutes autres techniques TDP.

Dans les faits, un moteur de jeu a pour objectif initial de créer un jeu vidéo. Dans ce sens, de nombreux outils comme la gestion d'évènements provenant de la souris ou du clavier apporte un grand apport ergonomique fonctionnel à l'outil de visualisation. De plus, la partie graphique est, elle

aussi, optimisée pour effectuer des centaines d'images par seconde. Un outil de visualisation fluide, interactive et facile à implémenter fournit un environnement idéal pour une visualisation qualitative.

L'outil qui a été développé prend le résultat du dynamic t-SNE ou de tout autre solution fournissant des projection bi-dimensionnelles. La résolution de l'application, la vitesse de défilement d'image et la précision sont les trois paramètres de configuration principaux de l'outil. La précision est une valeur d'extrapolation inter time step, par exemple : avec une valeur de précision de 10, 10 étapes intermédiaires vont être générée entre chaque time step bougeant de manière linéaire les points de leur position t à la position t+1. Ce paramètre permet d'obtenir une visualisation très fluide aidant parfois fortement à comprendre le mouvement des données, mais aucune interprétation ne doit être réalisée sur ces time steps interpolés car les positions calculées des instances sont complètement fictives.

Comme dit précédemment, un outil de visualisation basé sur un moteur de jeu est très pratique et, de ce fait, beaucoup de fonctionnalités supplémentaires ont été ajoutées à l'outil.

- La capacité de visualiser les vecteurs de mouvement des instances afin d'en dégager un champ vectoriel informant l'utilisateur de la direction de ces dernières au fil des times-steps ;

Figure 33 Visualisation d'une projection d'instances du dataset du Covid

Figure 34 Visualisation du champ vectoriel des vitesses de la projection en Figure 33

- La capacité de visualiser les vecteurs d'accélération des points ;

Figure 35 Visualisation d'une projection d'instances du dataset du Covid

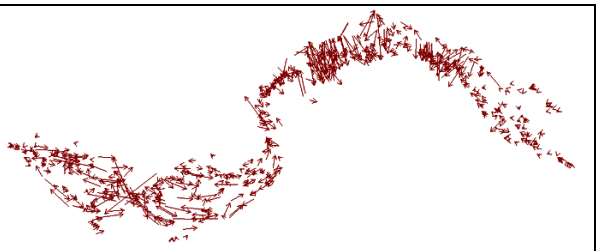


Figure 36 Visualisation du champ vectoriel des accélérations de la projection en Figure 35

- La possibilité de mettre en pause, d'avancer et de reculer image par image ;
- Identifier chaque instance. Par exemple : pour le dataset du covid, chaque instance correspond à une commune ;
- Afficher la reliability map à un time step précis ;
- Afficher pour chaque time step le score de Trustworthiness de la projection observée.

Avec un tel outil, il est très facile de naviguer parmi les time steps et d'exploiter les données sans problème. L'outil ne prend en moyenne pas plus de quelques secondes pour se préparer et est fluide.

5 Expériences sur dynamic t-SNE

Les expériences effectuées sur dynamic t-SNE présentent la démarche suivante : prendre un dataset ; préparer ses données ; lancer le dynamic t-SNE sur les données et envoyer le résultat dans l'outil de visualisation. Ces étapes permettent en premier lieu de paramétrer le dynamic t-SNE, c'est-à-dire de trouver le λ et la perplexité adéquate pour la meilleure visualisation possible. Après une phase de paramétrage, les véritables entraînements peuvent s'effectuer. C'est ainsi qu'avec le dataset SVHN, trois dynamic t-SNE sont effectués, un avec un λ nul, un second avec un λ à 0.1 (valeur par défaut recommandée dans le papier décrivant le dynamic t-SNE), et un dernier avec un λ plus élevé que la valeur indiquée dans l'état de l'art, 0.15.

5.1 Visualisation de la solution

Ci-dessous sont représentés les tableaux de figures reprenant les 12 premières images des animations de la solution sur le dataset du Covid et du SVHN.

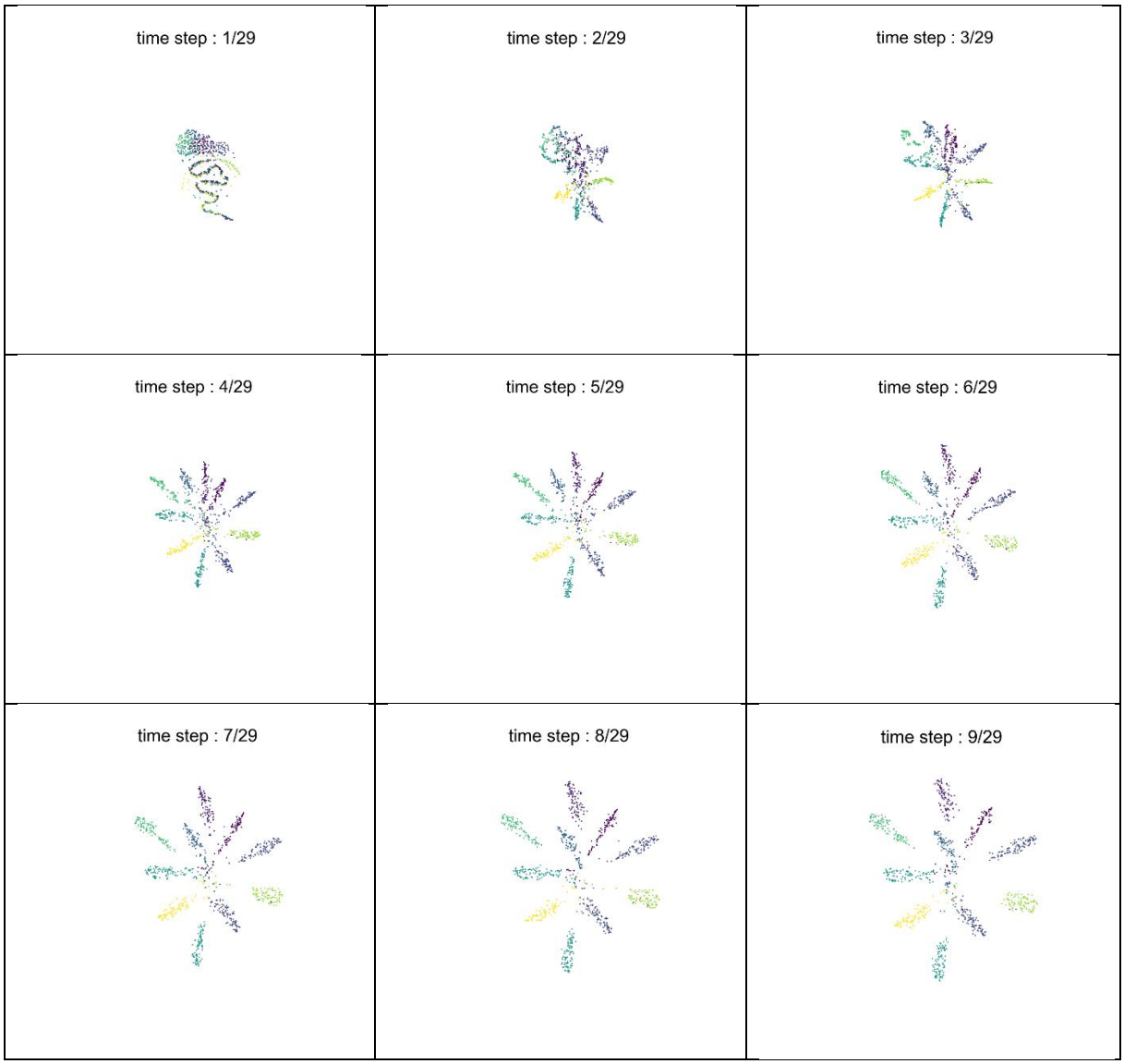
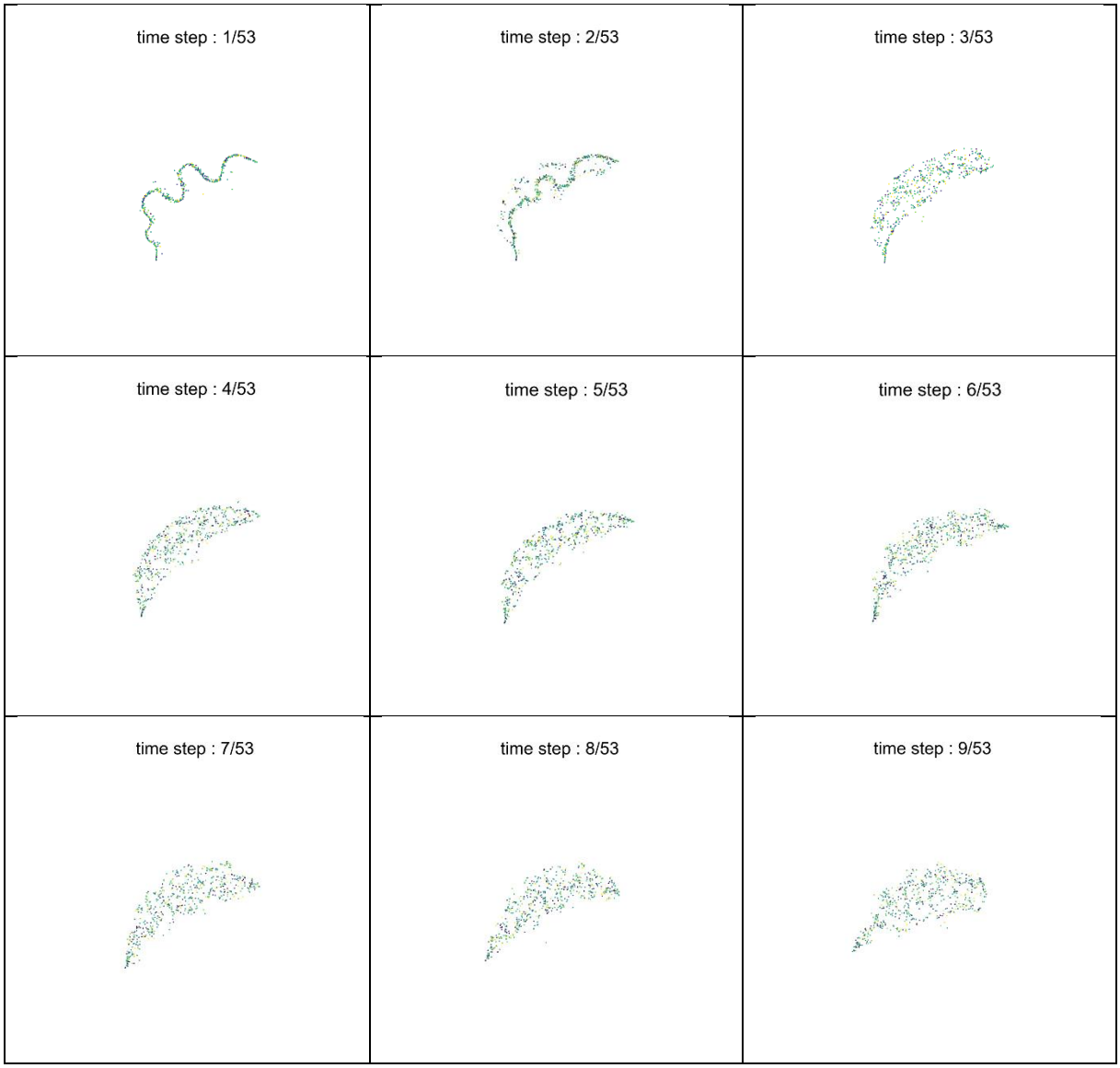


Figure 37 Animation d'une succession de projections générées par le dynamic t-SNE sur le dataset SVHN



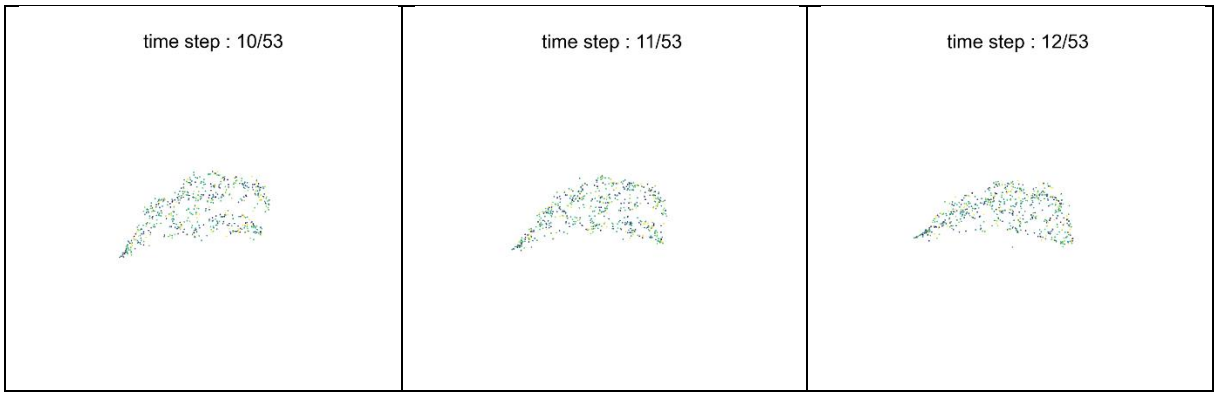


Figure 38 Animation d'une succession de projections générées par le dynamic t-SNE sur le dataset du Covid

La comparaison avec l'animation de la section 2.2 saute aux yeux, une cohérence globale et une fluidité de mouvement justifient largement l'existence de cette solution. Mais est-elle si performante ? C'est une réponse à cette question que les sections suivantes vont fournir.

5.2 Résultats des métriques

Après la configuration, l'exécution des entraînements, c'est au tour des métriques d'être récoltées. Les différentes métriques décrites aux chapitre 4.2 et 4.3 sont prélevées sur les projections. Les prochaines sections décrivent les performances du dynamic t-SNE pour chacune des métriques.

5.2.1 Conservation de la dynamique

Les résultats de la métrique de la conservation de la dynamique sont décrits dans cette section. Les deux tableaux qui suivent sont composés de quatre graphiques représentant :

- La dynamique des instances en haute dimension, ce graphique est le point de comparaison vers lequel il est intéressant de tendre avec la projection (en haut à gauche) ;
- La dynamique des instances projetées au travers d'un dynamic t-SNE non paramétré, son λ est nul (en haut à droite) ;
- La dynamique des instances projetées au travers d'un dynamic t-SNE paramétré avec le λ conseillé dans l'état de l'art, 0.1 (en bas à gauche);
- La dynamique des instances projetées au travers d'un dynamic t-SNE paramétré avec un λ supérieur à celui annoncé dans l'état de l'art. Cette expérience a pour but de déterminer par l'exagération l'effet de la pénalisation sur la projection. (en bas à droite)

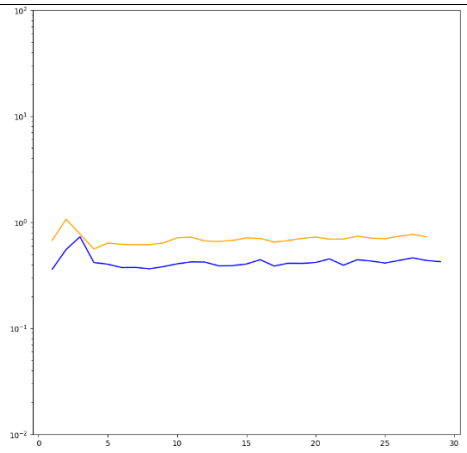


Figure 39 Dynamique en Haute dimension des instances du dataset SVHN

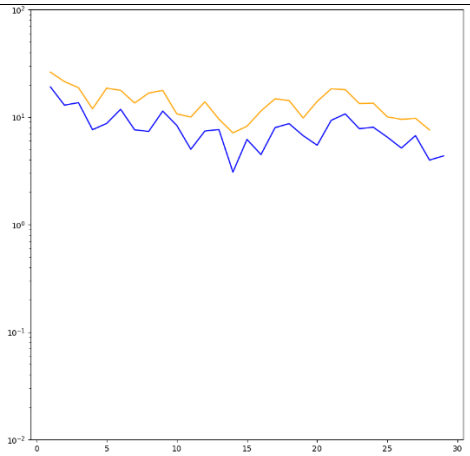


Figure 40 Dynamique des instances du TDP sans pénalisation sur le dataset SVHN

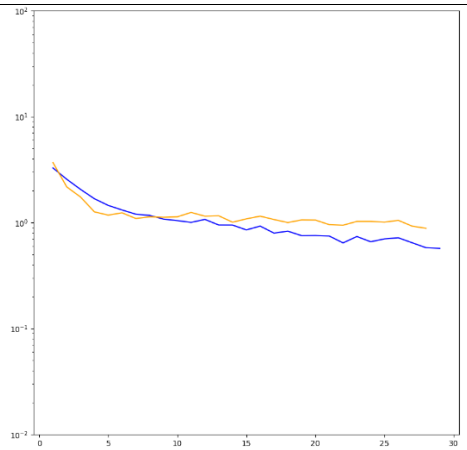


Figure 41 Dynamique des instances du TDP avec pénalisation (λ = 0.1) sur le dataset SVHN

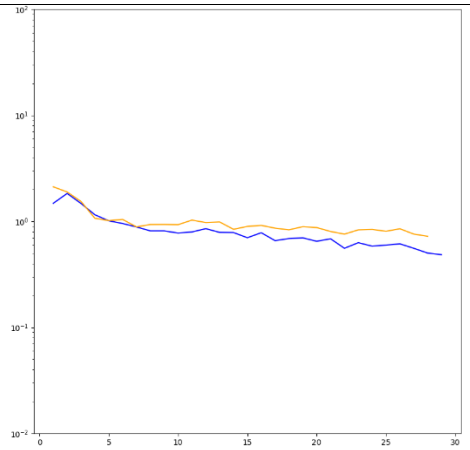


Figure 42 Dynamique des instances du TDP avec pénalisation (λ = 0.15) sur le dataset SVHN

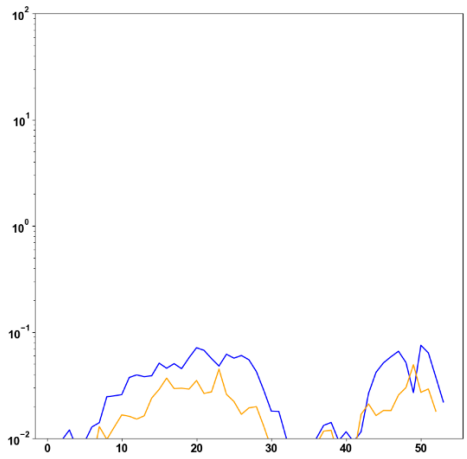


Figure 43 Dynamique en Haute dimension des instances du dataset Covid

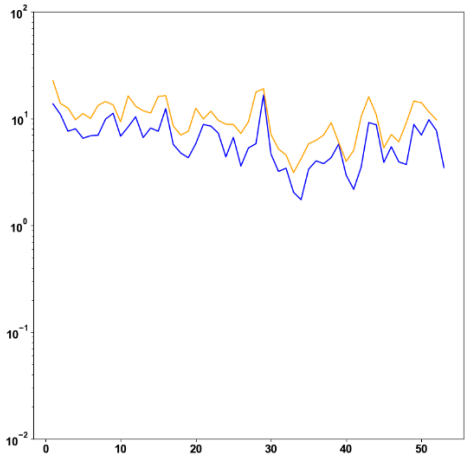


Figure 44 Dynamique des instances du TDP sans pénalisation sur le dataset Covid

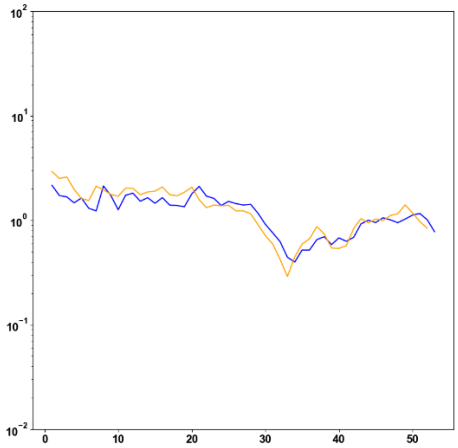


Figure 45 Dynamique des instances du TDP avec pénalisation (λ = 0.1) sur le dataset Covid

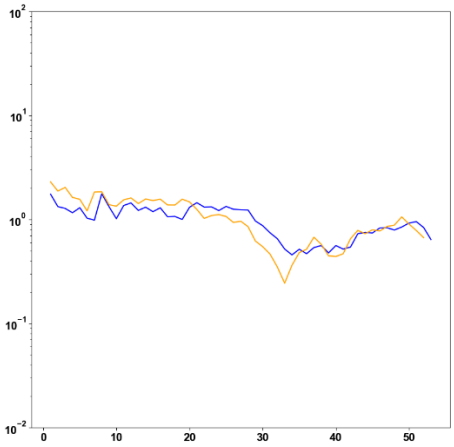


Figure 46 Dynamique des instances du TDP avec pénalisation (λ = 0.15) sur le dataset Covid

Il est à noter que les features présentes dans le dataset du Covid sont des valeurs assez faibles entre 0 et 1. La différence de déplacement entre deux time steps se fait également très légèrement ce qui induit un graphique de la dynamique aux valeurs très faibles expliquant la présence de la courbe aussi basse sur la Figure 43.

Les résultats montrent un fort impact du paramètre λ observé sur la dynamique générale des données. Comme prévu, pénaliser les vecteurs de mouvement (en bleu sur de la Figure 39 à la Figure 46) lors de l'entraînement du dynamic t-SNE induit une forte diminution de leur norme. Il est également observé que les vecteurs d'accélération (en orange sur les Figures Figure 39 à Figure 46), sont également impactés. La pénalisation des mouvements et le paramétrage du λ ont une influence non négligeable sur la norme moyenne finale des instances.

Lors des expériences, il a été observé qu'une valeur de λ supérieure à l'état de l'art ne changeait plus beaucoup la conservation de la dynamique. La forme de la courbe reste globalement la même mais une légère diminution de chaque valeur de time step est observée.

5.2.2 Les changements de direction

Ci-dessous sont présentés les graphiques pour la métrique sur les changements de direction.

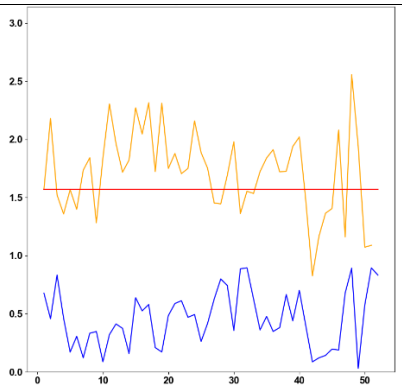


Figure 47 Changements de direction des instances en haute dimension du dataset COVID

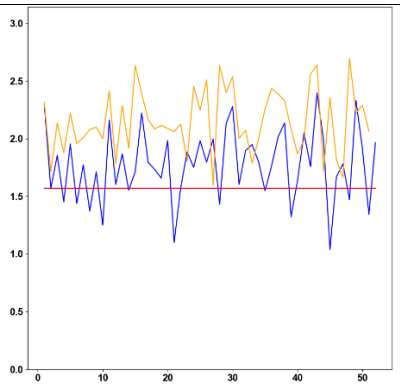


Figure 48 Changements de direction des instances du TDP sans pénalisation sur le dataset du COVID

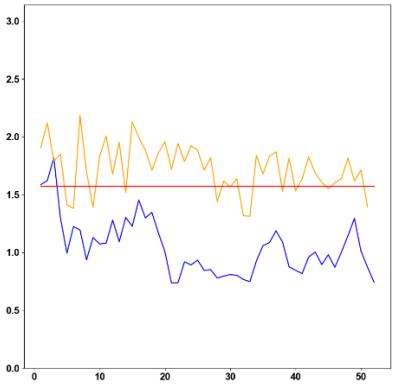


Figure 49 Changements de direction des instances du TDP avec pénalisation (λ = 0.1) sur le dataset du COVID

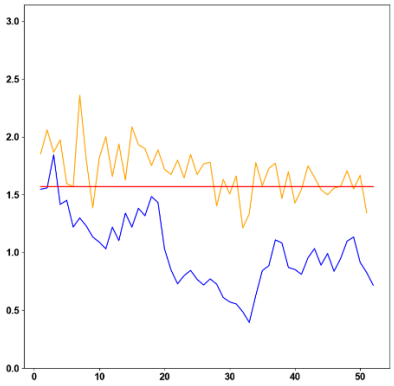


Figure 50 Changements de direction des instances du TDP avec pénalisation (λ = 0.15) sur le dataset du COVID

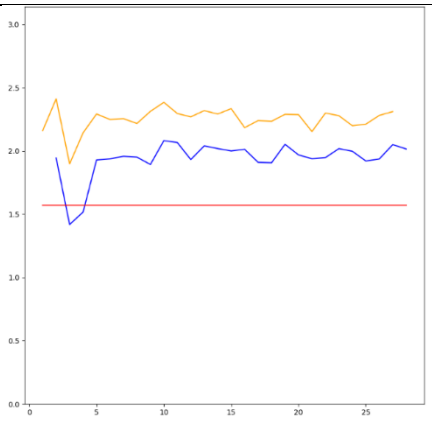


Figure 51 Changements de direction des instances en Haute dimension du dataset SVHN

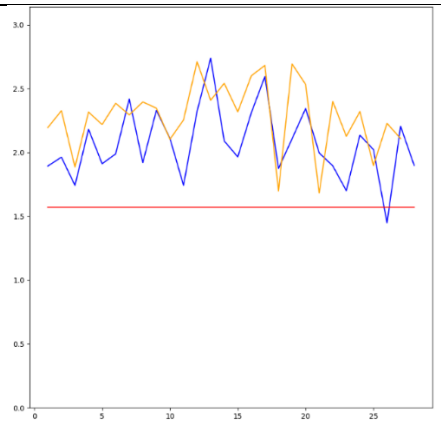


Figure 52 Changements de direction des instances du TDP sans pénalisation sur le dataset du SVHN

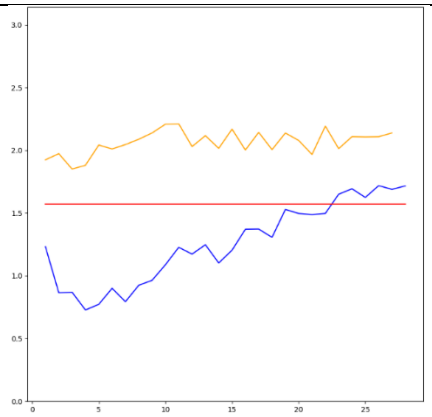


Figure 53 Changements de direction des instances du TDP avec pénalisation (λ = 0.1) sur le dataset du SVHN

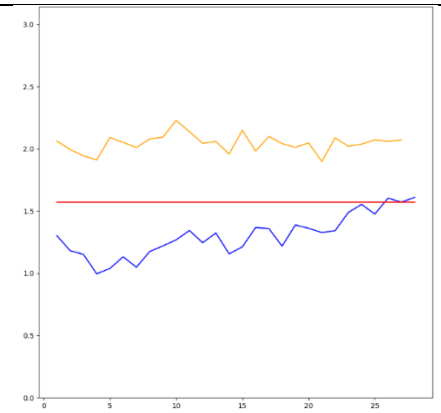


Figure 54 Changements de direction des instances du TDP avec pénalisation (λ = 0.15) sur le dataset du COVID

Il apparaît pour les expériences sur le dataset du COVID et du SVHN que la pénalisation des mouvements des instances induit bien une réduction de leur changement de direction. Il est également observable que plus le λ croît, plus les points se meuvent de façon fluide. Cependant, la présence d'un facteur de pénalisation de vitesse ne semble pas altérer les vecteurs d'accélération. Quand aucune pénalisation n'est appliquée, il est logique d'observer un mouvement plus imprévisible des instances.

5.2.3 Reliability map

Les résultats concernant la reliability map sont plus complexes à évaluer quand ils doivent être comparés entre eux. Cependant, une idée générale peut se faire par un jugement subjectif de la situation.

Figure 55 Reliability map des instances du TDP sans pénalisation sur le dataset du Covid

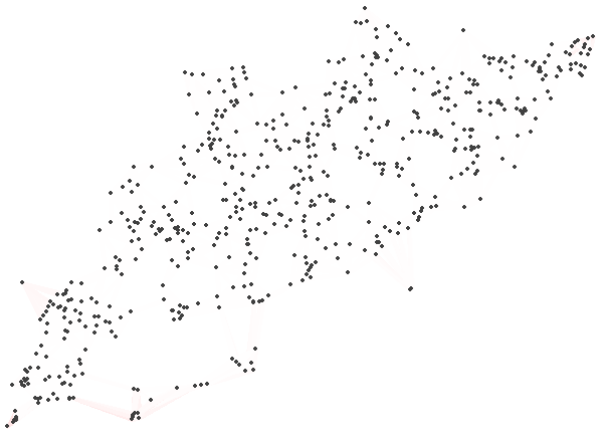


Figure 56 Reliability map des instances du TDP avec pénalisation (λ = 0.1) sur le dataset du Covid

Figure 57 Reliability map des instances du TDP avec pénalisation (λ = 0.15) sur le dataset du Covid

Dans l'ensemble, la Figure 55, la Figure 56 et la Figure 57 montrent des reliability maps performantes dans leur ensemble. Cependant, il est observable que quelques instances semblent s'être perdues dans les projections pénalisées d'un λ.

5.2.4 Trustworthisness

Les tableaux ci-dessous représentent un échantillon des valeurs de trustworthiness calculé avec le dynamic t-SNE sur les datasets du COVID et de SVHN avec pour valeurs de λ = 0 et 0.1. La moyenne des trustworthiness est calculée à partir de tous les time steps de la temporalité. La valeur T représente le nombre de time steps dans la temporalité, T/2 représente donc le time step se trouvant à la moitié de la temporalité.

	Dataset SVHN	Dataset Covid
λ = 0		
t = 0	0.9236	0.9901
t = T/4	0.7725	0.9662
t = T/2	0.8053	0.9762
t = 3T/4	0.8219	0.9809
t = T	0.7915	0.9766

Moyenne sur tous les time steps :	0.8033	0.9713

	Dataset SVHN	Dataset Covid
λ = 0.1		
t = 0	0.8729	0.9847
t = T/4	0.7974	0.9786
t = T/2	0.7833	0.9747
t = 3T/4	0.7800	0.9827
t = T	0.7835	0.9803
Moyenne sur tous les time steps :	0.7973	0.9752

Une forte différence de fiabilité est constatée entre la projection de SVHN et du Covid. Une si grande variation de fiabilité s'explique par la différence de dimensions de leur espace original. Le dataset du Covid est originalement constitué de trois dimensions. Une réduction de trois dimensions vers deux est triviale comparée au SVHN. Ce dernier étant constitué de 128 dimensions, une valeur moyenne de fiabilité de 0.79 pour une telle réduction est justifiée et est même un plutôt bon score.

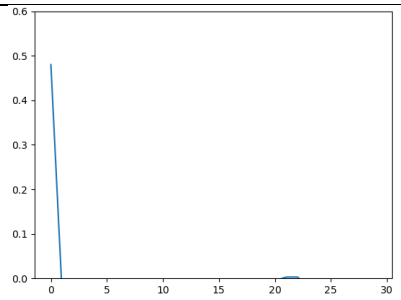
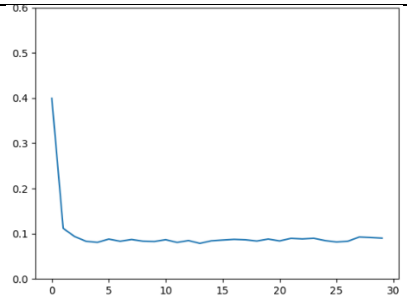
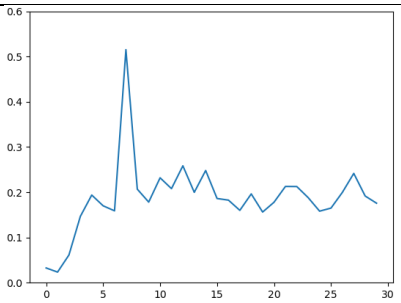
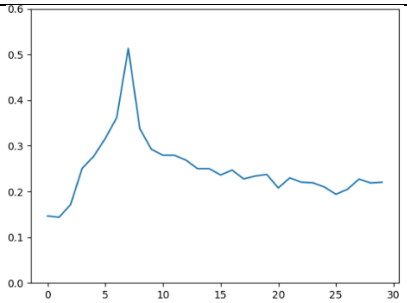
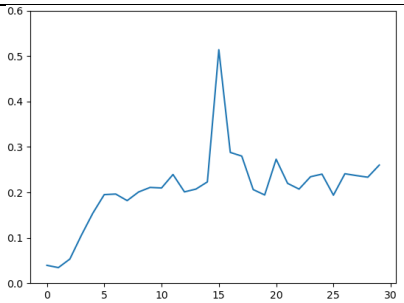
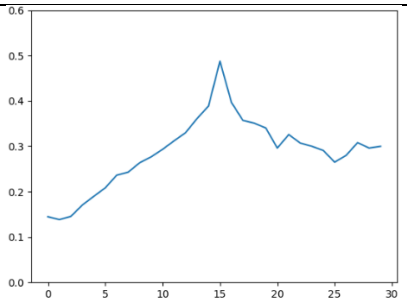
Entre une projection avec et sans pénalisation, une légère augmentation pour le dataset du covid et une légère augmentation de fiabilité pour le dataset du SVHN sont observées. Cette différence est cependant relativement faible et ignorable. De légères différences de fiabilité sont observées sur des projections utilisant le même lambda car il est possible que certaines configurations aléatoires des positions des points soient plus propices à fournir des projections finales plus fiables. Mais qu'en est-il des projections utilisant un lambda plus important que celui prescrit dans la littérature ?

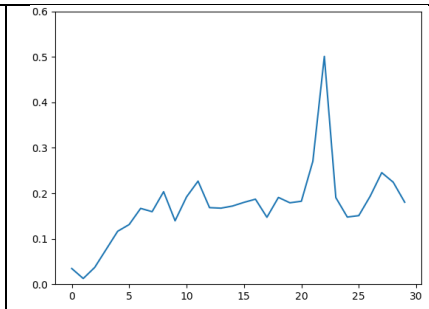
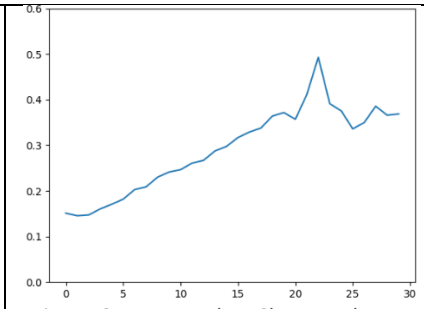
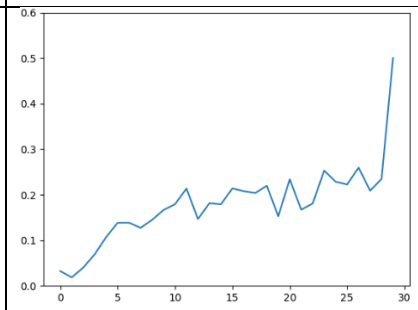
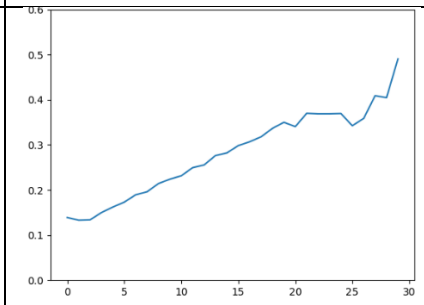
	Dataset SVHN	Dataset Covid
λ = 0.15		
t = 0	0.6493	0.9817
t = T/4	0.7852	0.9752
t = T/2	0.7962	0.9706
t = 3T/4	0.7846	0.9788
t = T	0.7823	0.9688
Moyenne sur tous les time steps :	0.7852	0.9647

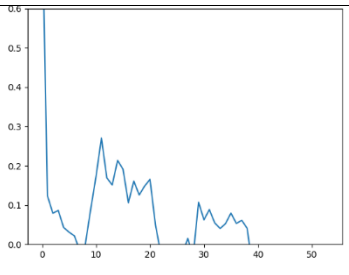
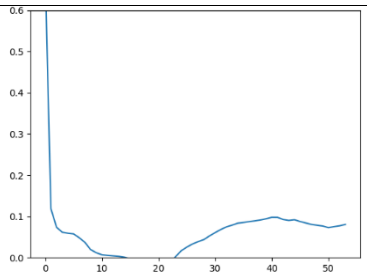
Cette fois, un impact plus important est observé sur les valeurs de trustworthiness et cet impact est plus que négligeable. Une diminution d'un centième est encore raisonnable mais, à choisir, il est plus intéressant de choisir une solution avec le λ recommandé. Il est également à noter qu'aucun des tests n'a permis d'obtenir un λ supérieur à un certain seuil propre à chaque dataset. Passé ce seuil, l'entraînement s'emballe et ne se termine pas car la fonction de coût diverge.

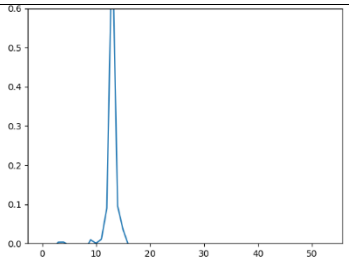
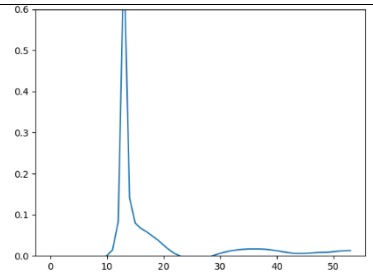
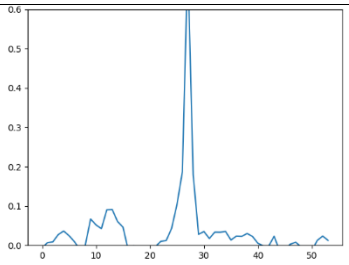
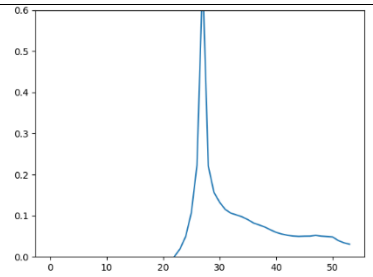
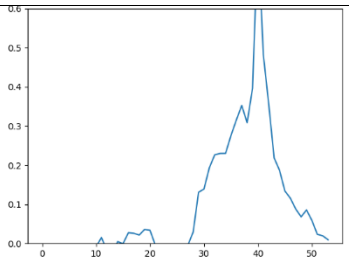
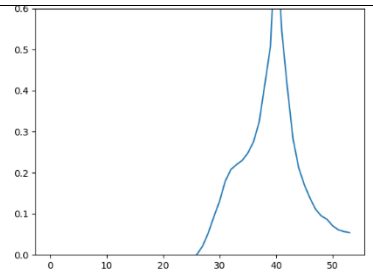
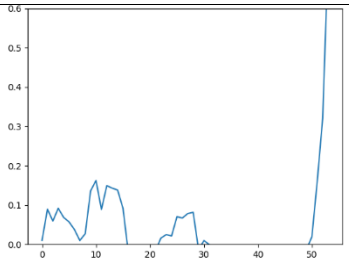
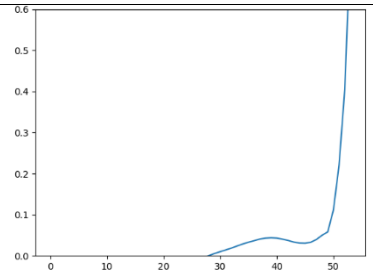
5.2.5 Temporal AUClogRNX

Deux interrogations sur les résultats du temporal AUClogRNX se sont posées. Comment se positionne l'état de l'art par rapport à cette métrique ? Est-ce possible d'améliorer ces résultats ? Voici les quelques observations qui ont été établies sur la métrique temporal AUClogRNX.

Dataset : SVHN Time step de référence	Dynamic t-SNE λ = 0	Dynamic t-SNE λ = 0.1
t=0	 *Figure 58 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = 0/T*	 *Figure 59 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = 0/T*
t=n/4	 *Figure 60 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = T/4*	 *Figure 61 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T/4*
t=n/2	 *Figure 62 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = T/2*	 *Figure 63 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T/2*

t=3n/4	 *Figure 64 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = 3*T/4*	 *Figure 65 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = 3*T/4*
t=n	 *Figure 66 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = T*	 *Figure 67 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T*

Dataset : Covid Time step de référence	Dynamic t-SNE λ = 0	Dynamic t-SNE λ = 0.1
t=0	 *Figure 68 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = 0/T*	 *Figure 69 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = 0/T*

t=n/4	 *Figure 70 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = T/4*	 *Figure 71 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T/4*
t=n/2	 *Figure 72 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = T/2*	 *Figure 73 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T/2*
t=3n/4	 *Figure 74 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = 3*T/4*	 *Figure 75 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = 3*T/4*
t=n	 *Figure 76 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = T*	 *Figure 77 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T*

Sur ces résultats, bien que les cohérences soient très faibles entre les time steps (observable par le pic de valeurs aux alentours du time step de référence), une tendance sur le dataset SVHN montre une amélioration assez marquée de la cohérence après pénalisation. Cependant, le dataset COVID fournit des graphiques à l'allure plus lissée par rapport à la solution sans pénalisation.

Dans les faits, SVHN est un dataset qui ressasse l'évolution de l'apprentissage d'un CNN. Donc, les données sont contraintes par le CNN à converger vers un état final représentant le choix qu'il va faire lors de sa prédiction (phénomène causé par l'apprentissage par descente de gradient du CNN). Une mise en ordre des données s'effectue dans le temps, ce qui explique la qualité des graphiques générés par le temporal AUClogRNX démontrant une conservation des voisins dans le temps très efficace. Pour le Covid, un phénomène similaire devrait apparaître, mais il est à noter que le dataset n'est pas complet. Pas complet dans le sens où les instances n'ont pas encore convergé vers leur état final (représentant une vaccination totale de la population). De ce fait, le phénomène présent sur le SVHN est attendu sur le dataset du covid mais probablement pas encore présent par manque de données.

5.3 Discussions des résultats

Après analyses des résultats sur les différents datasets, certaines conclusions ont été déclarées. Le dynamic t-SNE est une technique TDP qui projette les données de façon à favoriser certains paramètres au détriment d'autre.

Tout d'abord, la pénalisation de mouvement appliquée lors de l'entraînement aide fortement à la réalisation d'une animation finale interprétable lorsqu'un bon λ y est appliqué. Cette pénalisation force le dynamic t-SNE à s'adapter pour arranger les données malgré le fait que ces dernières puissent avoir une grande vélocité en haute dimension. Ce phénomène peut entraîner une désinformation quant à la vélocité réelle des données ; il est donc intéressant de trouver une alternative à ce défaut technique permettant de conserver la dynamique réelle des données.

Ensuite, la métrique sur le changement de direction informe de manière flagrante qu'une augmentation du λ entraîne une diminution directe des changements de directions des vecteurs de mouvements. Ce phénomène est la conséquence directe de la stabilité apportée à la visualisation discutée dans le paragraphe précédent. Avec un λ nul, les vecteurs de mouvement et d'accélération se meuvent chaotiquement comme prévu. Les vecteurs d'accélération, eux, ne semblent pas être affectés par la pénalisation. Une solution capable d'agir sur ces vecteurs est développée dans ce document au chapitre 6.

Toutes ces techniques apportées pour favoriser l'aspect visuel de la projection viennent avec un défaut non négligeable. Sur la reliability map, certaines zones rouges apparaissent dans les nuages de points indiquant que certaines instances sont étrangères à leurs voisins. Après réflexion, il est préférable d'accorder plus d'importance à l'aspect fiabilité plutôt qu'à l'aspect esthétique d'une visualisation. Dans les faits, si un dataset ayant des mouvements naturellement brusques est implémenté dans un dynamic t-SNE, il en ressortira une animation fortement fluidifiée par rapport à la réalité. Il est donc primordial de trouver une solution incluant cet aspect de respect du chaos pouvant exister dans la réalité.

Comme énoncé dans la section 5.2.4, lorsque le λ surpasse la valeur de 0.1, une légère « surcharge » de pénalisation provoque une faible chute de la fiabilité. Ce phénomène doit être évité car la fiabilité est la métrique la plus importante de toutes. Pour la solution de l'état de l'art, il suffit de limiter la valeur de λ à 0.1.

Les temporal AUClogRNX des deux datasets sont fortement différent malgré les similitudes du comportement des instances de chaque datasets. Le dataset du SVHN à fourni des courbes de temporal AUClogRNX aux profiles annonceur d'une bonne similitude de voisinage tout au long de la temporalité ce qui rappelle le principe de fonctionnement de la méthode d'apprentissage par descente de gradient.

D'autre solutions ont été envisagée pour résoudre un maximum de problème détecter lors de l'analyse des résultats des expériences sur dynamic t-SNE. La première solution consiste à retravailler la fonction de pénalisation du dynamic t-SNE pour tenter d'obtenir des résultats plus satisfaisants. La seconde consiste à retravailler la fonction objective au niveau de la formule de coût du t-SNE afin d'y insérer une notion de temporalité manquant dans le dynamic t-SNE. Ces solutions sont décrites dans de nouvelles implémentation décrites dans les chapitres 6, 7, 8 et 9.

6 Pénalisation des accélérations

La deuxième contribution de ce mémoire concernant les techniques de TDP réside dans la création de nouvelles solutions principalement basées sur le dynamic t-SNE. Quatre nouvelles solutions sont introduites dans les chapitres 6 à 9 ainsi que les expériences et discussions relatives à chacune d'entre elles.

Une première amélioration du dynamic t-SNE est la modification de la partie de son équation responsable de la pénalisation des vecteurs de vitesses. Comme spécifié dans l'analyse des résultats du dynamic t-SNE de l'état de l'art, la présence de cette pénalisation a une tendance prononcée à fluidifier les flux de déplacement, quitte à laisser tomber les mouvements potentiels rapides ou chaotiques que les données auraient dû avoir.

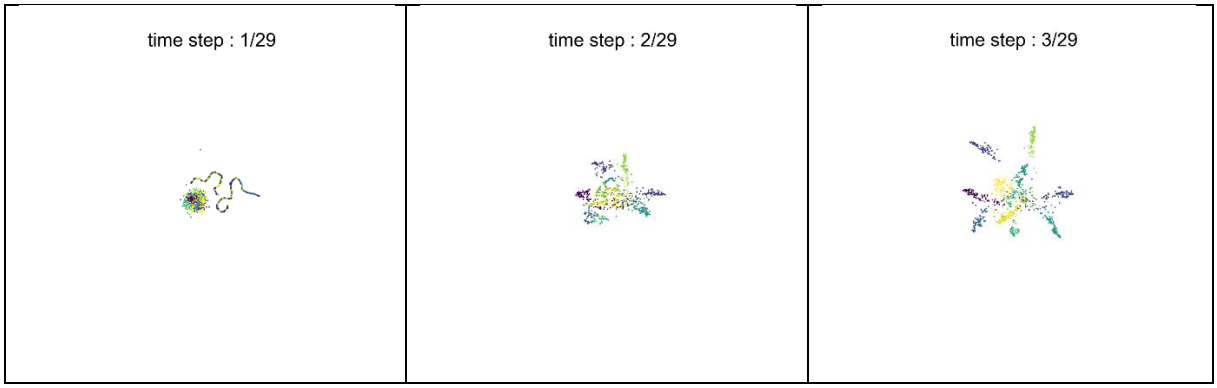
Notre proposition consiste à changer la cible de la pénalisation qui, cette fois, se tournera vers les vecteurs d'accélérations. $C = \sum_{t=1}^{T} C[t] + \frac{\lambda}{2N} \sum_{i=1}^{N} \sum_{t=1}^{T-1} ||p_i[t+1] - 2p_i[t] + p_i[t-1]||^2$ devient la fonction de coût de la solution.

La formule $||p_i[t+1] - 2p_i[t] + p_i[t-1]||^2$ est issue de la dérivée de la formule de la vitesse utilisée dans le dynamic t-SNE. Elle permet de calculer le carré de la norme du vecteur d'accélération du point i au time step t. L'idée, derrière cette implémentation, est d'essayer de maintenir la même force de pénalisation que la formule initialement utilisée, mais en laissant plus de possibilités aux données de se déplacer. En effet, les résultats attendus de cette modification se basent sur le principe que pénaliser l'accélération d'un point l'empêche de subir des accélérations trop importantes. Cela ne l'empêche pas de prendre énormément de vitesse si ce dernier en a besoin contrairement à la pénalisation des vecteurs de mouvement qui, eux, empêchent tout simplement la vitesse de croître.

Les mêmes tests sont effectués sur cette nouvelle implémentation et sur les mêmes datasets. Les résultats et observations de cette solution sont présentée dans les sections suivantes :

6.1 Visualisation de la solution

Ci-dessous sont représentés les tableaux de figures reprenant les 12 premières images des animations de la solution pénalisant les accélérations sur le dataset du Covid et du SVHN utilisant les meilleurs paramètres obtenus lors des expériences.



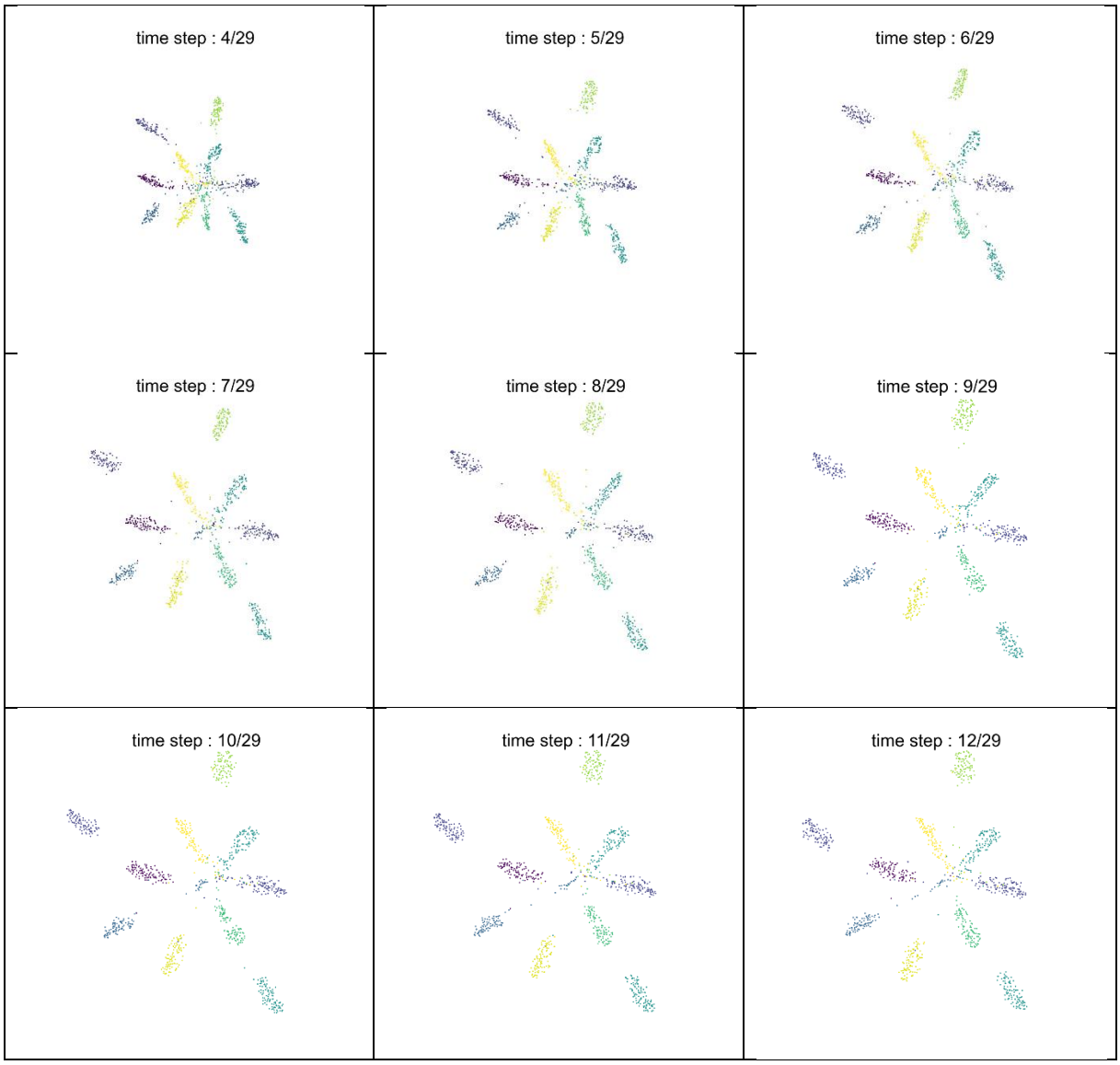
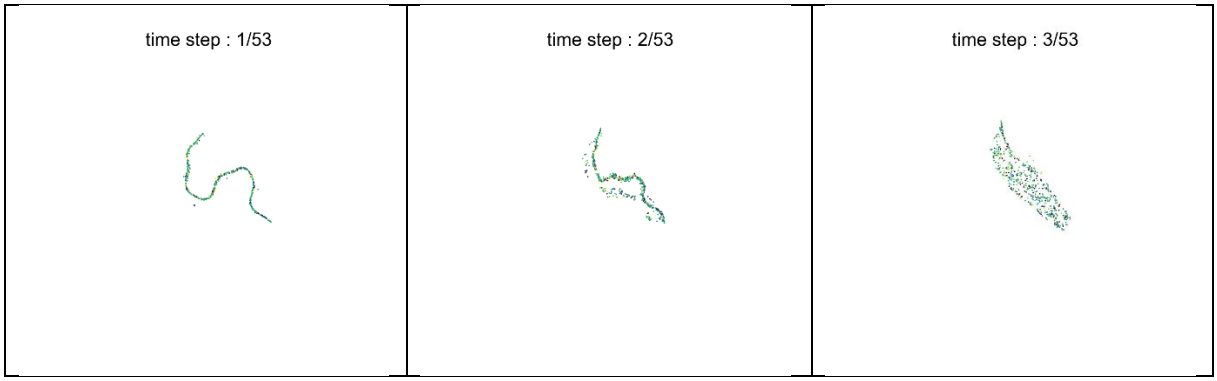


Figure 78 Animation d'une succession de projections générées par le dynamic t-SNE pénalisant les accélérations sur le dataset SVHN



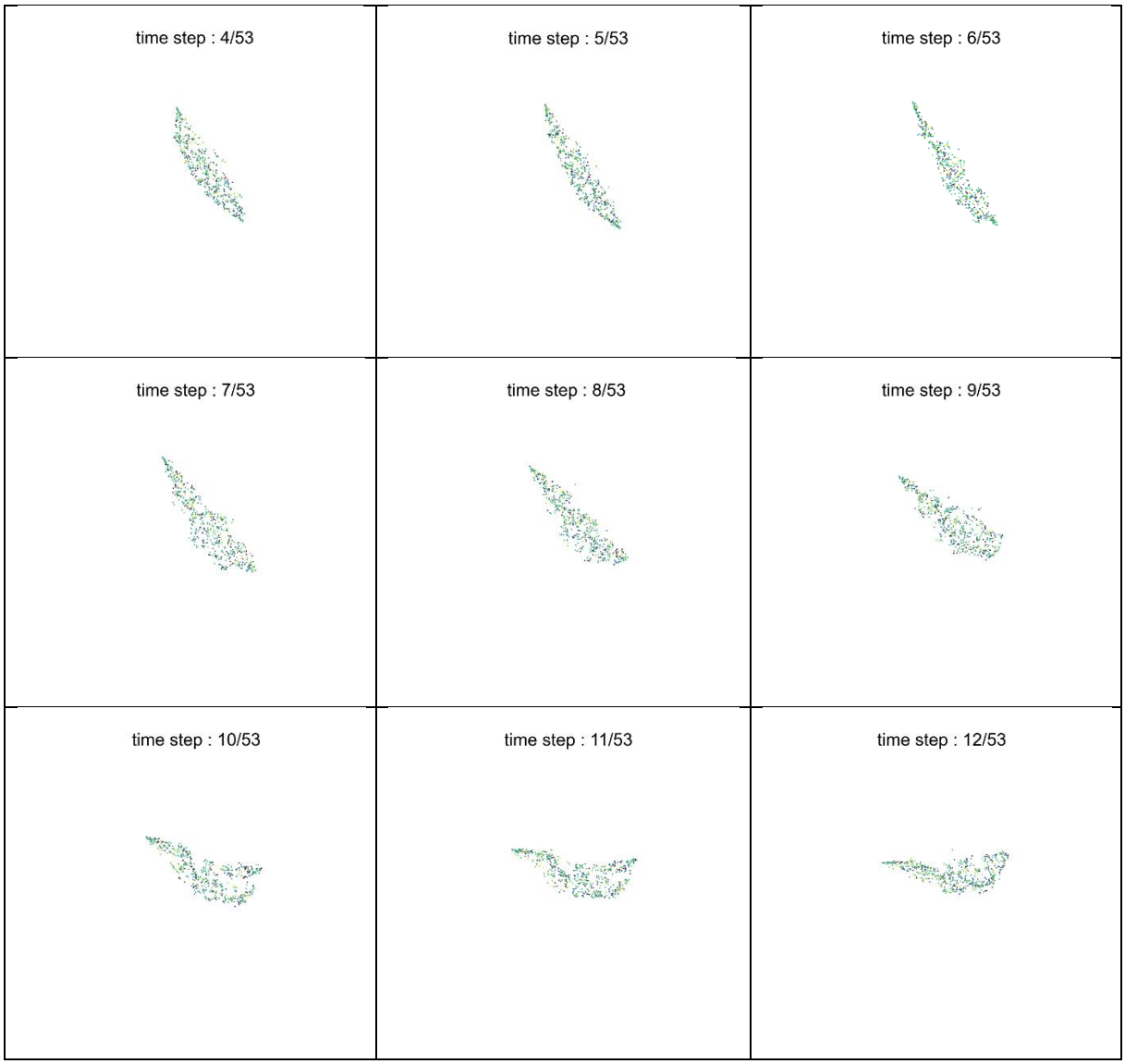


Figure 79 Animation d'une succession de projections générées par le dynamic t-SNE pénalisant les accélérations sur le dataset du Covid

6.2 Résultats des métriques

Un phénomène intéressant est apparu lors des expériences concernant la valeur du λ. Pour cette pénalisation, le λ doit être fortement réduit pour permettre à l'apprentissage de faire converger sa fonction de coût. En plus d'être inférieure, cette valeur change significativement selon le dataset, ce qui engendre un temps non négligeable de paramétrage pour chaque expérience à réaliser.

Il est à noter que les résultats pour un λ nul sont identiques aux résultats obtenus avec la pénalisation des vecteurs de vitesse car avec un λ nul, la fonction de coût des deux solutions est équivalente. Les résultats des métriques seront donc comparés entre la solution avec pénalisation des mouvements utilisant un λ = 0.1 et avec la pénalisation des accélérations utilisant le λ adapté.

6.2.1 Conservation de la dynamique

Pour cette solution, la dynamique des données est fortement modifiée par rapport à l'état de l'art. Comme prédit dans les hypothèses de changement de la formule de pénalisation, la dynamique est bien moins agressive sur les vitesses, leur permettant de s'accroître sans trop de limitations. Les accélérations, elles, sont fortement pénalisées, comme convenu. Pour rappel, il n'est pas nécessaire de comparer les valeurs des vecteurs entre les deux graphiques, mais plutôt la forme de la courbe qui en dit beaucoup sur le comportement des instances à chaque time step.

COVID : pénalisation des mouvements	COVID : pénalisation des accélérations
λ = 0.01 (faible pénalisation)	λ = 0.01 (faible pénalisation)

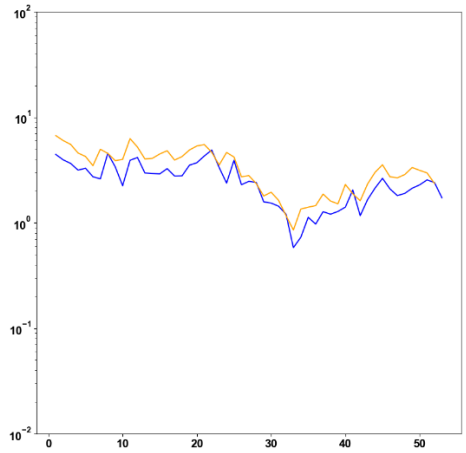
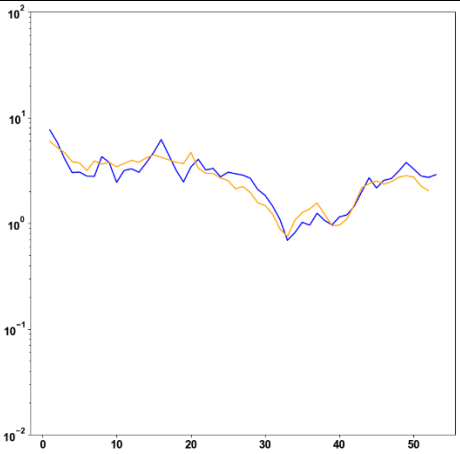

	

Figure 80 Dynamique des instances du TDP avec pénalisation des mouvements (λ = 0.01) sur le dataset Covid	*Figure 81 Dynamique des instances du TDP avec pénalisation des accélérations (λ = 0.01) sur le dataset Covid*
λ = 0.1 (Valeur par défaut)	λ = 0.045 (Pénalisation maximal avant divergence)

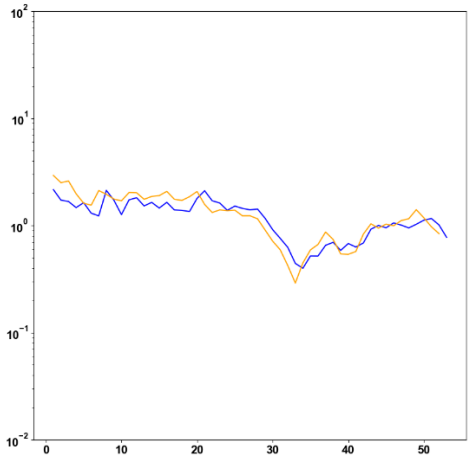
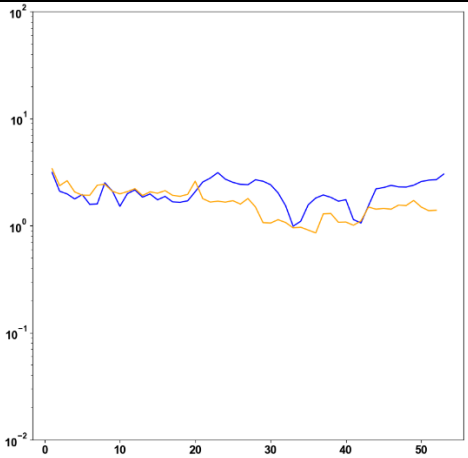

	

Figure 82 Dynamique des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid	*Figure 83 Dynamique des instances du TDP avec pénalisation des accélérations (λ = 0.04) sur le dataset Covid*

SVHN : pénalisation des mouvements	SVHN : pénalisation des accélérations
λ = 0.01 (faible pénalisation)	λ = 0.01 (faible pénalisation)

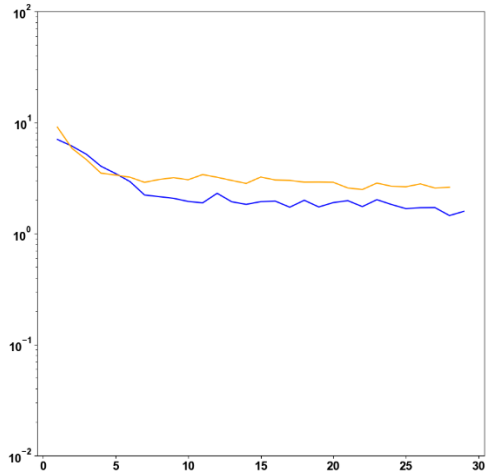
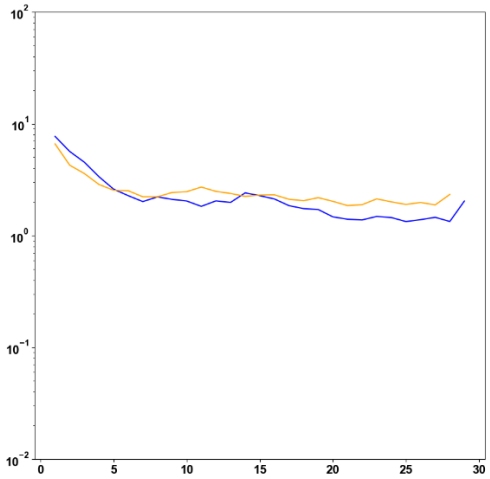

	

Figure 84 Dynamique des instances du TDP avec pénalisation des mouvements (λ = 0.01) sur le dataset SVHN	*Figure 85 Dynamique des instances du TDP avec pénalisation des accélérations (λ = 0.01) sur le dataset SVHN*
λ = 0.1 (Valeur par défaut)	λ = 0.07 (Pénalisation maximal avant divergence)

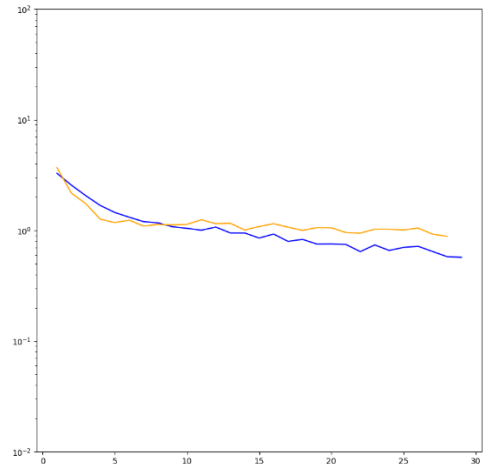
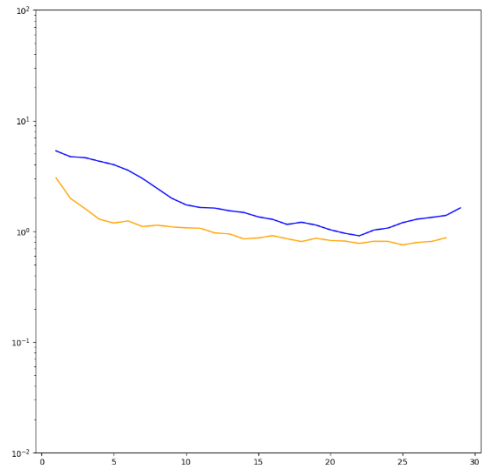

	

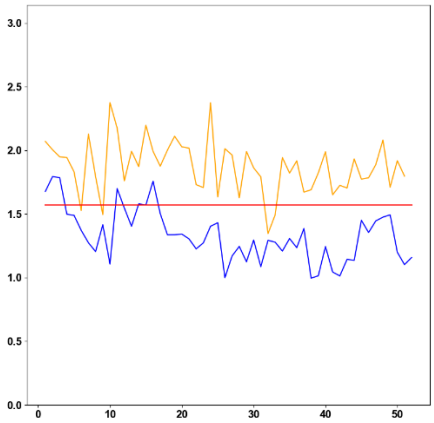
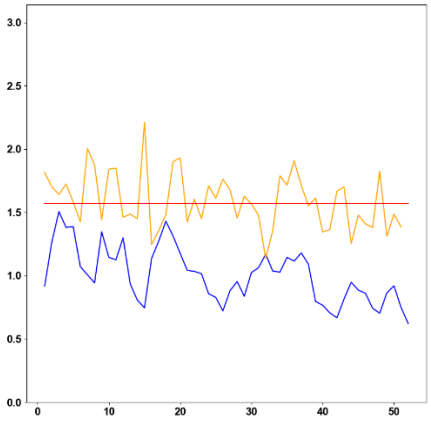
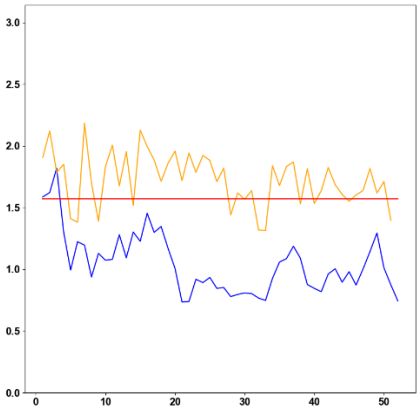
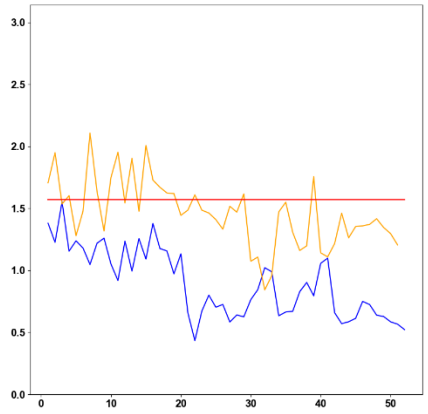
Figure 86 Dynamique des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN	*Figure 87 Dynamique des instances du TDP avec pénalisation des accélérations (λ = 0.07) sur le dataset SVHN*

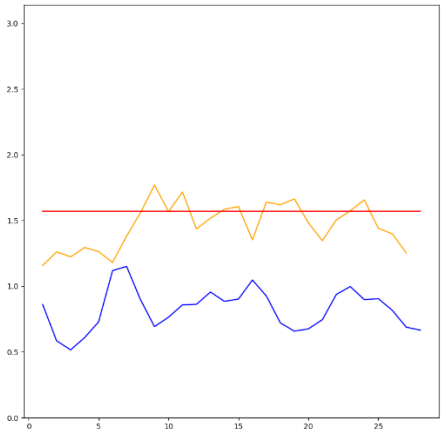
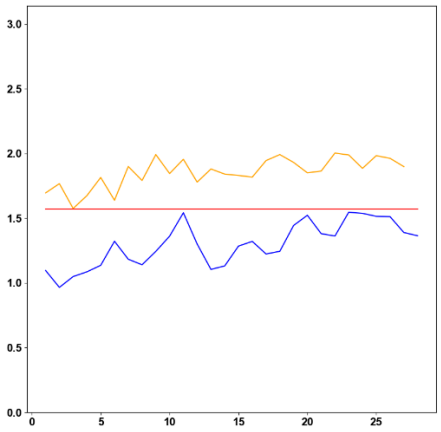
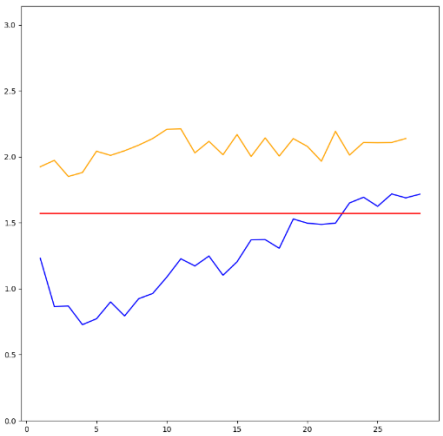
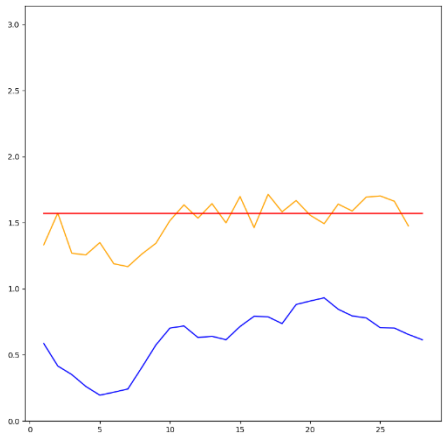
A faible pénalisation, les deux solutions fournissent une solution presque identique d'un point de vue de la conservation de la dynamique générale. Ce qui attendu car, à faible λ, les deux solutions se retrouvent avec une fonction de coût presque similaire. Mais lorsque le λ prend plus d'importance, les choses changent. Ce phénomène est plus visible pour le dataset du COVID que celui du SVHN car les instances de SVHN sont plus stables étant donné qu'elles convergent vers un état final sur plusieurs time steps.

De plus, limiter les accélérations permet à la dynamique d'évoluer de manière fluide fournissant des courbes plus douces et moins saccadées aux graphiques. Lors du visionnage des

projections, les données accélèrent et décélèrent bien plus lentement offrant des mouvements plus curvilignes aux données.

6.2.2 Conservation de la direction

COVID : pénalisation des mouvements	COVID : pénalisation des accélérations
λ = 0.01 (Faible pénalisation)	λ = 0.01 (Faible pénalisation)
 *Figure 88 Changements de direction des instances du TDP avec pénalisation des mouvements (λ = 0.01) sur le dataset Covid*	 *Figure 89 Changements de direction des instances du TDP avec pénalisation des accélérations (λ = 0.01) sur le dataset Covid*
λ = 0.1 (Spécifié comme valeur par défaut)	λ = 0.04 (Forte Pénalisation)
 *Figure 90 Changements de direction des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid*	 *Figure 91 Changements de direction des instances du TDP avec pénalisation des accélérations (λ = 0.04) sur le dataset Covid*

SVHN : pénalisation des mouvements	SVHN : pénalisation des accélérations
λ = 0.01 (Faible pénalisation)	λ = 0.01 (Faible pénalisation)
 *Figure Changements de direction des instances du TDP avec pénalisation des mouvements (λ = 0.01) sur le dataset SVHN*	 *Figure Changements de direction des instances du TDP avec pénalisation des accélérations (λ = 0.01) sur le dataset SVHN*
λ = 0.1 (Spécifié comme valeur par défaut)	λ = 0.07 (Forte Pénalisation)
 *Figure Changements de direction des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN*	 *Figure 92 Changements de direction des instances du TDP avec pénalisation des accélérations (λ = 0.07) sur le dataset SVHN*

Comme pour les mouvements, les directions sont également impactées par cette implémentation. Les instances subissant des accélérations plus faibles voient leur changement de direction devenir moins brusque ce qui est logique car la vitesse d'une instance au temps t+1 est déterminée par la vitesse de cette même instance au temps t plus un vecteur d'accélération réduit par la pénalisation. Cette accélération étant plus faible, elle impactera moins le vecteur de vitesse et aura plus de difficulté à le faire changer de direction. Plus le λ est grand, plus les directions et la dynamique semblent être affectées et semblent calmer la dynamique générale des instances, ce qui aide à fournir une animation fluide et plus stable. Ce comportement ne génère-t-il pas un inconvénient ? Celui de réduire la fiabilité des TDP produites ? Les deux prochaines métriques vont répondre à ces deux questions.

6.2.3 Reliability map

Les maps ci-dessous proviennent du même dataset et sont comparées selon différentes valeurs de λ. De ce fait, l'hypothèse selon laquelle le λ influence la fiabilité de la visualisation proportionnellement à sa taille pourra être vérifiée visuellement en tout cas.

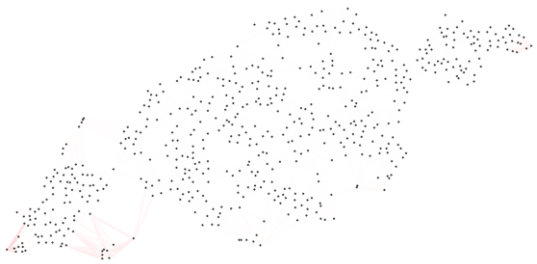


Figure 93 Reliability map des instances du TDP sans pénalisation sur le dataset du Covid

Figure 94 Reliability map des instances du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset du Covid

Figure 95 Reliability map des instances du TDP avec pénalisation des accélérations (λ = 0.04) sur le dataset du Covid

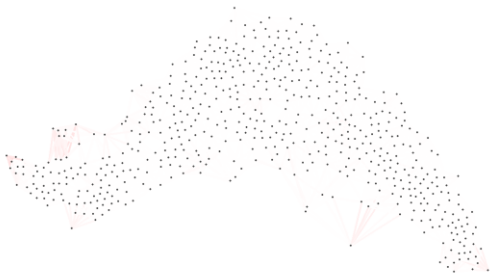


Figure 96 Reliability map des instances du TDP avec pénalisation des accélérations (λ = 0.045) sur le dataset du Covid

Les résultats présentés rencontrent des difficultés à valider l'hypothèse de la réduction de fiabilité liée au λ car très peu de différences sont notables sur les différentes maps. Les tableaux des scores de trustworthiness auront plus de précision pour analyser ce phénomène.

6.2.4 Trustworthiness

Les tableaux de la page suivante ont pour but de valider avec plus de certitude si un λ plus impactant réduit la fiabilité de la projection. De plus, il est intéressant de comparer les valeurs des λ intermédiaires avec les résultats de l'état de l'art. Les valeurs de fiabilité qu'auraient donné les projections avec les λ maximum sont également présentes pour se rendre compte du problème occasionné par un λ trop élevé.

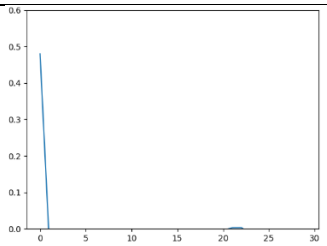
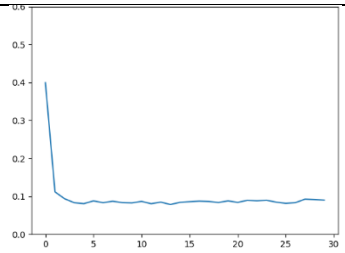
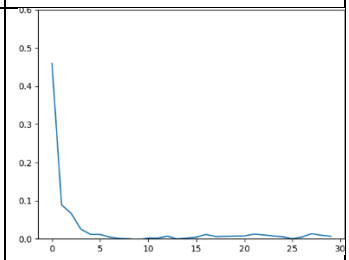
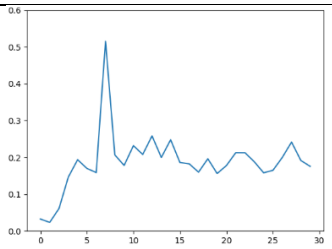
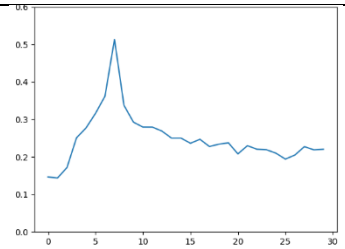
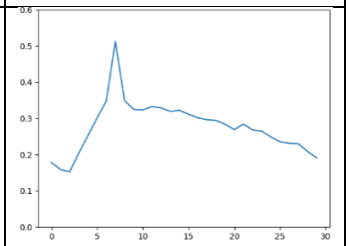
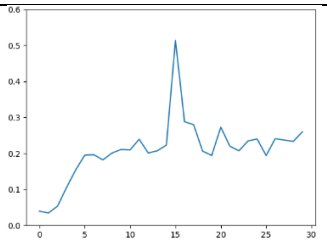
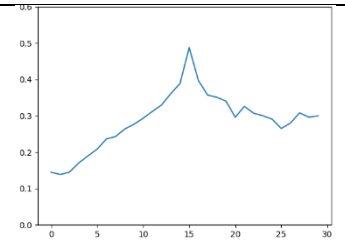
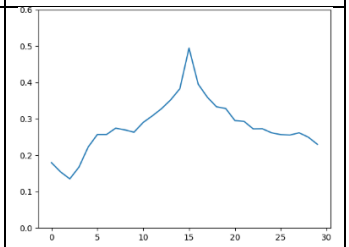
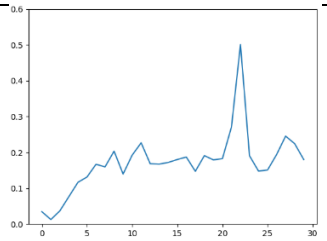
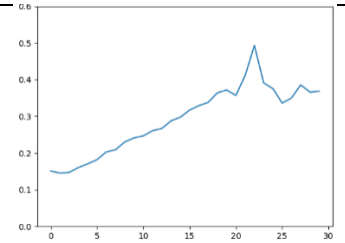
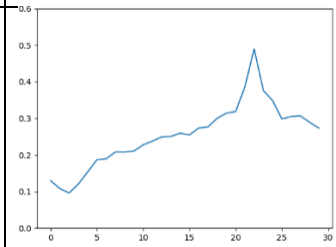
Dataset : Covid	Score de fiabilité [0,1]	Score de fiabilité [0,1]	Score de fiabilité [0,1]	Score de fiabilité [0,1]
λ =	0.1	0.03	0.04	0.045
Mode de pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	Pénalisation des accélérations	Pénalisation des accélérations
Time step concerné				
0/54	0.9847	0.9909	0.9908	0.9913
13/54	0.9786	0.9783	0.9768	0.9411
27/54	0.9747	0.9373	0.9729	0.8500
40/54	0.9827	0.9825	0.9674	0.8824
54/54	0.9803	0.9794	0.9563	0.9615
Moyenne sur tous les *t* :	0.9752	0.9669	0.9552	0.9023

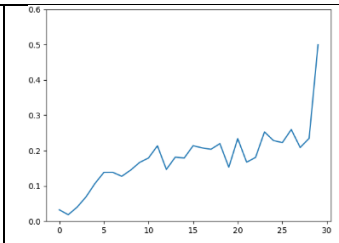
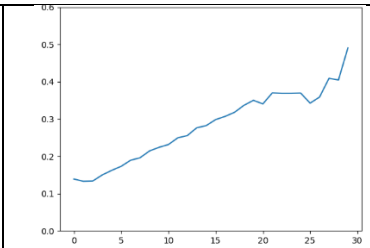
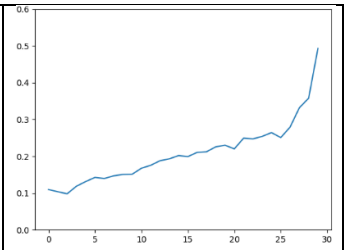
Dataset : SVHN	Score de fiabilité [0,1]	Score de fiabilité [0,1]	Score de fiabilité [0,1]	Score de fiabilité [0,1]
λ	0.1	0.02	0.05	0.07
Mode de pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	Pénalisation des accélérations	Pénalisation des accélérations
Time step t concerné				
0/29	0.8729	0.9161	0.9092	0.9139
7/29	0.7974	0.7921	0.7703	0.7447
15/29	0.7833	0.8090	0.7678	0.7429
22/29	0.7800	0.7883	0.7470	0.7519
29/29	0.7835	0.7705	0.7696	0.7611
Moyenne sur tous les *t* :	0.7973	0.7938	0.7729	0.7579

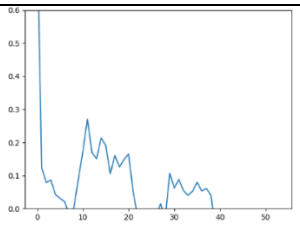
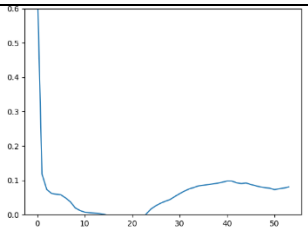
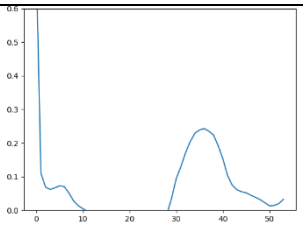
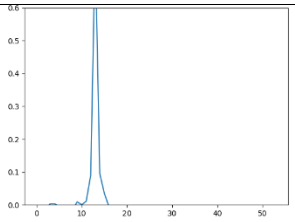
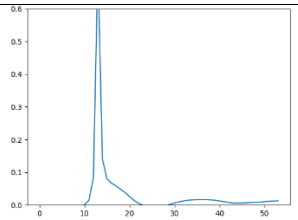
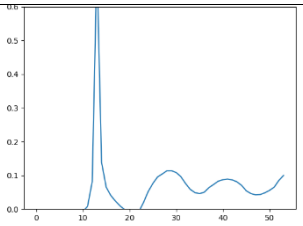
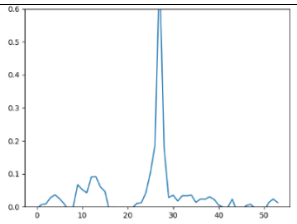
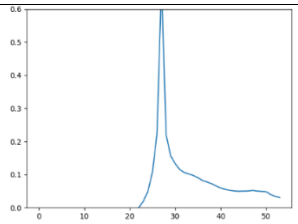
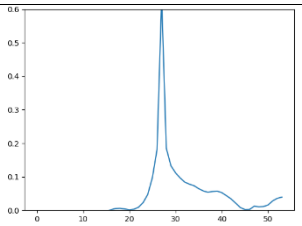
Les résultats obtenus sont parlants, une très légère différence entre la solution de l'état de l'art et la nouvelle est observée lorsque le λ a été réduit, ce qui positionne la solution à une bonne position d'un point de vue fiabilité lorsque le λ est bien configuré. Avec des valeurs trop grandes de λ, les fiabilités chutent. L'hypothèse de l'impact du λ sur la fiabilité des TDP est vérifiée. Une attention particulière doit être prise lors de l'utilisation de cette solution car il n'existe pas officiellement de λ par défaut.

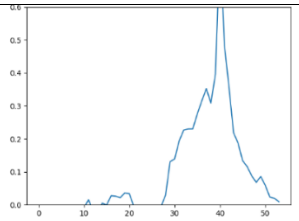
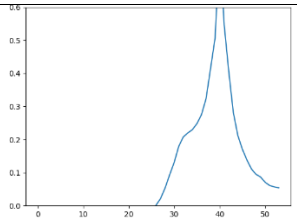
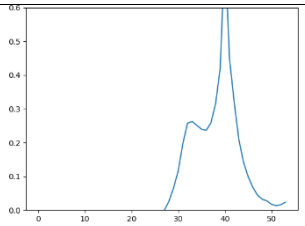
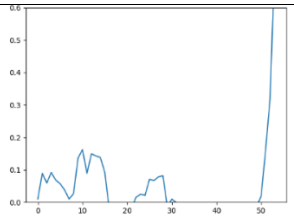
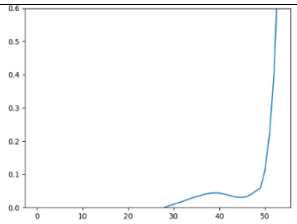
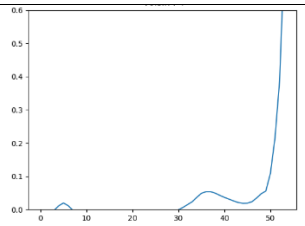
6.2.5 Temporal AUClogRNX

Les résultats présentés à la page suivante reprennent les performances des cohérences des méthodes sans pénalisation, avec pénalisation de mouvement et les comparent avec la nouvelle méthode utilisant un λ sous le seuil de surcharge.

Dataset : SVHN Time step de référence	Dynamic t-SNE Pas de pénalisation λ = 0	Dyn t-SNE Pénalisation mouvement λ = 0.1	Dynamic t-SNE Pénalisation accélération λ = 0.05
t=0	 *Figure 97 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = 0/T*	 *Figure 98 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = 0/T*	 *Figure 99 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = 0/T*
t= T/4	 *Figure 100 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = T/4*	 *Figure 101 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T/4*	 *Figure 102 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T/4*
t= T/2	 *Figure 103 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = T/2*	 *Figure 104 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T/2*	 *Figure 105 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T/2*
t=3*T/4	 *Figure 106 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = 3*T/4*	 *Figure 107 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = 3*T/4*	 *Figure 108 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = 3*T/4*

t=n	 *Figure 109 Temporal AUClogRNX du TDP sans pénalisation sur le dataset SVHN avec pour time step de référence = T*	 *Figure 110 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T*	 *Figure 111 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T*

Dataset : Covid Time step de référence	Dynamic t-SNE λ = 0	Dyn t-SNE Pénalisation mouvement λ = 0.1	Dynamic t-SNE Pénalisation accélération λ = 0.03
t=0	 *Figure 112 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = 0/T*	 *Figure 113 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = 0/T*	 *Figure 114 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence = 0/T*
t=T/4	 *Figure 115 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = T/4*	 *Figure 116 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T/4*	 *Figure 117 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence =T/4*
t= T/2	 *Figure 118 Temporal AUClogRNX du TDP sans pénalisation sur le*	 *Figure 119 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset*	 *Figure 120 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le*

	dataset Covid avec pour time step de référence = T/2	*Covid avec pour time step de référence = T/2*	*dataset Covid avec pour time step de référence =T/2*
t=3*T/4	 *Figure 121 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = 3*T/4*	 *Figure 122 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = 3*T/4*	 *Figure 123 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence =3*T/4*
t=T	 *Figure 124 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Covid avec pour time step de référence = T*	 *Figure 125 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T*	 *Figure 126 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence = T*

Il est observé que la pénalisation des accélérations est meilleure que l'état de l'art pour conserver le voisinage sur la temporalité des datasets. Il est notamment observable sur la Figure 117 que la courbe est plus élevée sur la droite du pic par rapport à la Figure 116. Après quelques réflexions, la pénalisation des accélérations est censée augmenter la liberté des instances à se déplacer selon leur déplacement en haute dimension. Cette liberté augmente donc les chances que le voisinage soit mieux projeté. Les résultats sont meilleurs mais encore loin du comportement attendu de la solution.

6.3 Discussion des résultats

Les résultats des expériences sont satisfaisants sur plusieurs plans. D'un point de vue visuel, les résultats observés montrent des mouvements plus harmonieux et fluides que l'état de l'art. Le principe de moins contraindre les données à prendre de la vitesse contrecarre le problème de la pénalisation des mouvements.

Néanmoins, comme constaté lors du relevé des mesures des scores de fiabilité, une attention particulière doit être prise lors du choix du λ, ce dernier ne peut pas toujours se permettre d'être le plus grand possible au risque de déclencher une chute de la fiabilité des projections produites.

Le temporal AUClogRNX montre également que la pénalisation de l'accélération fournit des projections présentant des voisinages plus intéressants que celles de l'état de l'art. Cependant, les résultats fournis sont encore loin des résultats attendus. Une nouvelle « famille » de solutions est présentée dès le chapitre 50 tentant de résoudre les problèmes de similarité de voisinages.

7 TCP

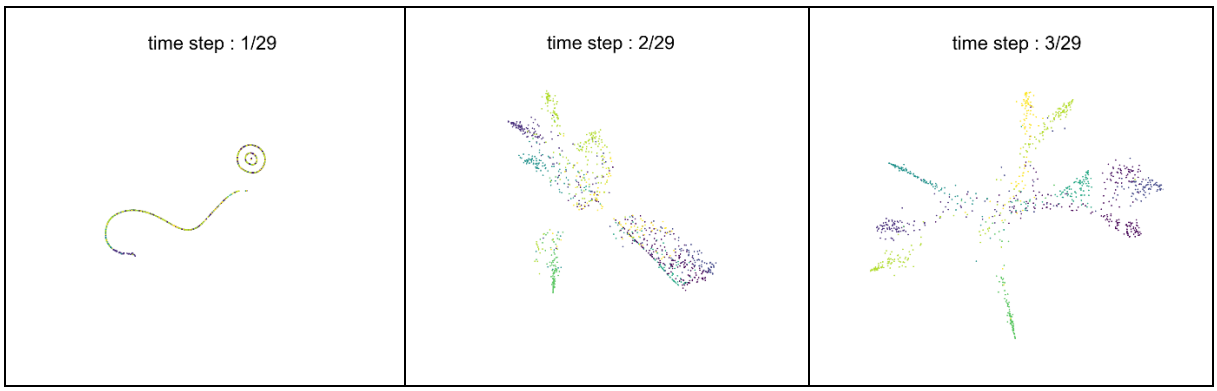
La fonction de coût est aussi améliorable. Après avoir effectué tous les tests nécessaires sur la pénalisation de l'accélération, un changement de la façon dont le coût d'un t-SNE a été rediscuté. Serait-il possible de confectionner une fonction de coût ne pénalisant aucun comportement des données tout en conservant la cohérence temporelle de time step en time step. Que donneraient les résultats sur les autres métriques ? C'est la partie que va aborder ce chapitre.

La fonction calculant actuellement le coût de chaque t-SNE de la temporalité le fait de manière très indépendante et non corrélée. Aucune mécanique ne pousse l'entraînement à se régler automatiquement sur le résultat du time step précédent. Cette manière d'opérer rend l'apprentissage plus complexe et contraignant pour la technique de TDP. Alors de quelle manière est-il possible de fournir cette fonctionnalité à la formule de régression ? La réponse peut se trouver dans la manière dont sont traitées les données lors de l'apprentissage. A chaque itération (ou « époque »), chaque t-SNE est entraîné indépendamment l'un de l'autre. Il serait intéressant de pouvoir apprendre et adapter les positions des instances par rapport à leur position au time step suivant. Ainsi, la position des instances à l'étape t+1 influence la position des points à l'étape t, ce qui, indirectement, force tous les points à « s'auto-corréler » temporellement. Notre méthode s'appelle le temporal coherency preservation (ou encore TCP).

Un avantage avec cette technique est l'absence du paramètre λ. Puisqu'aucune pénalisation supplémentaire n'est prévue, plus aucun facteur ne justifie la présente du λ dans la fonction. Cependant, un désavantage certain apparaît pour les datasets d'une certaine envergure. En effet, une grande partie des calculs se base sur le calcul de matrice de distance. Ces matrices représentent toutes les distances euclidiennes respectives de toute paire d'instance réalisable à un même time step. Pour un dataset de 1000 éléments, cette matrice est composée d'1 million de valeurs. Pour l'implémentation de cette solution, le nombre d'instances à considérer est doublé car deux time steps sont fusionnés en un. Le dataset de 1000 occurrences se retrouve donc doublé ce qui va quadrupler la quantité de valeurs à calculer. Il faut donc s'attendre à des temps d'apprentissage prenant jusqu'à 4 fois plus de temps.

7.1 Visualisation de la solution

Ci-dessous sont représentés les tableaux de figures reprenant les 12 premières images des animations de la solution TCP sur le dataset du Covid et du SVHN.



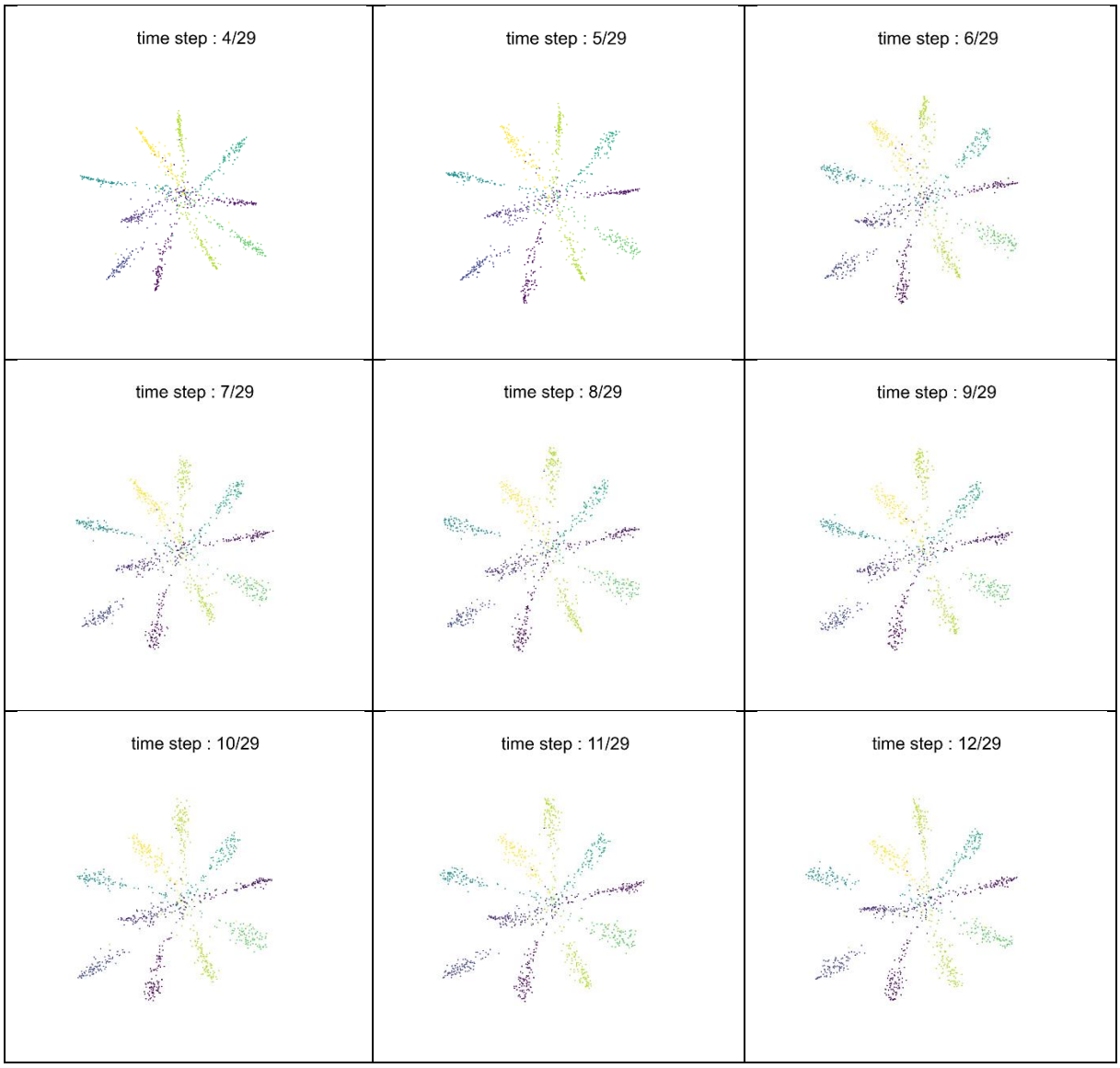
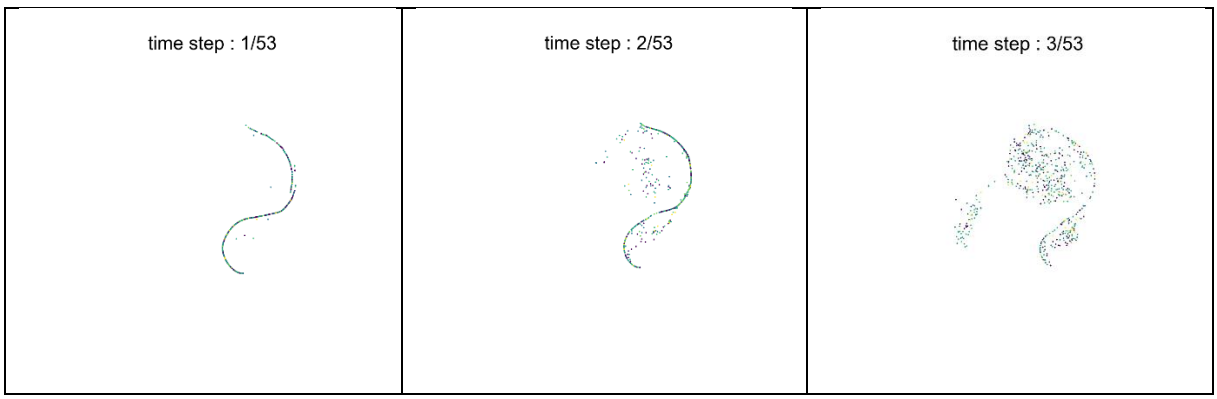


Figure 127 Animation d'une succession de projections générées par le TCP sur le dataset SVHN



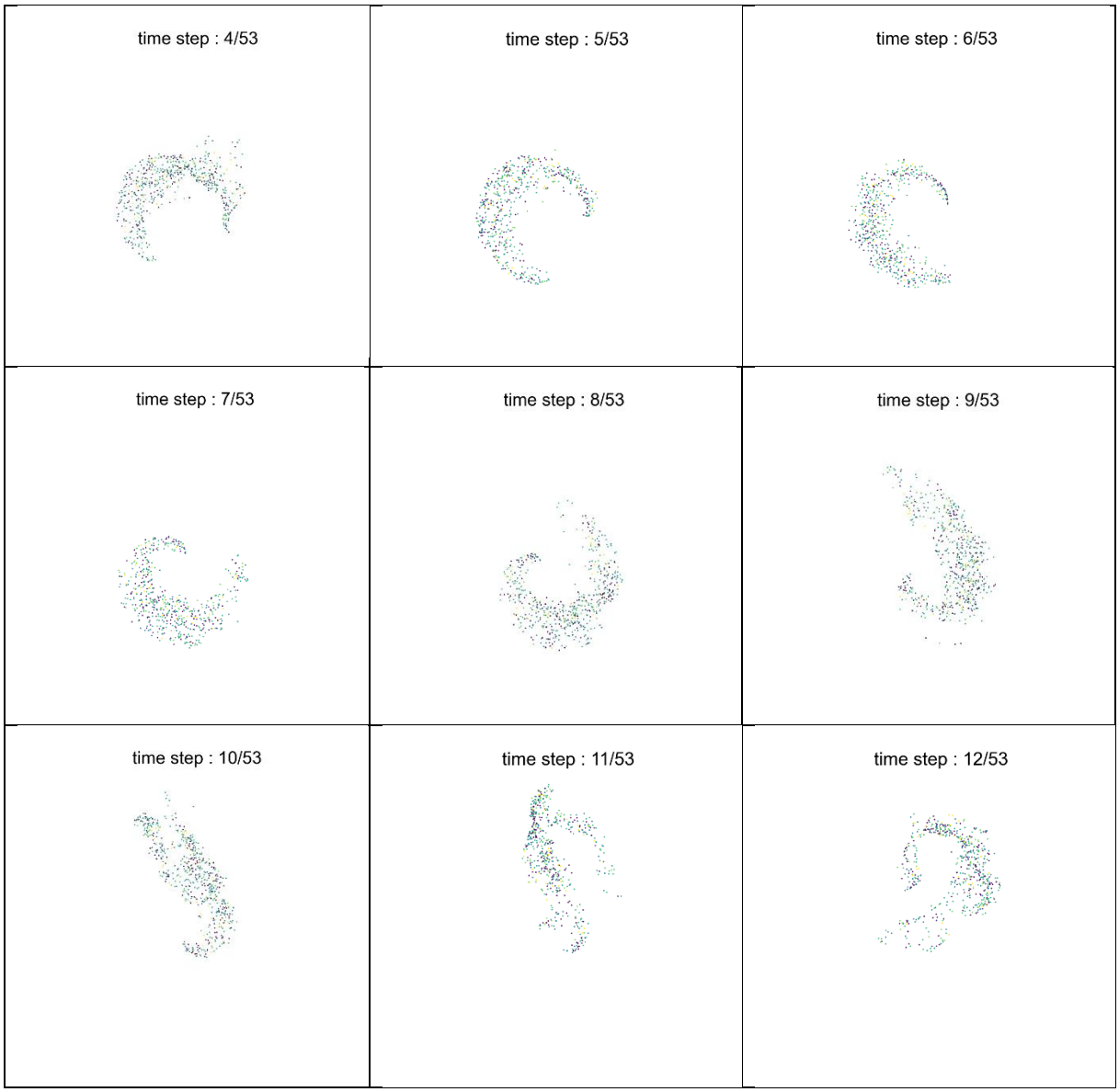


Figure 128 Animation d'une succession de projections générées par le TCP sur le dataset du Covid

7.2 Résultats des métriques

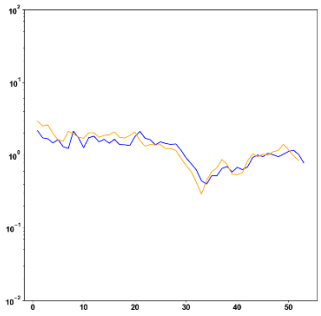
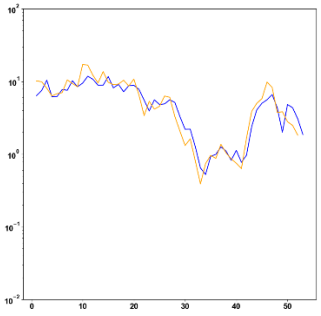
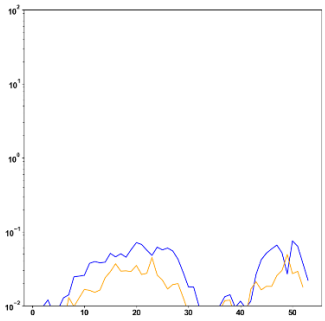
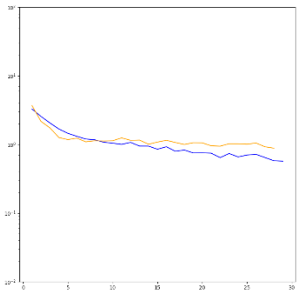
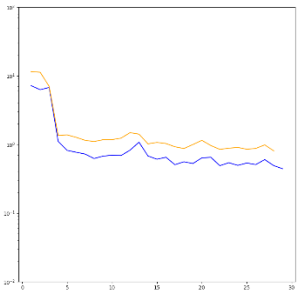
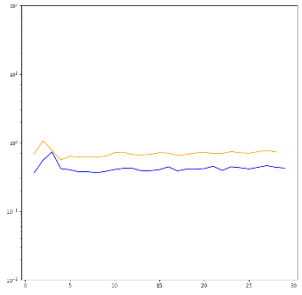
Cette section apporte une interprétation des résultats obtenus à partir de l'implémentation de la nouvelle formule de coût. Comme les deux dernières solutions, des comparaisons et interprétations sont effectuées pour chacune des métriques.

Durant les tests, les résultats visuels et les valeurs de fiabilités des visualisations ont fourni des résultats particulièrement insatisfaisants, voire complètement ridicules. Avec des valeurs moyennes aux alentours de 0.5 pour les fiabilités et des animations très éparpillées sans vraiment de structure. Quelque chose perturbe les expériences. Après plusieurs phases de recherches et d'expérimentations, il s'avère que c'est la perplexité qui pose un problème. Dans l'état actuel des choses, la perplexité est définie une fois pour toutes pour chaque dataset. Mais le fait d'avoir modifié la manière dont la fonction de coût gérait les données d'un time rend la perplexité beaucoup trop faible pour fonctionner comme attendu. La perplexité doit au moins être doublée pour obtenir des résultats intéressants.

Les expériences et résultats suivants sont comparés avec les résultats de l'état de l'art utilisant le λ recommandé, 0.1. La dynamique et les changements de direction sont également comparés avec les mêmes métriques des instances en haute dimension.

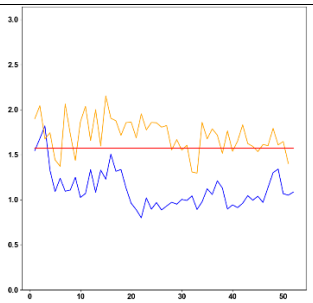
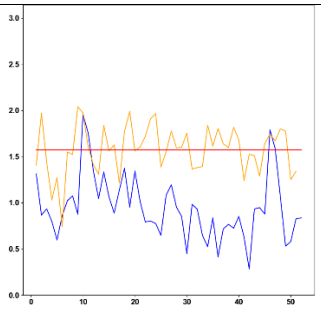
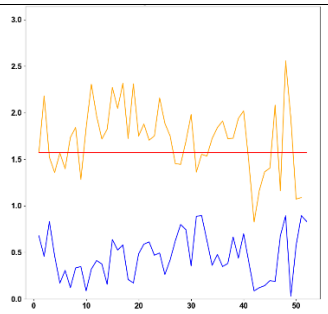
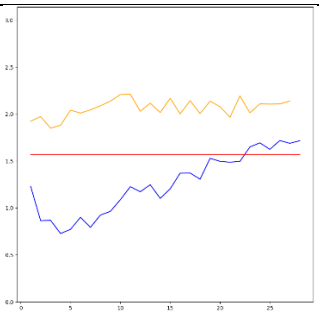
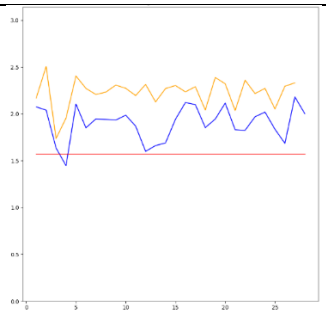
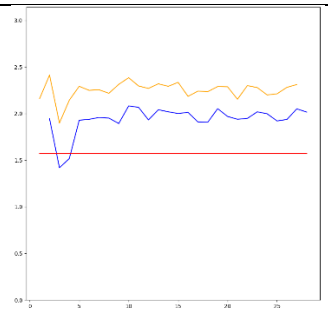
7.2.1 Conservation de la dynamique

Le TCP devrait fournir des résultats suffisamment différents des deux autres solutions par la structure différente de sa fonction de coût.

Dataset	Pénalisation des mouvements (Etat de l'art) λ = 0.1	TCP	Dynamique en haute dimension
Covid	 *Figure 129 Dynamique des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid*	 *Figure 130 Dynamique des instances du TDP avec TCP sur le dataset Covid*	 *Figure 131 Dynamique en Haute dimension des instances du dataset Covid*
SVHN	 *Figure 132 Dynamique des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN*	 *Figure 133 Dynamique des instances du TDP avec TCP sur le dataset SVHN*	 *Figure 134 Dynamique en Haute dimension des instances du dataset SVHN*

La préservation de la dynamique n'a pas vraiment de spécificité dans le cas du TCP. Une certaine similitude avec la dynamique de la pénalisation des mouvements apparaît mais les vecteurs de vitesse et d'accélération restent bien plus élevés comparés à l'état de l'art. Pour le dataset du Covid, une diminution de la dynamique est observée vers les deux tiers du graphique. Cet endroit indique le moment où la vaccination de la première et deuxième dose a commencé à ralentir et la hausse de la dynamique qui suit cette réduction est due au moment où la campagne de la troisième dose a débuté.

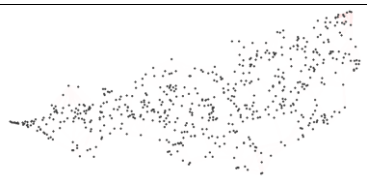
7.2.2 Conservation de la direction

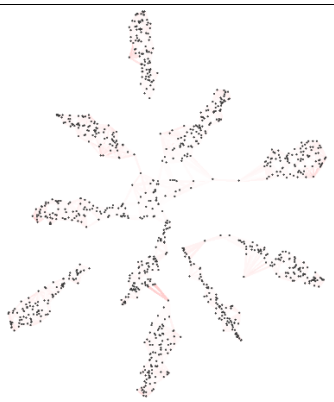
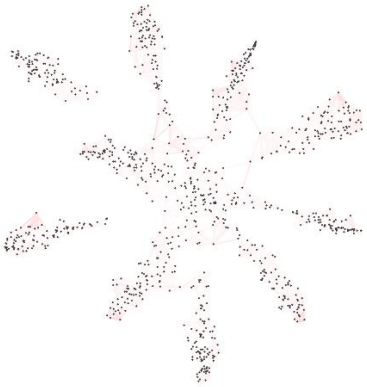
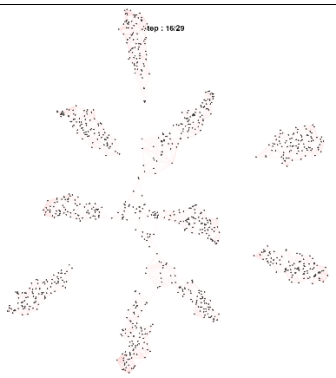
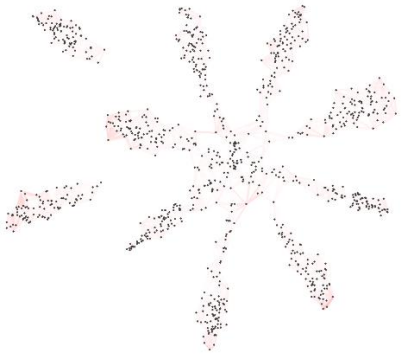
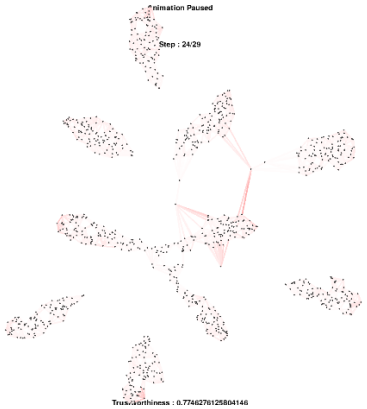
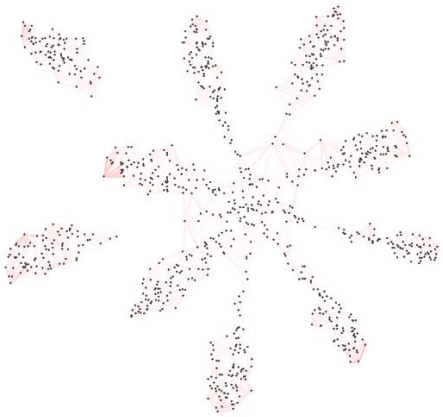
Dataset	Pénalisation des mouvements (Etat de l'art) λ = 0.1	TCP	Dynamique en haute dimension
Covid	 *Figure 135 Changements de direction des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid*	 *Figure 136 Changements de direction des instances du TDP avec TCP sur le dataset Covid*	 *Figure 137 Changements de direction des instances en haute dimension sur le dataset Covid*
SVHN	 *Figure 138 Changements de direction des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN*	 *Figure 139 Changements de direction des instances du TDP avec TCP sur le dataset SVHN*	 *Figure 140 Changements de direction des instances en haute dimension sur le dataset SVHN*

La similitude entre la courbe des changements de direction en haute dimension et celle du TCP est impressionnante. Alors que la pénalisation des accélérations se rapproche de la réalité, c'est avec une étonnante précision que la solution réussit à retransmettre la réalité des changements de direction des instances dans sa projection (surtout présente sur le dataset du SVHN).

7.2.3 Reliability map

Changer la fonction de régression change-t-il en mieux les visualisations des instances projetées ? C'est via les reliability maps de l'état de l'art que la comparaison va se faire.

Dataset : Covid	Pénalisation des mouvements λ = 0.1	TCP
Time step = 15	 *Figure 141 Reliability map des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset du Covid au time step 15*	 *Figure 142 Reliability map des instances du TDP avec TCP sur le dataset du Covid au time step 15/54*
Time step = 30	 *Figure 143 Reliability map des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset du Covid au time step 30/54*	 *Figure 144 Reliability map des instances du TDP avec TCP sur le dataset du Covid au time step 30/59*
Time step = 45	 *Figure 145 Reliability map des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset du Covid au time step 45/54*	 *Figure 146 Reliability map des instances du TDP avec TCP sur le dataset du Covid au time step 45/54*

Dataset : SVHN	Pénalisation des mouvements λ = 0.1	TCP
Time step = 8	 *Figure 147 Reliability map des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN au time step 8/29*	 *Figure 148 Reliability map des instances du TDP avec TCP sur le dataset SVHN au time step 8/29*
Time step = 16	 *Figure 149 Reliability map des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN au time step 16/29*	 *Figure 150 Reliability map des instances du TDP avec TCP sur le dataset SVHN au time step 16/29*
Time step = 24	 *Figure 151 Reliability map des instances du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN au time step 24/29*	 *Figure 152 Reliability map des instances du TDP avec TCP sur le dataset SVHN au time step 24/29*

Il est encore compliqué de déterminer quelle solution fournit la projection la plus fiable. La métrique de la reliability map trouve sa faiblesse quand deux projections comparées se ressemblent,

car, pour rappel, l'interprétation de la reliability map est subjective. Pour rappel, la manière de pallier au problème de visualisation des différences de fiabilité est de se tourner vers les valeurs de trustworthiness, ce qui est le cas pour ces deux solutions.

7.2.4 Trustworthiness

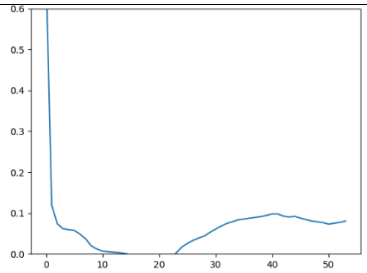
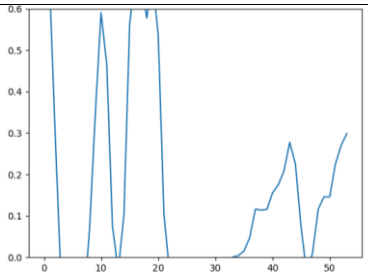
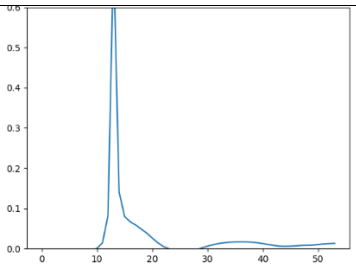
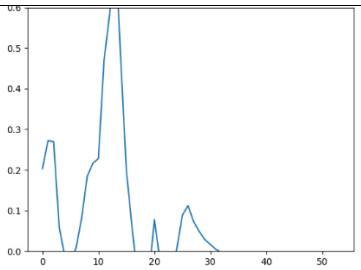
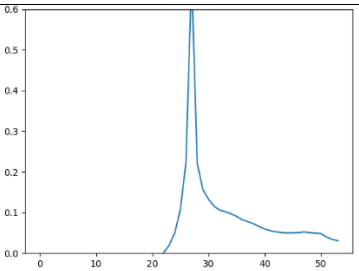
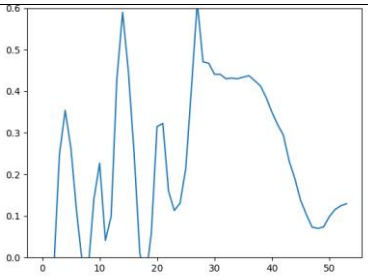
Pour pallier le problème d'interprétation de la reliability map, un tableau comparatif entre les scores de trustworthiness de l'état de l'art et du TCP est dressé.

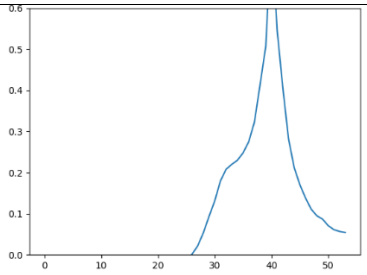
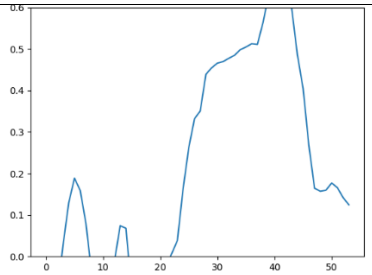
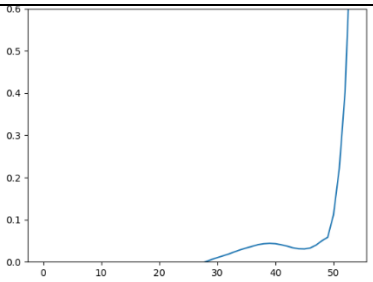
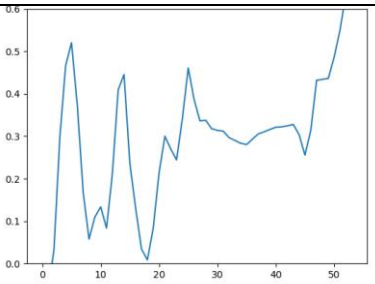
Dataset	Covid	Covid
Mode de pénalisation	Pénalisation des vecteurs de mouvement λ = 0.1	TCP
Time step t concerné / Temporalité T		
0/54	0.9847	0.9945
13/54	0.9786	0.9534
27/54	0.9747	0.9732
40/54	0.9827	0.9837
54/54	0.9803	0.9751
Moyenne sur tous les time steps :	0.9752	0.9720

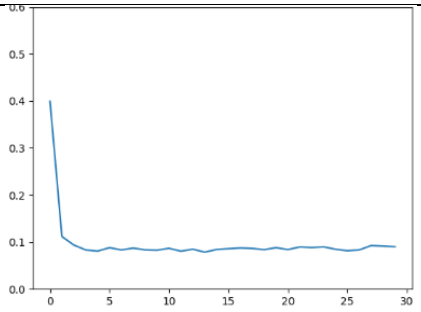
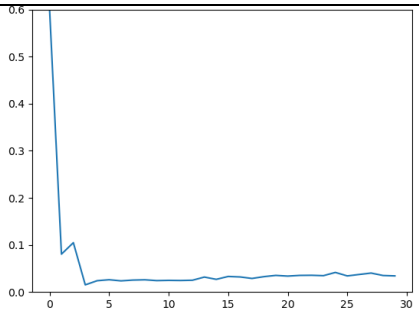
Dataset	SVHN	SVHN
Mode de pénalisation	Pénalisation des vecteurs de mouvement	TCP
Time step t concerné / Temporalité T		
0/29	0.8729	0.9438
7/29	0.7974	0.8084
15/29	0.7833	0.8046
22/29	0.7800	0.8130
29/29	0.7835	0.8175
Moyenne sur tous les time steps :	0.7973	0.8231

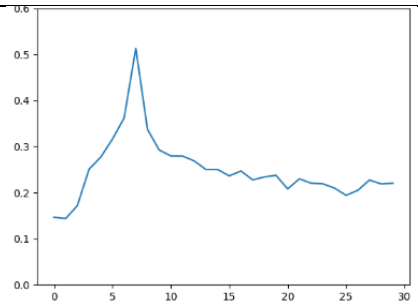
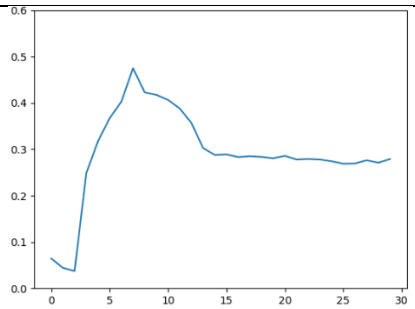
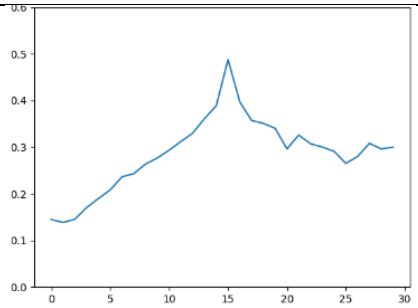
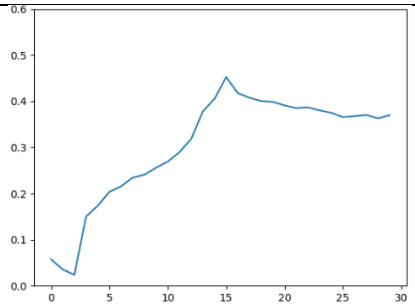
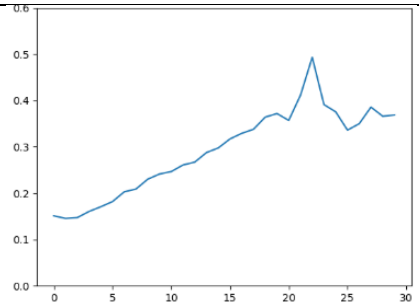
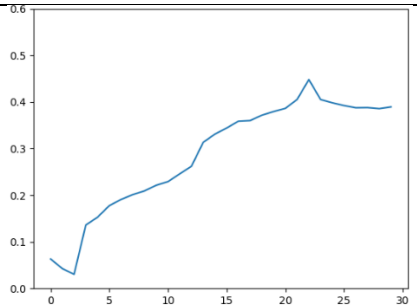
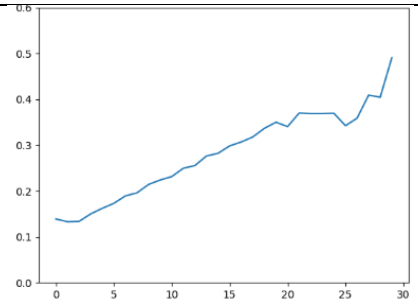
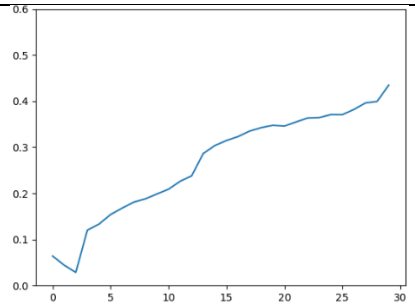
Les résultats comparés sont intéressants. Le TCP est une solution aux résultats similaires à ceux de l'état de l'art concernant le dataset du Covid. Pour le SVHN, la solution est même plus efficace que l'état de l'art. C'est un très bon point pour la solution car, comme spécifié dans les expériences sur l'état de l'art, il est très important d'être performant dans cette métrique et, pour l'instant, c'est la solution la plus efficace pour retranscrire avec le plus de fiabilité les voisins des instances.

7.2.5 Temporal AUClogRNX

Dataset: Covid Time step de référence	Dyn t-SNE Pénalisation mouvement λ = 0.1	TCP
t=0	 *Figure 153 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = 0/T*	 *Figure 154 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = 0/T*
t=T/4	 *Figure 155 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T/4*	 *Figure 156 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = T/4*
t=T/2	 *Figure 157 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T/2*	 *Figure 158 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = T/2*

t=3T/4	 *Figure 159 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = 3*T/4*	 *Figure 160 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = 3*T/4*
t=T	 *Figure 161 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset Covid avec pour time step de référence = T*	 *Figure 162 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = T*

Dataset: SVHN Time step de référence	Dyn t-SNE Pénalisation mouvement λ = 0.1	Temporal dimension reduction
t=0	 *Figure 163 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = 0/T*	 *Figure 164 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = 0/T*

t=T/4	 *Figure 165 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T/4*	 *Figure 166 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = T/4*
t=T/2	 *Figure 167 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T/2*	 *Figure 168 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = T/2*
t=3T/4	 *Figure 169 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = 3*T/4*	 *Figure 170 Temporal AUClogRNX | TCP | SVHN time step de référence = 3*T/4*
t=T	 *Figure 171 Temporal AUClogRNX du TDP avec pénalisation des mouvements (λ = 0.1) sur le dataset SVHN avec pour time step de référence = T*	 *Figure 172 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = T*

Cette fois, AUClogRNX s'exprime différemment. Pour SVHN, les valeurs et les courbes restent équivalentes à celles de l'état de l'art. Ce comportement est rassurant, son comportement vis-à-vis d'un dataset a priori ordonné est similaire à toutes les autres solutions décrites. Cependant, pour les données du Covid, les graphiques sont complètement différents. AUClogRNX a fourni des résultats composés de nombreux pics de similarité. Est-ce un comportement plus efficace que l'état de l'art ? Trouver la réponse à cette question n'est pas si simple, mais il est possible de trouver des pistes de réponses en analysant le comportement de la métrique sur les datasets « blob ». Ces analyses sont présentes dans le chapitre 10. Les résultats provenant des autres solutions indiquent que, d'un time step au suivant, presque aucun voisin n'est le même ; ce qui est étonnant à la vue de données concernant la vaccination cumulée de chaque commune de Belgique. Il est donc plus probable que le TCP préserve mieux les voisinages sur le dataset du Covid.

7.3 Discussion des résultats

La dynamique n'est pas la métrique la plus touchée par le TCP ; une ressemblance se dessine avec la courbure de la dynamique des instances projetées avec l'état de l'art. Pour s'en assurer, une vérification est réalisée sur les datasets « blob ». Ces résultats sont disponibles dans le chapitre 10.

Les changements de direction sont fortement préservés par le TCP. La grande force relevée du TCP est sa capacité à conserver les changements de directions avec une étonnante précision. L'hypothèse expliquant ce phénomène proviendrait du fait qu'inclure les positions des instances au temps t+1 influence d'une certaine manière le comportement des instances. Pour vérifier cette hypothèse, il faut analyser les résultats sur les datasets « blob » au chapitre 10.

Les observations sur la reliability map et les mesures de scores de trustworthiness sont unanimes ; TCP est la meilleure solution pour retranscrire le voisinage des instances dans ses projections parmi les solutions décrites dans ce document.

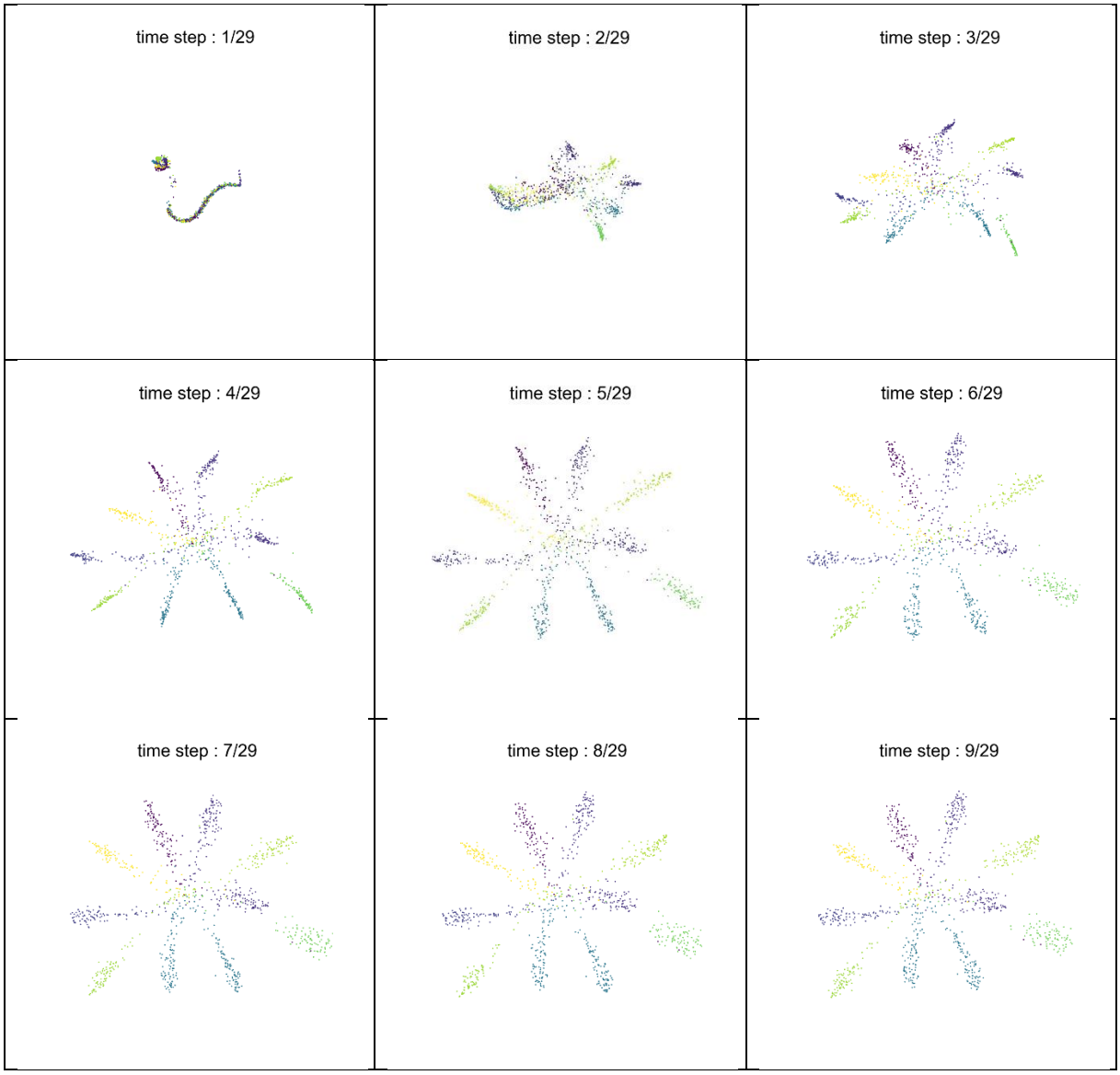
Une légère hausse des similarités temporelles est observée sur le temporal AUClogRNX justifiant l'usage de cette solution par rapport aux autres solutions.

8 Solution hybride

Les inconvénients des précédentes techniques font réfléchir. La fonction de pénalisation des accélérations permet de récupérer une fiabilité et une visualisation plus accrues. Le TCP ajoute un facteur de cohérence temporel et un respect plus important de la structure des données en haute dimension. La solution se base sur la fusion du TCP et sur la pénalisation des accélérations. La principale idée derrière la solution hybride est la fusion entre le TCP et la pénalisation des accélérations. Cette expérience a pour but de tester une solution pouvant surpasser une des deux, voire les deux solutions dont elle est composée.

8.1 Visualisation de la solution

Ci-dessous sont représentés les tableaux de figures reprenant les 12 premières images des animations de la solution hybride sur le dataset du Covid et du SVHN.



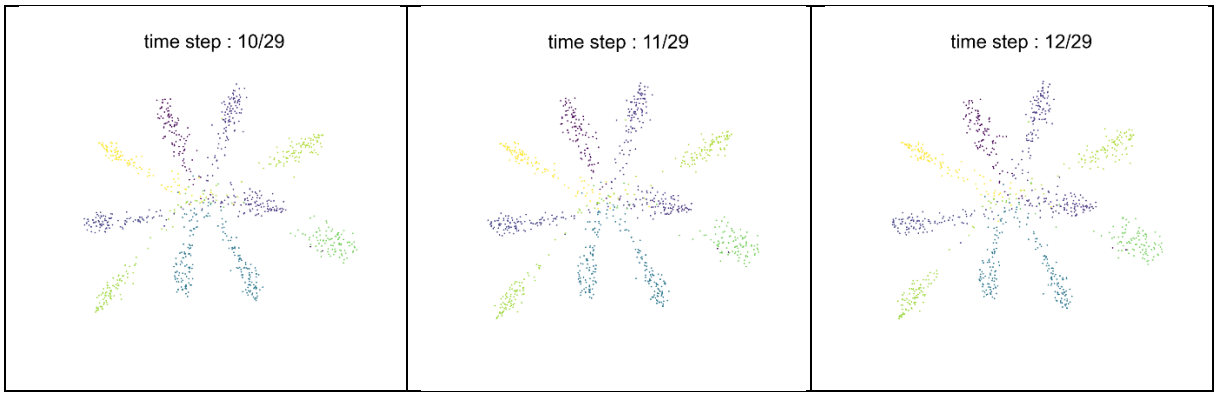
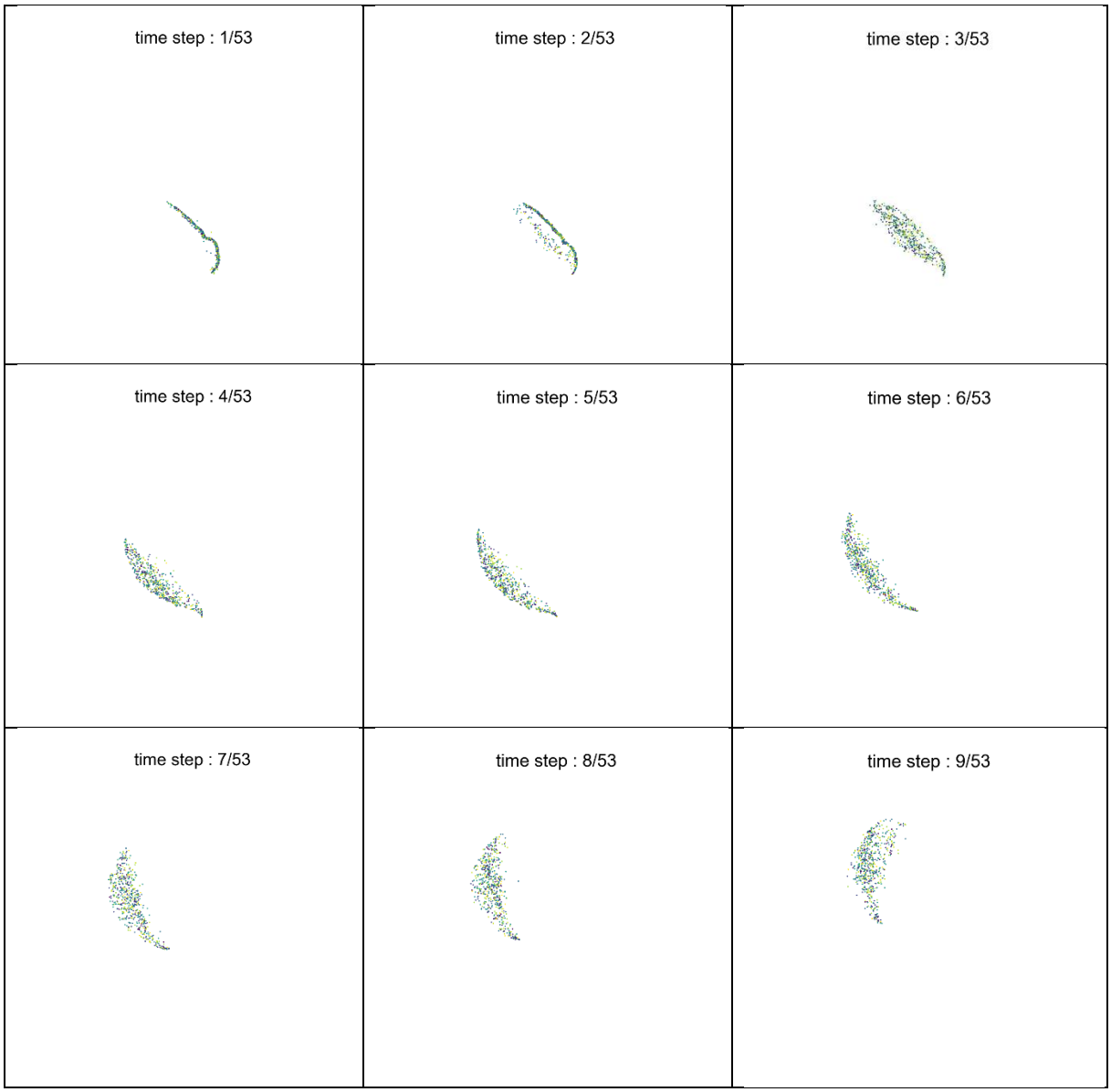


Figure 173 Animation d'une succession de projections générées par la solution hybride sur le dataset SVHN



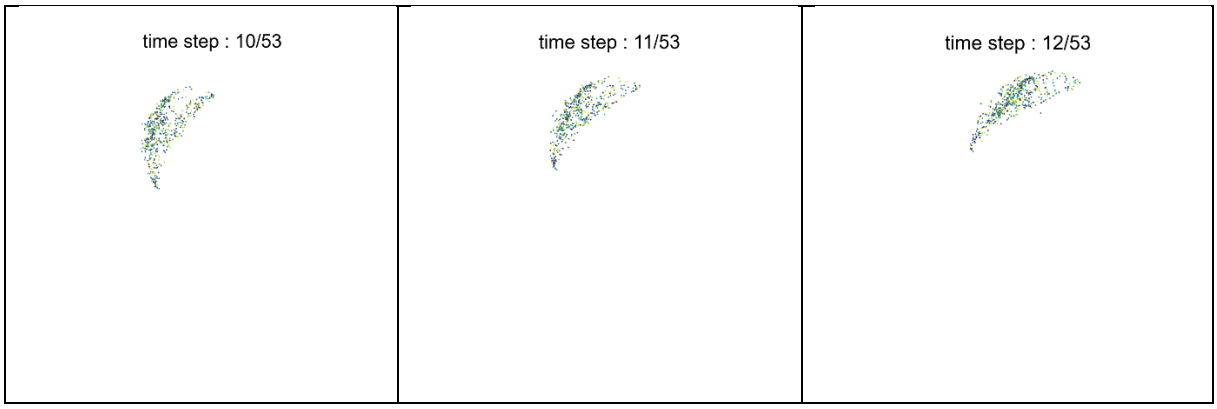


Figure 174 Animation d'une succession de projections générées par la solution hybride sur le dataset du Covid

8.2 Résultats des métriques

Les expériences et résultats suivants sont comparés aux résultats des solutions sur la pénalisation des vecteurs d'accélérations et du TCP. Le but de cette comparaison est d'essayer de capter les contributions de chaque technique dans la version hybride et d'essayer d'obtenir une solution mixant les avantages de chacun tout en atténuant au maximum leurs inconvénients respectifs.

Pour la configuration de l'apprentissage de la solution hybride, un λ de 0.03 pour le Covid et 0.05 pour le SVHN a été choisi sur base des résultats de la solution pénalisant les accélérations. Il a été remarqué que la tranche de valeurs encadrant le λ était relativement similaire à celle de la pénalisation des accélérations ce qui a grandement accéléré sa configuration lors des tests.

8.2.1 Conservation de la dynamique

Dans cette section, les observations sur la dynamique des instances de la solution hybride sont présentées. Un comparatif avec les solutions composites de la version hybride est réalisé.

Dataset : Covid		
Pénalisation des accélérations λ = 0.03	Solution hybride λ = 0.03	TCP

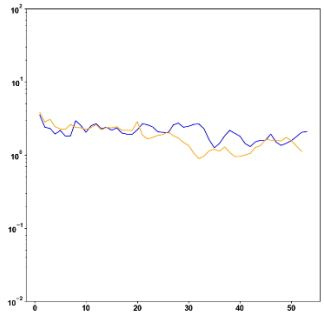
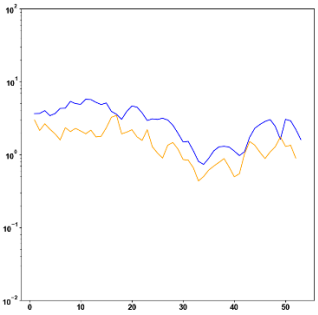
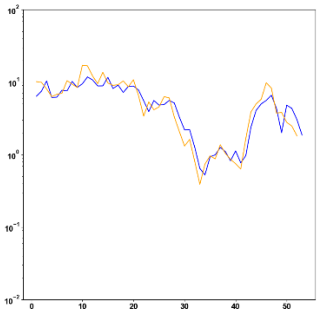

		

Figure 175 Dynamique des instances du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid	*Figure 176 Dynamique des instances du TDP avec la solution Hybride (λ = 0.03) sur le dataset Covid*	*Figure 177 Dynamique des instances du TDP avec TCP sur le dataset Covid*
Dataset : SVHN		

Pénalisation des accélérations λ = 0.05	Solution hybride λ = 0.05	TCP

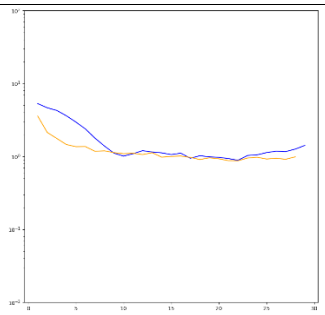
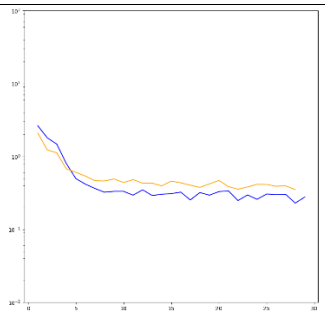
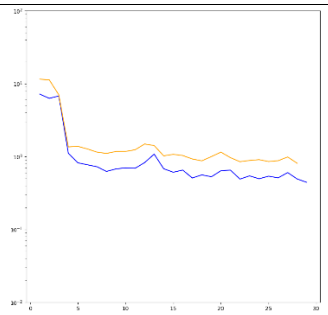

		

Figure 178 Dynamique des instances du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN	*Figure 179 Dynamique des instances du TDP la version hybride (λ = 0.05) sur le dataset SVHN*	*Figure 180 Dynamique des instances du TDP avec TCP sur le dataset SVHN*

Comme anticipé, la dynamique des instances projetées par la solution hybride est un entre-deux des deux autres solutions. Elle conserve la force de préservation du TCP tout en fournissant une visualisation un peu plus intéressante d'un point de vue déplacement (la fluidité du mouvement est en hausse grâce à la pénalisation).

8.2.2 Conservation de la direction

Dataset : Covid		
Pénalisation des accélérations λ = 0.03	Solution hybride λ = 0.03	TCP

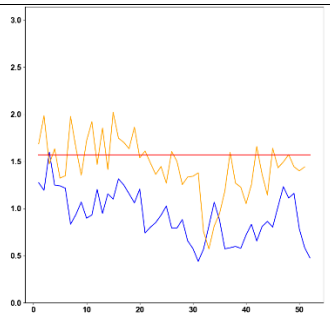
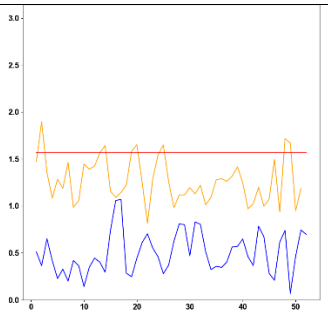
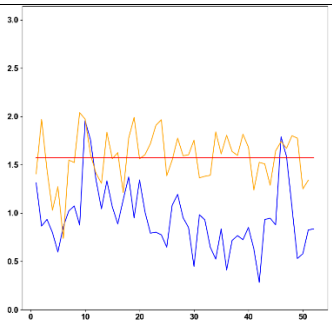

		

Figure 181 Changements de direction des instances du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid	*Figure 182 Changements de direction des instances du TDP avec la version hybride (λ = 0.03) sur le dataset Covid*	*Figure 183 Changements de direction des instances du TDP avec le TCP sur le dataset Covid*
Dataset : SVHN		
Pénalisation des accélérations λ = 0.05	Solution hybride λ = 0.05	TCP

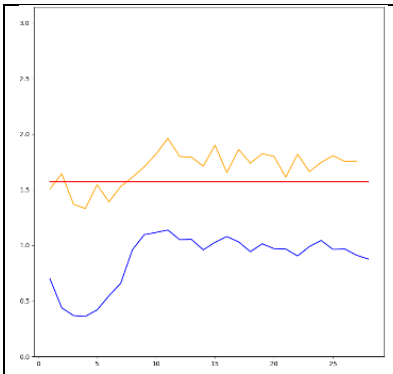


Figure 184 Changements de direction des instances du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN

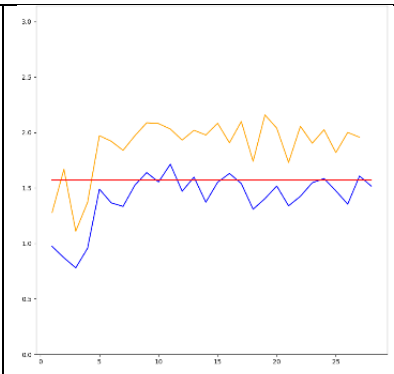


Figure 185 Changements de direction des instances du TDP avec la version hybride (λ = 0.05) sur le dataset SVHN

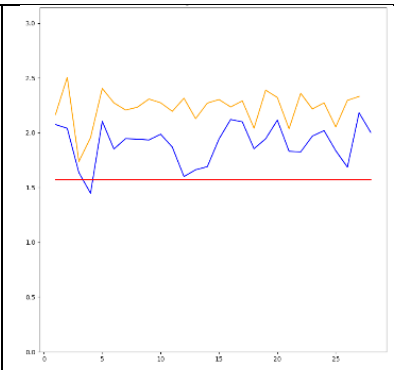
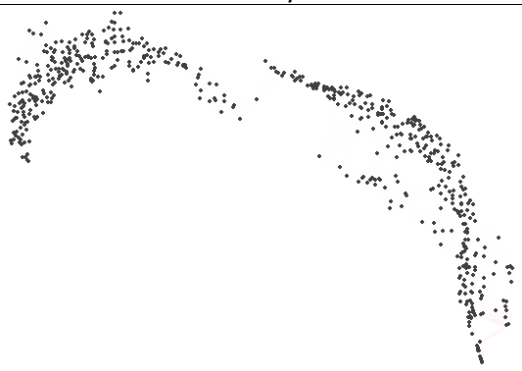


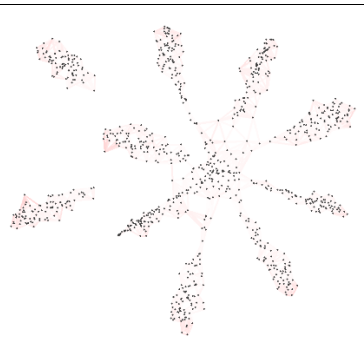
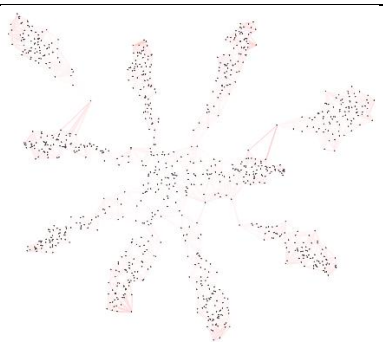
Figure 186 Changements de direction des instances du TDP avec le TCP sur le dataset SVHN

Pour SVHN, la version hybride se situe juste entre les deux solutions, elle préserve les directions presque aussi bien que le TCP, mais est en partie influencée par l'ajout de fluidité de sa pénalisation. Le dataset du Covid fournit une visualisation beaucoup plus impactée au niveau des vecteurs de vitesse et d'accélération. Cette réaction rend la version hybride moins fiable que le TCP qui égalait déjà très bien la réalité des changements de direction des instances du dataset. Il a été prouvé précédemment que toute pénalisation arrivait avec un coût, une chute notable des changements de direction (section 6.2.2).

8.2.3 Reliability map

Les performances de la version hybride comparées au TCP concernant la fiabilité ne peuvent pas être positives. L'ajout de la pénalisation a déjà été la source d'une fiabilité en baisse dans les autres solutions, et la version hybride n'y échappe pas non plus. Comme présenté sur les quelques exemples ci-dessous, ce sont réellement les scores de trustworthiness qui vont départager les deux solutions car il est visuellement complexe de déterminer à quel point la version hybride est plus faible que le TCP.

	Solution hybride	TCP
Dataset : Covid	 *Figure 187 Reliability map des instances du TDP avec la solution hybride (λ = 0.03) sur le dataset Covid d'un time step aléatoire*	 *Figure 188 Reliability map des instances du TDP avec TCP sur le dataset du Covid d'un time step aléatoire*

Dataset : SVHN	 *Figure 189 Reliability map des instances du TDP avec la solution hybride (λ = 0.05) sur le dataset SVHN d'un time step aléatoire*	 *Figure 190 Reliability map des instances du TDP avec TCP sur le dataset SVHN d'un time step aléatoire*

8.2.4 Trustworthiness

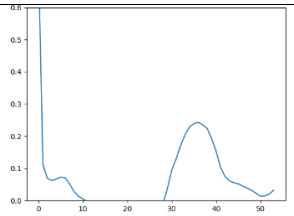
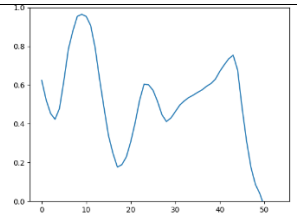
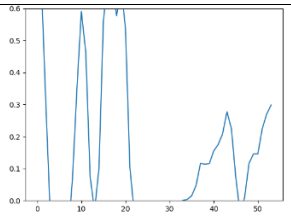
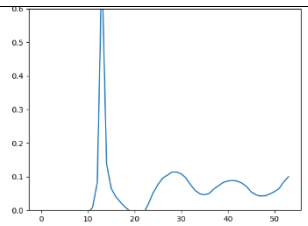
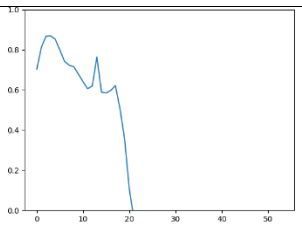
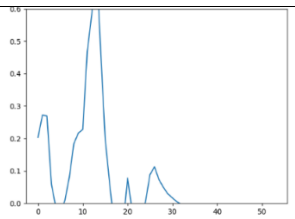
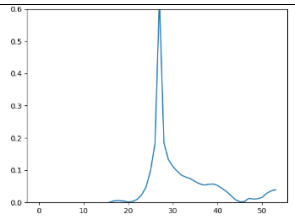
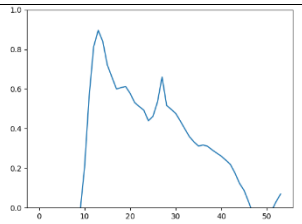
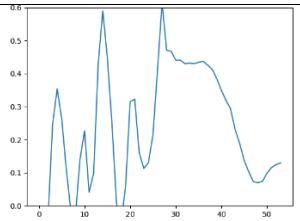
Le tableau comparatif des scores de trustworthiness ci-dessous met en relation des scores relatifs des trois solutions concernées. L'idée ici est de déterminer si la version hybride est capable d'égaler les compétences du TCP malgré les doutes. Dans un second temps, il est intéressant également de comparer les résultats de la solution avec celle de la pénalisation des accélérations dont est issue la solution hybride.

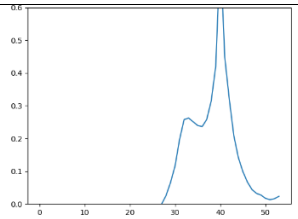
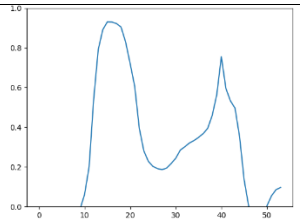
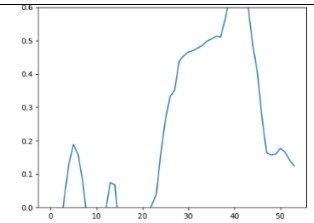
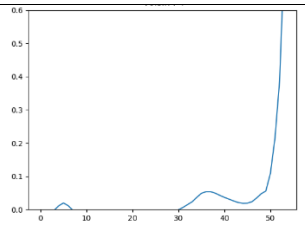
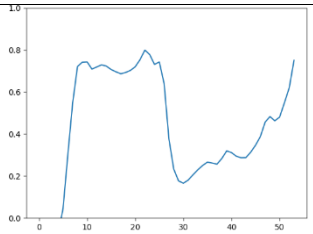
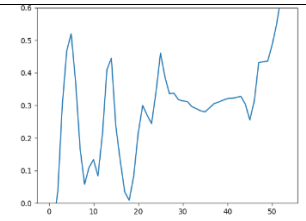
Dataset	Covid	Covid	Covid
Mode de pénalisation	Pénalisation des accélérations λ = 0.03	Solution hybride λ = 0.03	TCP
Time step t concerné / Temporalité T			
0/54	0.9909	0.9942	0.9945
13/54	0.9783	0.9903	0.9534
27/54	0.9373	0.9799	0.9732
40/54	0.9825	0.9820	0.9837
54/54	0.9794	0.9783	0.9751
Moyenne sur tous les time steps :	0.9669	0.9821	0.9720

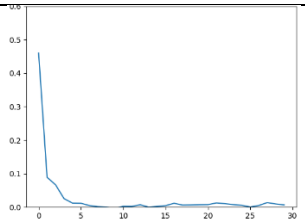
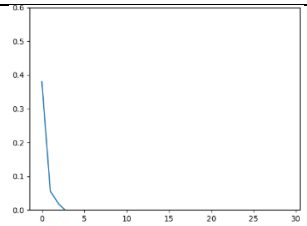
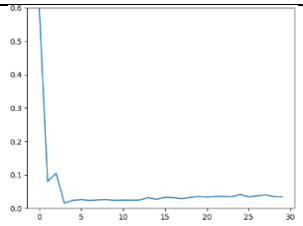
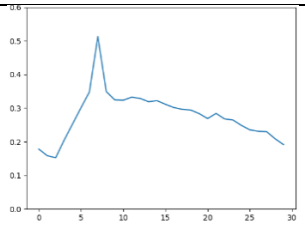
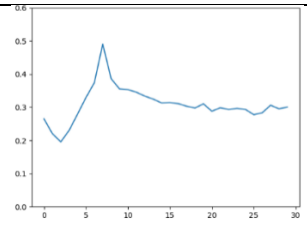
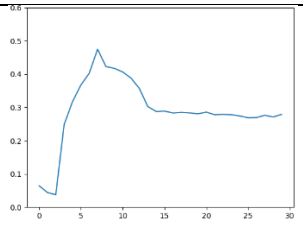
Dataset	SVHN	SVHN	SVHN
Mode de pénalisation	Pénalisation des accélérations λ = 0.05	Solution hybride λ = 0.05	TCP
Time step t concerné / Temporalité T			
0/29	0.9092	0.9351	0.9438
7/29	0.7703	0.8251	0.8084
15/29	0.7678	0.7993	0.8046
22/29	0.7470	0.7910	0.8130
29/29	0.7696	0.7976	0.8175
Moyenne sur tous les time steps :	0.7729	0.8138	0.8231

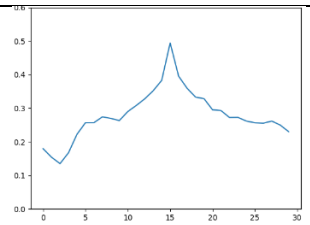
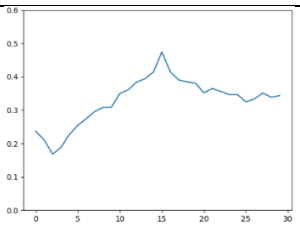
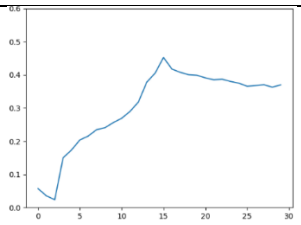
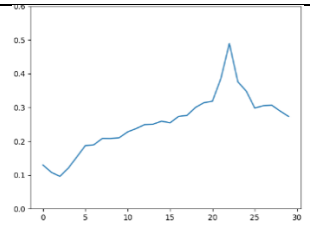
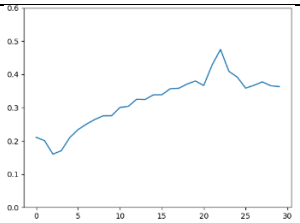
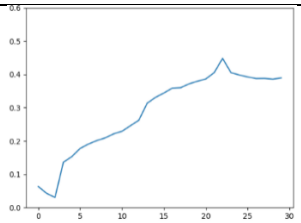
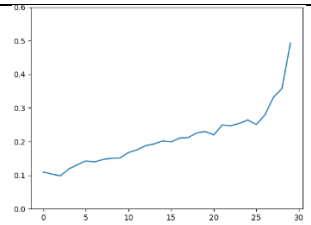
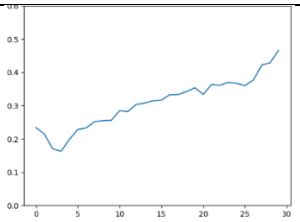
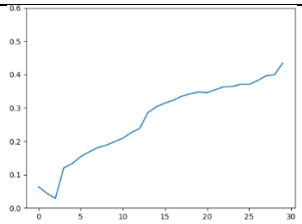
Comme attendu, la version hybride surpasse en tout point la pénalisation des accélérations mais n'égale pas la précision de retranscription des voisinages du TCP sur le dataset SVHN. Cependant, pour le dataset du Covid, les scores de trustworthiness sont plus élevés que le TCP, ce qui était complètement inattendu. Avec une marge d'erreur, le TCP et la version hybride semblent se ressembler d'un point de vue fiabilité ; une différence de 1% de fiabilité est suffisamment faible pour déclarer les deux solutions équivalentes sur cette métrique.

8.2.5 Temporal AUClogRNX

Dataset: Covid Time step de référence	Pénalisation des accélérations λ = 0.03	Solution hybride λ = 0.03	TCP
t=0	 *Figure 191 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence = 0/T*	 *Figure 192 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.03) sur le dataset Covid avec pour time step de référence = 0/T*	 *Figure 193 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = 0/T*
t=n/4	 *Figure 194 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence = T/4*	 *Figure 195 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.03) sur le dataset Covid avec pour time step de référence = T/4*	 *Figure 196 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = T/4*
t=n/2	 *Figure 197 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence = T/2*	 *Figure 198 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.03) sur le dataset Covid avec pour time step de référence = T/2*	 *Figure 199 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = T/2*

t=3n/4	 *Figure 200 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence = 3*T/4*	 *Figure 201 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.03) sur le dataset Covid avec pour time step de référence = 3*T/4*	 *Figure 202 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = 3*T/4*
t = n	 *Figure 203 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.03) sur le dataset Covid avec pour time step de référence = T*	 *Figure 204 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.03) sur le dataset Covid avec pour time step de référence = T*	 *Figure 205 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence = T*

Dataset: SVHN Time step de référence	Pénalisation des accélérations λ = 0.05	Solution hybride λ = 0.05	TCP
t=0	 *Figure 206 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = 0/T*	 *Figure 207 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.05) sur le dataset SVHN avec pour time step de référence = 0/T*	 *Figure 208 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = 0/T*
t=n/4	 *Figure 209 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T/4*	 *Figure 210 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T/4*	 *Figure 211 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = T/4*

t=n/2	 *Figure 212 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T/2*	 *Figure 213 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T/2*	 *Figure 214 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = T/2*
t=3n/4	 *Figure 215 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = 3*T/4*	 *Figure 216 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.05) sur le dataset SVHN avec pour time step de référence = 3*T/4*	 *Figure 217 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = 3*T/4*
t=n	 *Figure 218 Temporal AUClogRNX du TDP avec pénalisation des accélérations (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T*	 *Figure 219 Temporal AUClogRNX du TDP avec la solution hybride (λ = 0.05) sur le dataset SVHN avec pour time step de référence = T*	 *Figure 220 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence = T*

Les résultats sont ceux attendus pour le dataset SVHN, les courbes de la version hybride se sont logées entre les deux courbes des deux solutions, un peu en dessous du TCP car elle est composée d'une pénalisation qui réduit les similarités et un peu au-dessus de la pénalisation des accélérations qui, elle, est dépourvue de la force de préservation de la solution TCP. Cependant, le dataset du Covid a des résultats bien différents. La version hybride est la première à offrir une cohérence temporelle meilleure que le TCP pour ce dataset.

8.3 Discussion des résultats

L'idée derrière la version hybride est de déterminer si le meilleur de deux solutions pouvait se retrouver dans une seule. La version hybride montre des signes de faiblesse lors des résultats sur le dataset SVHN mais se défend bien sur le dataset du Covid. Comme dit précédemment, les résultats sur les dataset « blob » détermineront les réels tenants et aboutissants de chaque technique.

9 TCP+

Une dernière solution basée sur le TCP est à tester. Pour rappel, TCP se base sur le principe d'inclure aux instances d'un time step t les instances du time step suivant t+1. L'entraînement se fait ensuite sur deux fois plus d'instances que prévu, mais ce processus permet de mettre de côté la partie pénalisation de la fonction de coût.

Après toutes les expériences effectuées, le comportement du dataset du SVHN est facile à comprendre. Ce dernier se basant sur un apprentissage à descente de gradients, les positions de chaque instance tendent donc à se stabiliser et très rapidement, de même pour le voisinage de chaque instance. Il faut environ une dizaine de time steps pour que la convergence s'observe. Dès lors, les voisinages se ressemblent et le temporal AUClogRNX tend vers une courbe constante. Ci-dessous, sur la Figure 221 et la Figure 222 avec un point de référence sur le premier et le dernier time step de la temporalité du dataset SVHN, une estimation de ce qu'il est attendu des courbes temporal AUClogRNX est présentée. Cette estimation a été réalisée à la main et n'a pour but que d'illustrer le style de comportement attendu des courbes du temporal AUClogRNX.

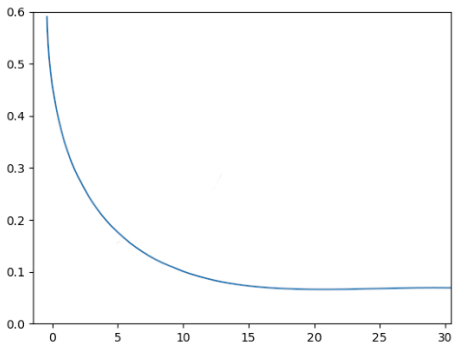


Figure 221 Temporal AUClogRNX parfait attendu d'une projection sur le dataset SVHN avec pour time step de référence 0/29

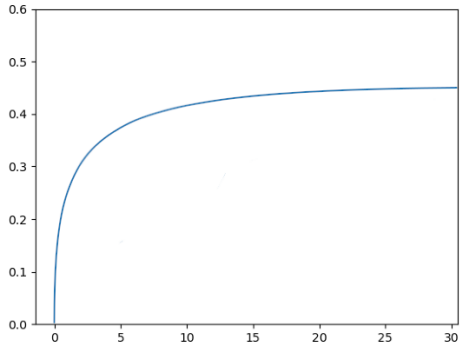


Figure 222 Temporal AUClogRNX parfait attendu d'une projection sur le dataset SVHN avec pour time step de référence 29/29

Malheureusement, une telle estimation est plus compliquée à réaliser sur le dataset du Covid car sa dynamique est moins stable que celle du SVHN.

L'estimation de la Figure 221 et de la Figure 222 est la base de motivation de la nouvelle solution TCP+. Le principe est le même que TCP mais au lieu d'inclure uniquement le time step t+1 au time step t, l'ajout du dernier time step du dataset y est également inclus.

Le TCP+ essaye d'ajouter un point d'ancrage servant d'objectif de voisinage supplémentaire à atteindre sur le dernier time step de la temporalité. TCP+ fait sens avec le dataset SVHN, mais ce n'est pas la meilleure approche concernant le dataset du Covid.

Cette stratégie vient avec un désavantage très impactant. Comme spécifié dans le chapitre 7, fusionner deux time steps le temps de l'apprentissage rend les calculs 4 fois plus lourds (par les matrices de distance étant carrées). Or, pour le TCP+, ce sont 3 time steps qui sont fusionnés ! Le temps de calcul explose donc par un facteur de 9. À titre indicatif, le dataset du Covid prend environ 17 minutes pour s'entraîner sur l'ordinateur qui a servi de test avec le dynamic t-SNE de l'état de l'art. Avec TCP+, cet entraînement passe à 2h25 soit environ 9 fois plus longtemps. Cette solution sera donc fortement handicapée par la taille du dataset qui lui sera fourni.

9.1 Visualisation de la solution

Ci-dessous sont représentés les tableaux de figures reprenant les 12 premières images des animations de la solution TCP+ sur le dataset du Covid et du SVHN.

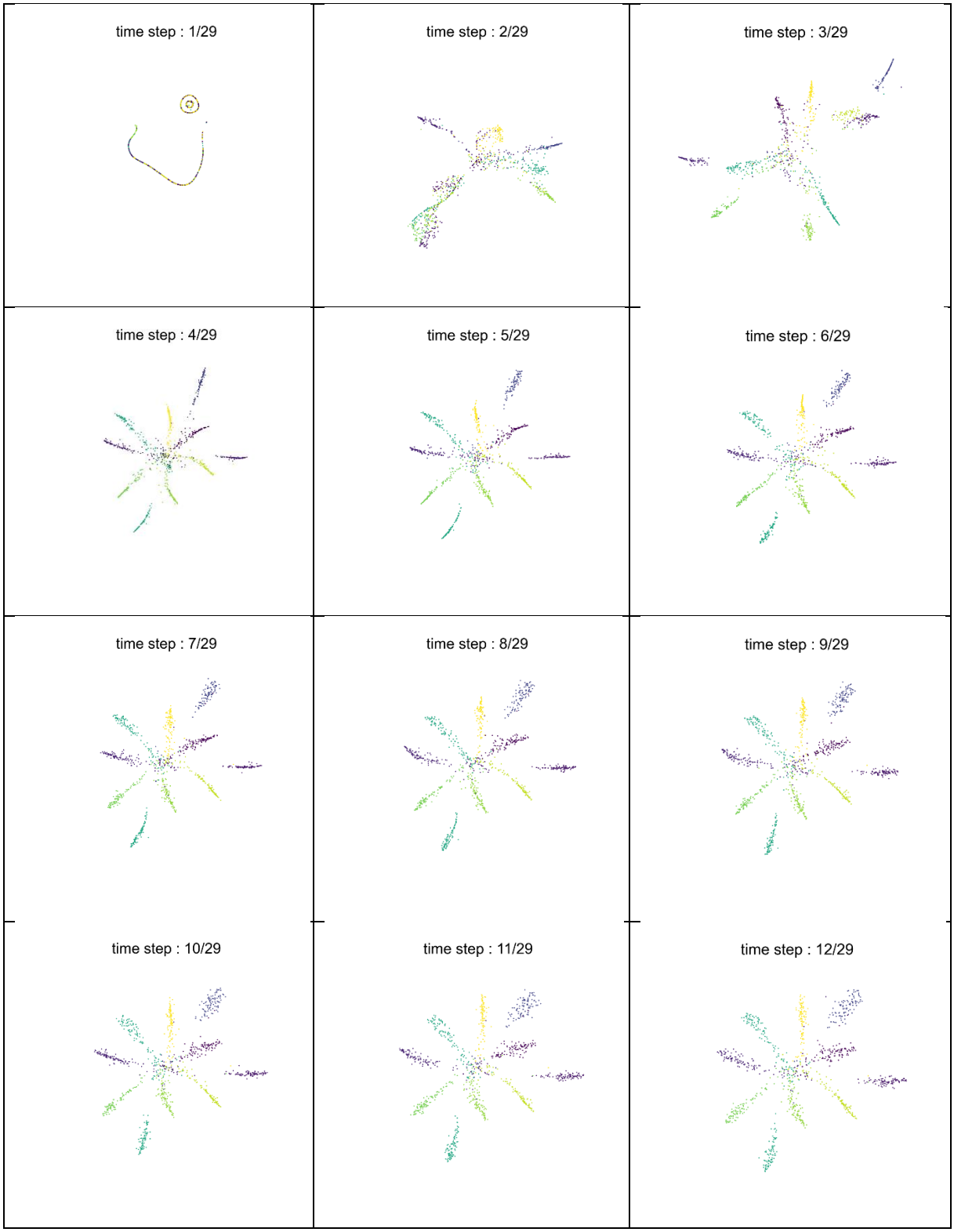


Figure 223 Animation d'une succession de projections générées par le TCP+ sur le dataset SVHN

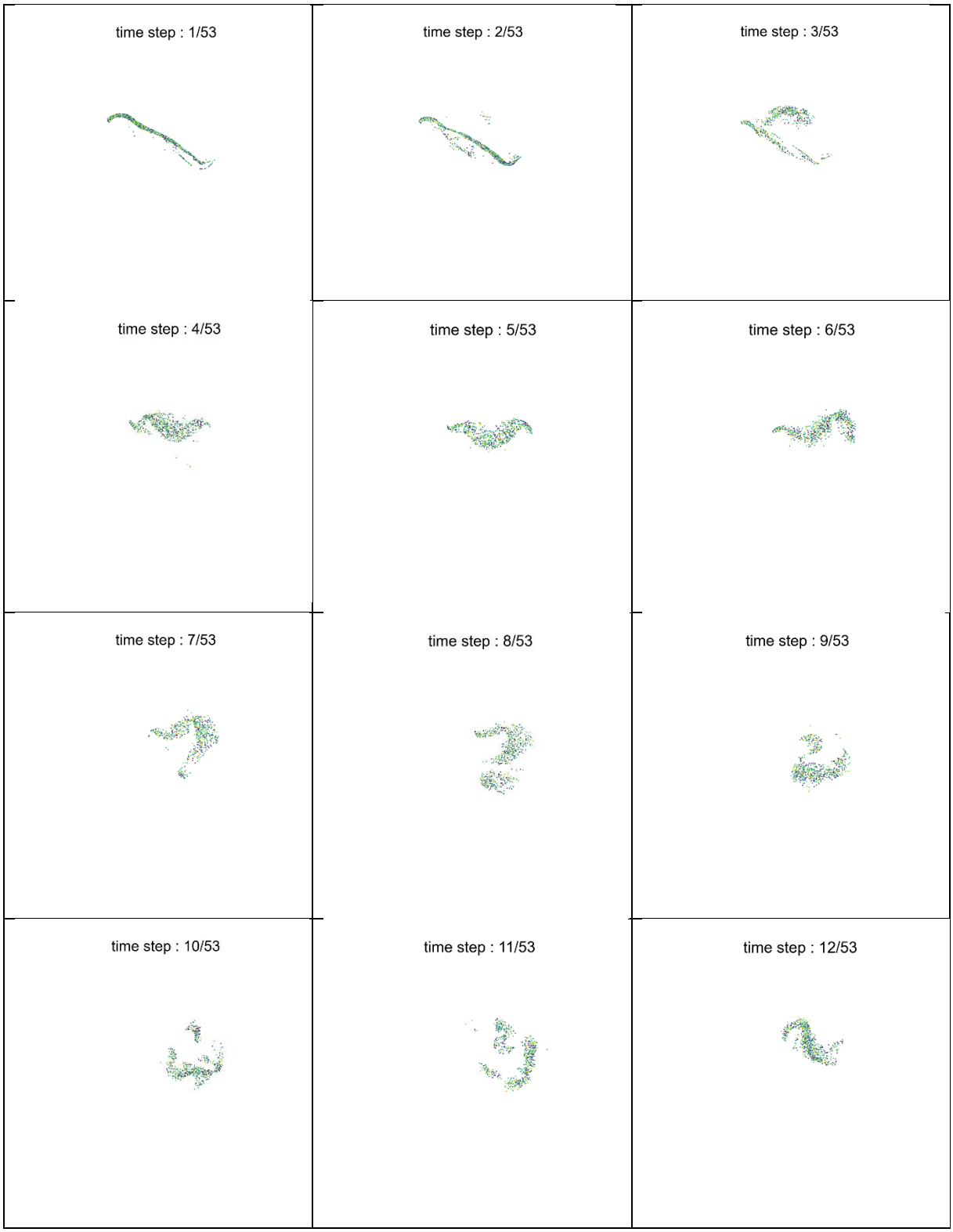


Figure 224 Animation d'une succession de projections générées par le TCP+ sur le dataset du Covid

9.2 Résultats des métriques

La comparaison des résultats s'est faite par rapport aux résultats du TCP qui a, pour l'instant, fourni la majorité des meilleurs résultats. Une hausse des performances est attendue de cette solution du côté du dataset SVHN.

9.2.1 Conservation de la dynamique

TCP	TCP+	Dynamique en haute dimension

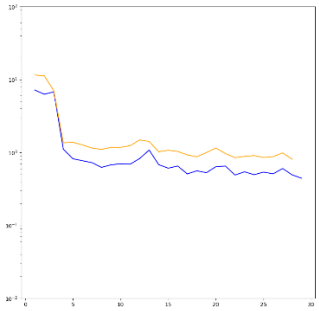
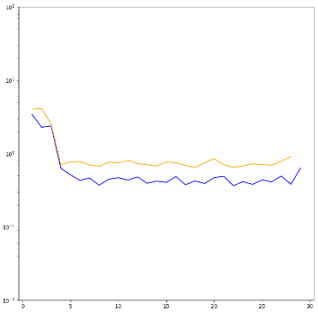
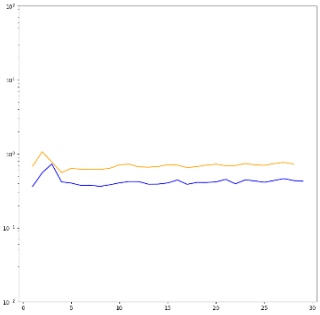

		

Figure 225 Dynamique des instances du TDP avec TCP sur le dataset SVHN	*Figure 226 Dynamique des instances du TDP avec TCP+ sur le dataset SVHN*	*Figure 227 Dynamique en haute dimension des instances du dataset SVHN*

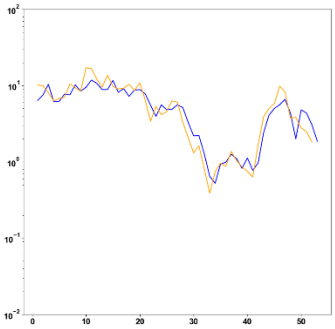
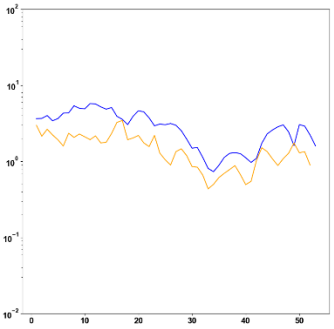
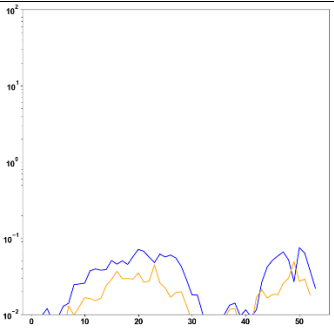

		

Figure 228 Dynamique des instances du TDP avec TCP sur le dataset Covid	*Figure 229 Dynamique des instances du TDP avec TCP+ sur le dataset Covid*	*Figure 230 Dynamique en haute dimension des instances du dataset Covid*

TCP+ semble avancer d'un pas supplémentaire vers la dynamique en haute dimension pour les deux datasets. Plus de précision était un des effets prévus lors de la réalisation de cette technique.

9.2.2 Conservation de la direction

TCP	TCP+	Dynamique en haute dimension

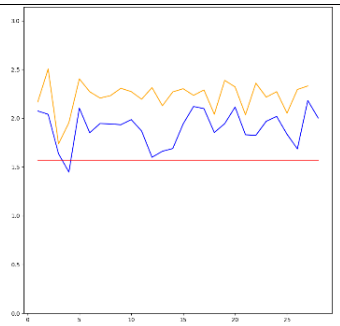
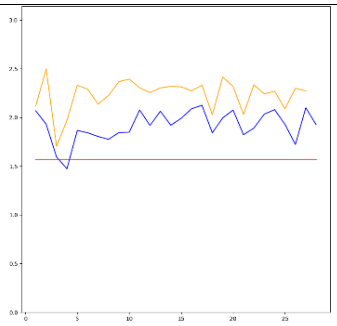
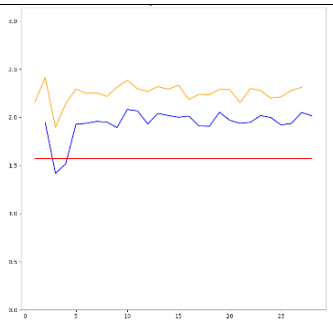

		

Figure 231 Changements de direction des instances du TDP avec TCP sur le dataset SVHN	*Figure 232 Changements de direction des instances du TDP avec TCP+ sur le dataset SVHN*	*Figure 233 Changements de direction des instances en haute dimension du dataset SVHN*

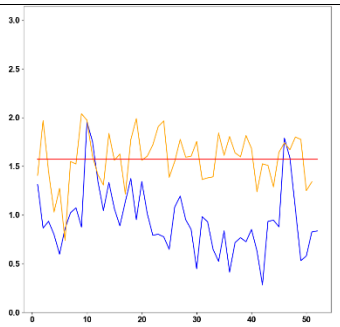
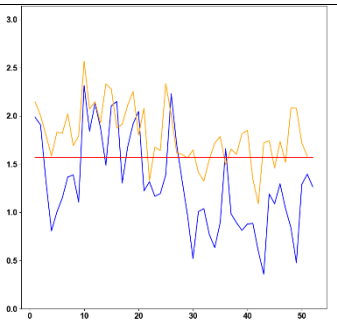
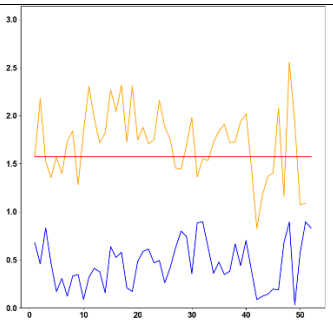

		

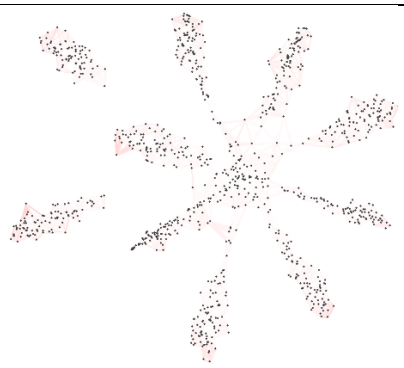
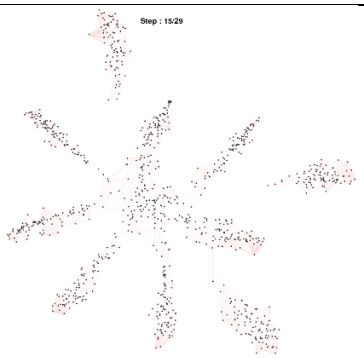
Figure 234 Changements de direction des instances du TDP avec TCP sur le dataset Covid	*Figure 235 Changements de direction des instances du TDP avec TCP+ sur le dataset Covid*	*Figure 236 Changements de direction des instances en haute dimension du dataset Covid*

Pour le SVHN, les changements de direction sont aussi bien conservés avec TCP qu'avec TCP+.

Le dataset du Covid voit les changements de direction de ses vecteurs de mouvement (courbe bleue) moins influencés par TCP+ qu'avec TCP. Ce comportement s'explique par le fait que le dernier time step ajouté à la fonction de coût a pour but de calmer les instances en les entraînant plus facilement vers leur état stable final (comme avec SVHN). Or, le dernier time step du Covid n'est pas un état final, ce qui résulte en une agitation supplémentaire des instances.

9.2.3 Reliability map

	TCP	TCP+

Dataset : SVHN	 *Figure 237 Reliability map des instances du TDP avec TCP sur le dataset SVHN d'un time step aléatoire*	 *Figure 238 Reliability map des instances du TDP avec TCP+ sur le dataset SVHN d'un time step aléatoire*
Dataset : Covid	 *Figure 239 Reliability map des instances du TDP avec TCP sur le dataset Covid d'un time step aléatoire*	 *Figure Reliability map des instances du TDP avec TCP+ sur le dataset Covid d'un time step aléatoire*

La reliability map du TCP+ sur SVHN est fortement similaire à la reliability map du TCP. Il est même compliqué de détecter que les deux visualisations viennent de deux techniques différentes. Cependant, à la vue des résultats sur le Covid, il semble que le TCP reste plus efficace que le TCP+. Les résultats du trustworthiness vont départager les similarités entre TCP et TCP+ pour le SVHN.

9.2.4 Trustworthiness

La comparaison entre trois techniques est présente dans les tableaux suivants, TCP, TCP+ et la version hybride. Une hypothèse selon laquelle la version hybride serait plus efficace que le TCP+ est à démontrer pour le dataset du Covid.

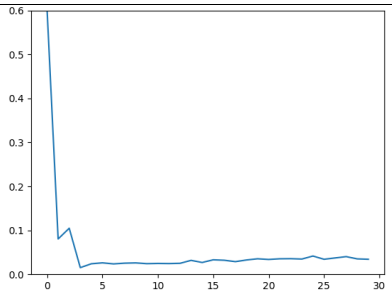
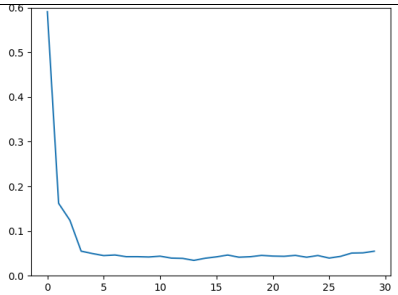
Dataset	SVHN	SVHN	SVHN
Mode de pénalisation	TCP	TCP+	Hybride λ = 0.05

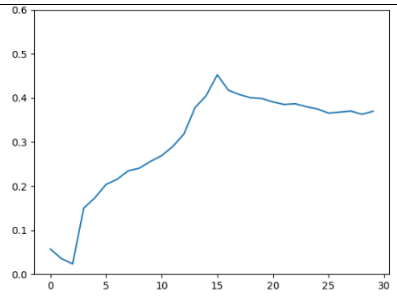
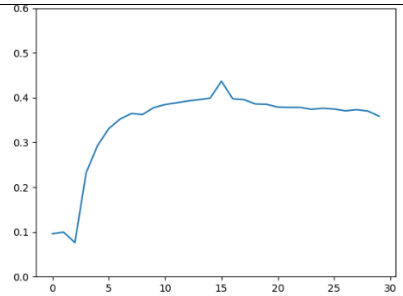
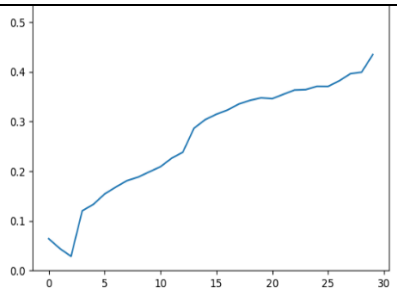
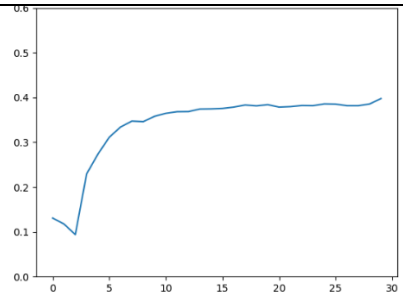
Time step t concerné / Temporalité T			
0/29	0.9438	0.9477	0.9351
7/29	0.8084	0.8145	0.8251
15/29	0.8046	0.801	0.7993
22/29	0.813	0.798	0.7910
29/29	0.8175	0.7899	0.7976
Moyenne sur tous les time steps :	0.8231	0.8169	0.8138

Dataset	Covid	Covid	Covid
Mode de pénalisation	TCP	TCP+	Hybride λ = 0.03
Time step t concerné / Temporalité T			
0/54	0.9945	0.9941	0.9942
13/54	0.9534	0.9538	0.9903
27/54	0.9732	0.9707	0.9799
40/54	0.9837	0.9838	0.9820
54/54	0.9751	0.9696	0.9783
Moyenne sur tous les time steps :	0.9720	0.9670	0.9821

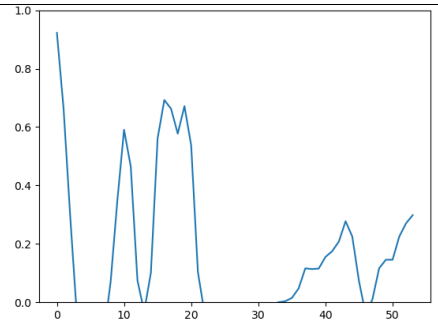
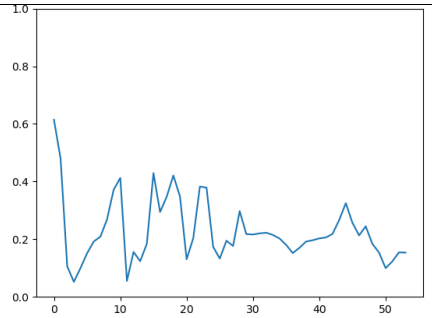
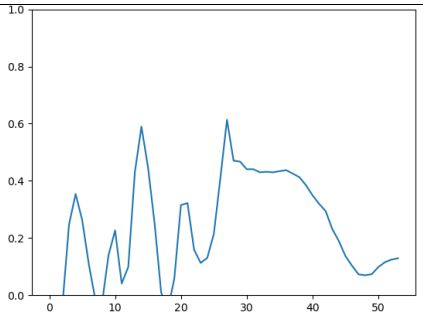
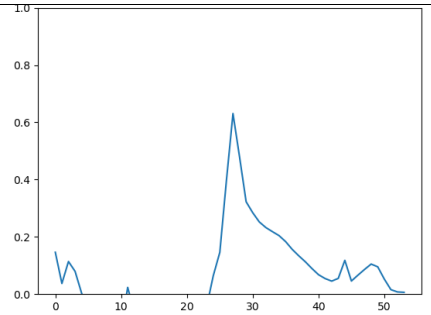
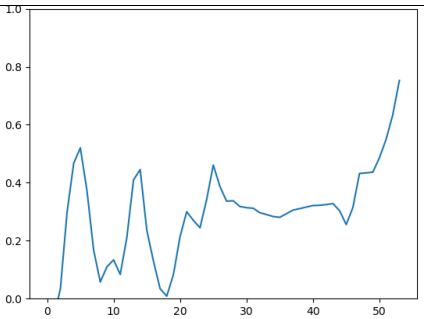
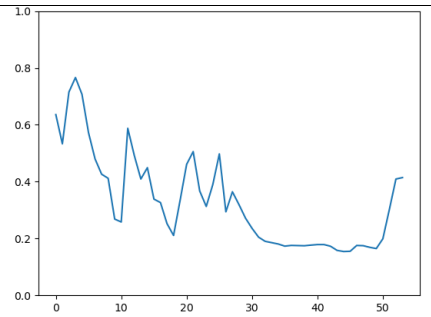
Une légère baisse en efficacité est enregistrée du côté du TCP+ en ce qui concerne la fiabilité de la projection du SVHN, cette différence est cependant négligeable à la vue du centième de différence séparant les deux moyennes. Le dataset du Covid, lui, a subi une pareille réduction de sa fiabilité avec TCP+. Cependant, la version hybride semble être une meilleure alternative pour ce dataset. L'analyse sur les similarités de voisinage devrait décider définitivement de la meilleure solution pour chaque dataset.

9.2.5 Temporal AUClogRNX

Dataset: SVHN Time step de référence	TCP	TCP+
t=0	 *Figure 240 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence 0/29*	 *Figure 241 Temporal AUClogRNX du TDP avec TCP+ sur le dataset SVHN avec pour time step de référence 0/29*

t = T/2	 *Figure 242 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence 15/29*	 *Figure 243 Temporal AUClogRNX du TDP avec TCP+ sur le dataset SVHN avec pour time step de référence 15/29*
t=T	 *Figure 244 Temporal AUClogRNX du TDP avec TCP sur le dataset SVHN avec pour time step de référence 29/29*	 *Figure 245 Temporal AUClogRNX du TDP avec TCP+ sur le dataset SVHN avec pour time step de référence 29/29*

Pour le SVHN, les résultats voulus ont été obtenus. Pour rappel, la Figure 221 et la Figure 222 représentent l'objectif de voisinage pour le dataset. Et le TCP+ s'approche fortement de ces estimations. TCP présente certains problèmes pour maintenir la similarité des voisinages sur une longue durée, en témoigne la Figure 244 où la courbe au début de temporalité n'a pas assez de similarité avec les derniers time steps. La Figure 245, grâce au système d'ancrage du TCP+, permet à tous les time steps aux voisinages similaires d'être poussés vers le haut et de tendre vers la valeur maximale de similarité.

Dataset: Covid Time step de référence	TCP	TCP+
t=0	 *Figure 246 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence 0/54*	 *Figure 247 Temporal AUClogRNX du TDP avec TCP+ sur le dataset Covid avec pour time step de référence 0/54*
t = T/2	 *Figure 248 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence 27/54*	 *Figure 249 Temporal AUClogRNX du TDP avec TCP+ sur le dataset Covid avec pour time step de référence 27/54*
t=T	 *Figure 250 Temporal AUClogRNX du TDP avec TCP sur le dataset Covid avec pour time step de référence 54/54*	 *Figure 251 Temporal AUClogRNX du TDP avec TCP+ sur le dataset Covid avec pour time step de référence 54/54*

Le comportement du temporal AUClogRNX du TCP+ sur la projection du dataset du Covid est bien différent du comportement sur la projection du TCP. Le TCP+ n'est pas capable d'améliorer les performances du temporal AUClogRNX sur le dataset du Covid. Il est complexe d'interpréter le comportement de la métrique sur ce dataset, mais il n'est pas nécessaire de se lancer dans cet exercice car ce n'est pas initialement pertinent d'appliquer cette solution sur ce genre de dataset dû à la nature non convergente du dataset (ou en tout cas, non suffisamment convergente par manque de données). Les résultats restent également très loin des résultats fournis par la solution hybride.

9.3 Discussion des résultats

La création du TCP+ a pour unique objectif d'améliorer la performance de la métrique temporal AUClogRNX sur des datasets comme le SVHN. Cette spécificité rend la solution très peu généralisée sur l'ensemble des datasets temporels pouvant exister. La solution présente de moins bons résultats que le TCP classique ou la version hybride lorsque le dataset n'est pas convergent.

TCP+ est très performant pour projeter les instances en respectant la convergence des voisinages du dataset sur lequel il travaille. Cet avantage permet également d'obtenir des visualisations très fiables mais pas tout à fait au niveau du TCP classique qui, ayant moins de contraintes, a moins de difficulté pour projeter ses instances. Cette différence de fiabilité n'est toutefois pas à prendre en compte, les deux solutions étant considérées comme équivalentes d'un point de vue de la fiabilité.

La conservation de la dynamique et des directions est efficace sur les deux datasets. Une réduction de la dynamique est observée mais la forme de la courbe reste la même et est similaire à la dynamique en haute dimension. Les changements de directions respectent les changements en haute dimension uniquement si le dataset est convergent, sinon il va être biaisé par l'état des instances du dernier time step qui n'ont pas encore, ou ne vont jamais, converger.

10 Vérifications des hypothèses sur les datasets « blob »

De nombreuses expériences ont été réalisées sur les datasets du Covid et du SVHN, cependant ces dernières ont un défaut. Il est très compliqué de vérifier certaines des conclusions à cause d'un manque d'informations sur la réelle structure des données. Le besoin de réaliser des expériences sur des données connues de toutes parts, c'est-à-dire créées de toutes pièces et conçues pour mettre en évidence certains comportements sur la dynamique des instances doit s'effectuer.

Ce chapitre abordera certains des résultats des métriques des expériences des 5 solutions décrites précédemment, mais en utilisant les datasets « blob » pour certains cas de figures où des doutes sont présents. Tout cela afin de déceler les similarités et différences entre les solutions et pour s'assurer de ne pas avoir fait fausse route sur certaines conclusions. L'objectif des sections suivantes est de pouvoir dresser un tableau de compétences mettant en relation toutes les solutions, les comparant par rapport à leurs résultats sur les différentes métriques.

Toutes les hésitations pour les interprétations tournent autour de la métrique temporal AUClogRNX. Les configurations de chaque solution pour chaque dataset sont déterminées selon les paramètres (perplexité et le λ) maximisant le trustworthiness. Les graphiques sur la dynamique et les changements de direction ont fourni des conclusions similaires à celles prises lors des expérimentations.

10.1 Choix des paramètres

Ci-dessous sont repris les scores de trustworthiness de chaque solution pour chaque dataset « blob ». Ces paramètres sont considérés comme les plus optimaux pour cette métrique.

Blob_MRU	Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	TCP	Hybride	TCP+
λ ->	0	0.04	0.01	Pas de λ	0.01	Pas de λ
Time step						
0/29	0.9816	0.9806	0.9934	0.9916	0.9911	0.9821
7/29	0.9380	0.9626	0.9749	0.9737	0.9746	0.9582
15/29	0.9511	0.9612	0.9665	0.9738	0.9724	0.9431
22/29	0.9042	0.9579	0.9741	0.9711	0.9696	0.9620
29/29	0.9829	0.9827	0.9214	0.9908	0.9910	0.9472
Moyenne sur tous les time steps :	0.9426	0.9624	0.9728	0.9760	0.9767	0.9633

Blob_MRUA	Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	TCP	Hybride	TCP+
λ ->	0	0.05	0.02	Pas de λ	0.01	Pas de λ
Time step						
0/29	0.9734	0.9936	0.9929	0.9935	0.9932	0.9749
7/29	0.9715	0.9915	0.9913	0.9911	0.9910	0.9591
15/29	0.9722	0.9918	0.9916	0.9914	0.9914	0.9747
22/29	0.9454	0.9861	0.9851	0.9850	0.9834	0.9705
29/29	0.9702	0.9904	0.9903	0.9860	0.9876	0.9555
Moyenne sur tous les time steps :	0.9683	0.9907	0.9904	0.9897	0.9896	0.9710

Blob_MCu	Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	TCP	Hybride	TCP+
λ ->	0	0.05	0.02	Pas de λ	0.02	Pas de λ
Time step						
0/29	0.8639	0.9934	0.9933	0.9923	0.9923	0.9630
7/29	0.8878	0.9916	0.9918	0.9900	0.9913	0.9777
15/29	0.8658	0.9931	0.9930	0.9903	0.9890	0.9705
22/29	0.9181	0.9926	0.9926	0.9904	0.9904	0.9738
29/29	0.9058	0.9930	0.9933	0.9911	0.9912	0.9477
Moyenne sur tous les time steps :	0.8854	0.9929	0.9929	0.9909	0.9912	0.9722

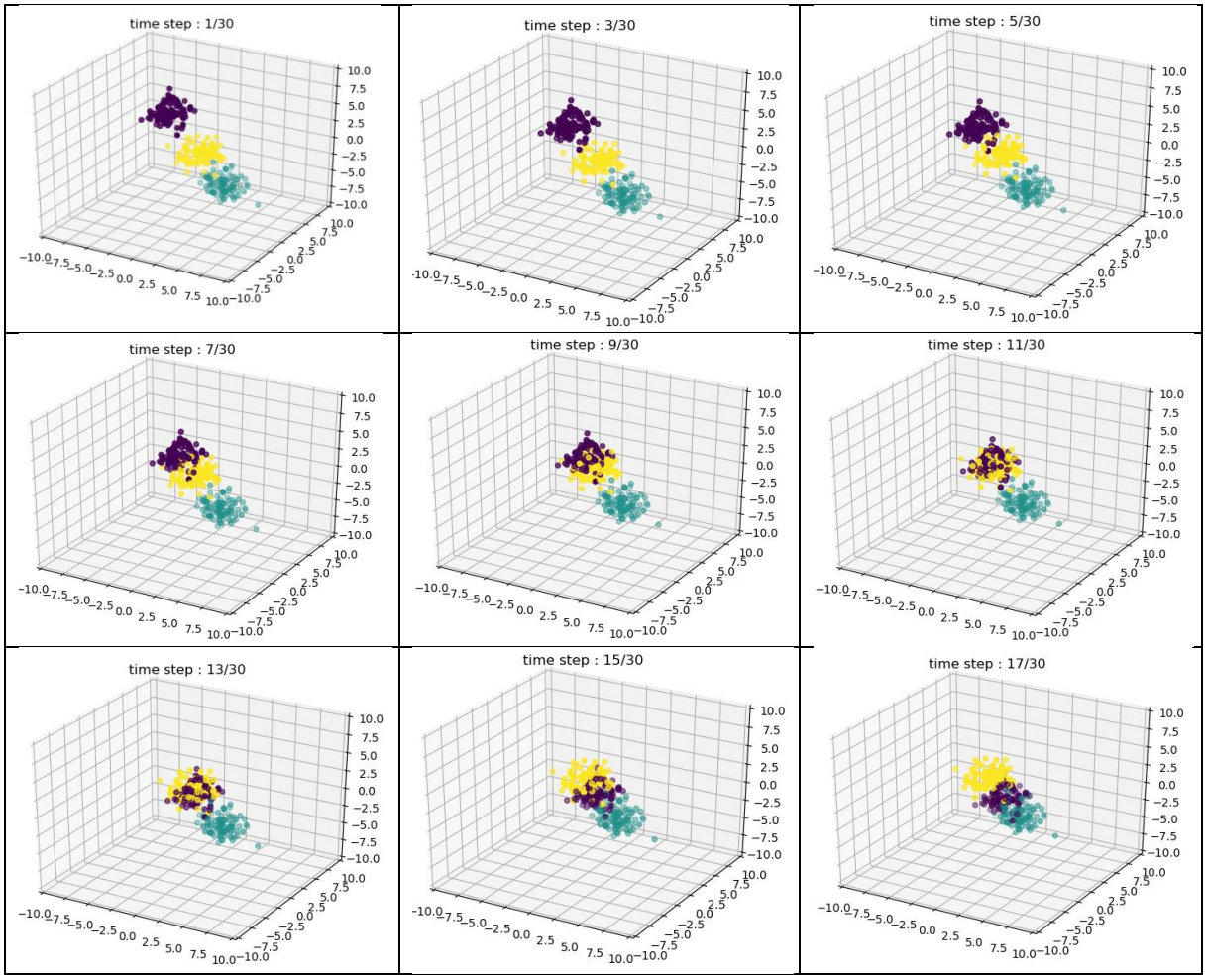
Blob_Rand	Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	TCP	Hybride	TCP+
λ ->	0	0.05	0.02	Pas de λ	0.02	Pas de λ
Time step						
0/29	0.7959	0.9817	0.9829	0.9825	0.9821	0.9706
7/29	0.8589	0.9934	0.9931	0.9929	0.9923	0.9740
15/29	0.8203	0.9919	0.9921	0.9906	0.9908	0.9751
22/29	0.8909	0.9933	0.9935	0.9927	0.9925	0.9740
29/29	0.8749	0.9933	0.9933	0.9926	0.9930	0.9453
Moyenne sur tous les time steps :	0.8829	0.9923	0.9924	0.9916	0.9913	0.9701

10.2 Expériences

Maintenant que les solutions sont paramétrées, l'analyse des résultats de la métrique temporal AUClogRNX est la prochaine étape. Chaque tableau présent dans les sections relatives à chaque dataset « blob » reprend les résultats de la métrique temporal AUClogRNX sur chaque solution au time step de référence centrale à la temporalité.

10.2.1 Analyse du dataset Blob_MRU

Le tableau ci-dessous présente 9 time steps représentant la dynamique des trois nuages du dataset Blob_MRU en haute dimension.



Pour rappel, le dataset Blob_MRU est composé de 3 nuages de points se déplaçant en mouvement rectiligne uniforme sur une temporalité de 30 time steps.

Concernant les résultats sur le dataset Blob_MRU, TCP+ a enregistré des résultats intéressants. Dans les faits, c'est la seule solution à avoir retranscrit au mieux la dynamique des clusters et les modifications des voisinages. Pour le démontrer, il suffit d'analyser la dynamique du Blob_MRU en haute dimension, ce qui est facile car les blobs sont construits en trois dimensions. Une vidéo a donc été réalisée représentant le déplacement des nuages de points. Dans cette vidéo, deux collisions entre deux paires de clusters se produisent. La première entre le cluster A et le cluster B sur une durée assez longue (du time step 7 au 15). Et une deuxième entre le cluster B et le cluster C (du time step 16 au 25).

Le Blob_MRU intervertit les positions de ses clusters, le cluster A prend la place du cluster B, le cluster B prend la place du cluster C et le cluster C prend la place du cluster A. Cette dynamique très singulière est bien représentée sur le TDP du TCP+ en témoignent les figures suivantes.

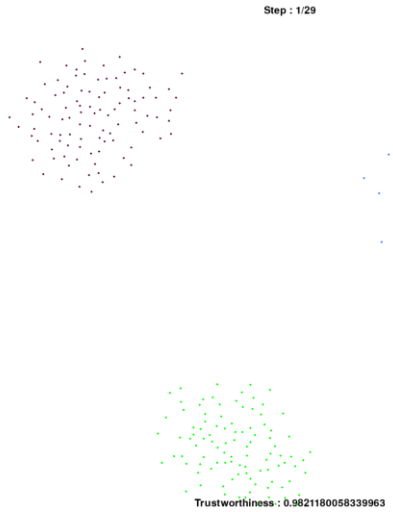


Figure 252 TDP des instances du dataset Blob_MRU avec TCP+ au time step 1/29

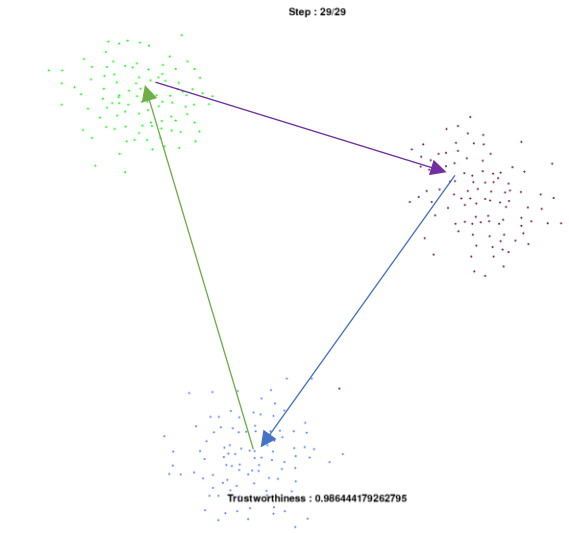


Figure 253 Résumé du mouvement complet des clusters du dataset Blob_MRU durant la temporalité du TDP avec TCP+ au time step 29/29

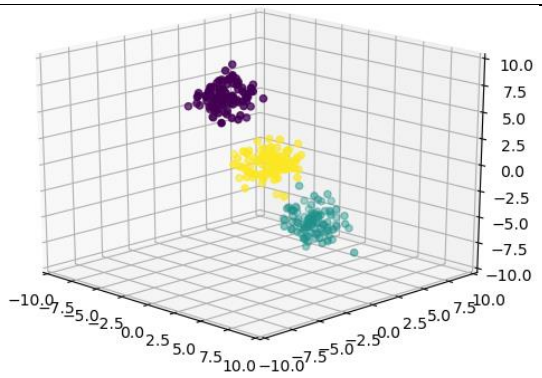


Figure 254 État initial des clusters du dataset Blob_MRU en haute dimension

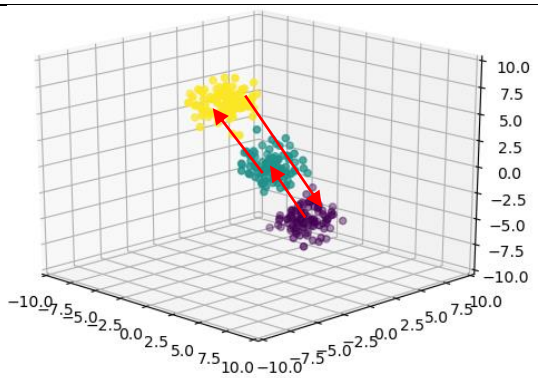


Figure 255 Résumé du mouvement complet des clusters en haute dimension du dataset Blob_MRU durant la temporalité

La dynamique est respectée et la présence des collisions également, en témoignent la Figure 256 et la Figure 257 suivantes comparant la projection de TCP+ et le time step équivalent en haute dimension. Ce comportement est très intéressant, TCP+ est capable de projeter des mouvements simples mais avec une très grande précision.

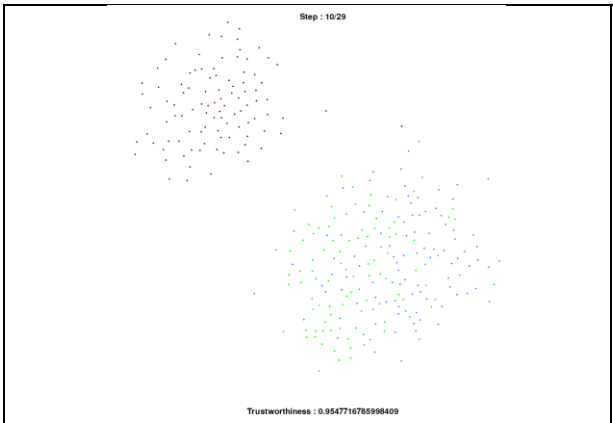


Figure 256 Première collision entre deux clusters du dataset Blob_MRU projeté avec TCP+ au time step 10/29

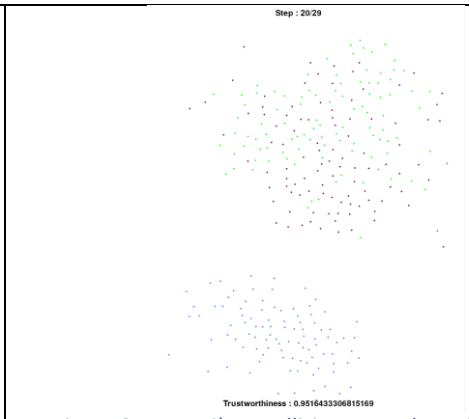


Figure 257 Deuxième collision entre deux clusters du dataset Blob_MRU projeté avec TCP+ au time step 20/29

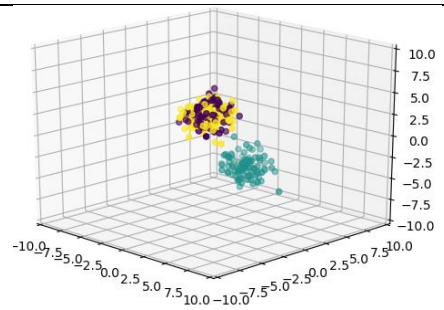


Figure 258 Collision entre deux clusters du dataset Blob_MRU en haute dimension au time step 10/29

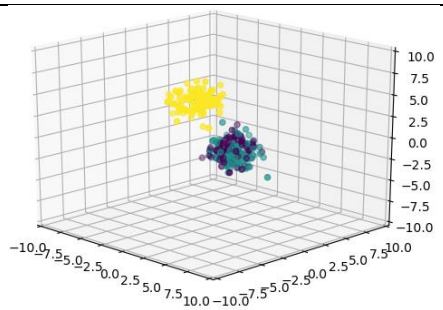


Figure 259 Collision entre deux clusters du dataset Blob_MRU en haute dimension au time step 20/29

Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations

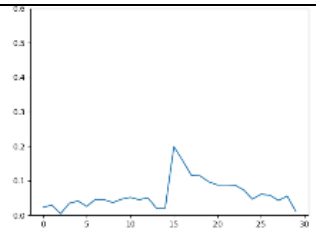
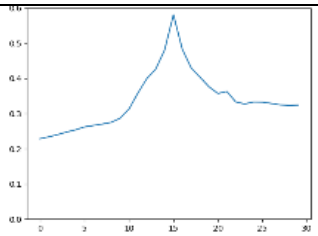
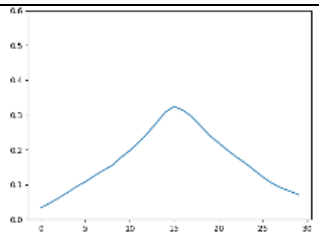

		

Figure 260 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Blob_MRU sur le time step de référence 15/29	*Figure 261 Temporal AUClogRNX du TDP avec pénalisation des mouvements sur le dataset Blob_MRU sur le time step de référence 15/29*	*Figure 262 Temporal AUClogRNX du TDP avec pénalisation des accélérations sur le dataset Blob_MRU sur le time step de référence 15/29*
TCP	**Hybride**	**TCP+**

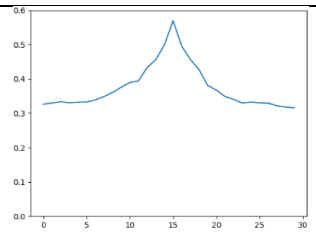
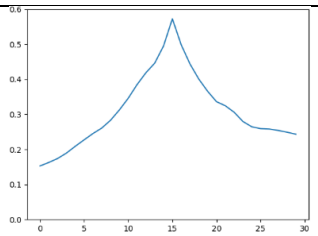
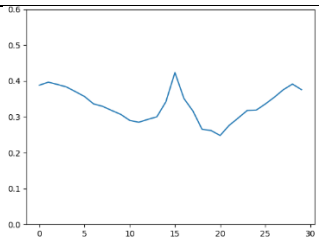
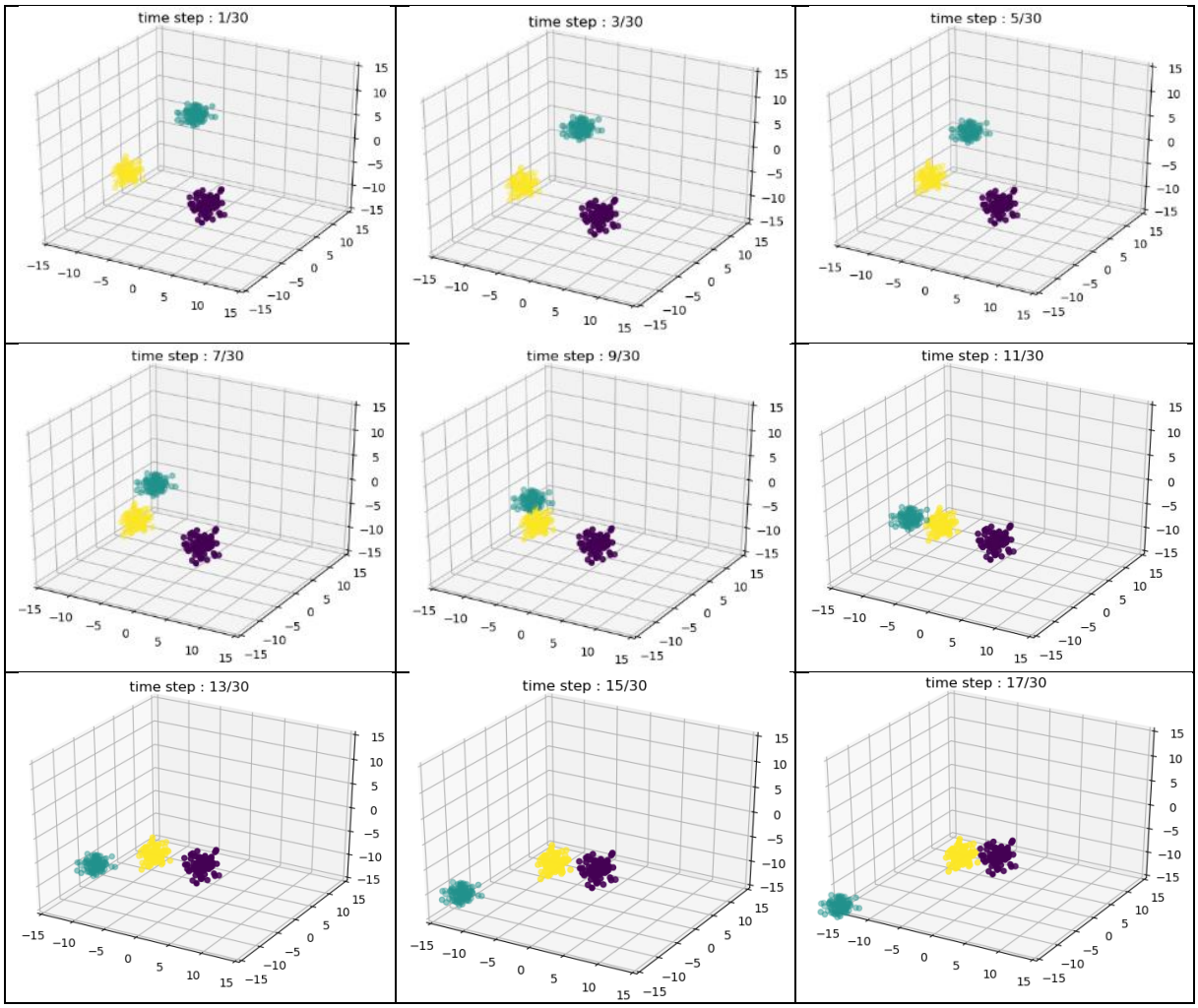

		

Figure 263 Temporal AUClogRNX du TDP avec TCP sur le dataset Blob_MRU sur le time step de référence 15/29	*Figure 264 Temporal AUClogRNX du TDP avec la version hybride sur le dataset Blob_MRU sur le time step de référence 15/29*	*Figure 265 Temporal AUClogRNX du TDP avec TCP+ sur le dataset Blob_MRU sur le time step de référence 15/29*

De plus, à la vue du graphique de temporal AUClogRNX en Figure 265 TCP+ est la solution la plus efficace comparé à toutes les autres présentes dans ce document. Le time step 15 étant le time step entre les deux collisions, les voisinages à ce moment précis sont très similaires aux moments avant et après les collisions ce qui explique la forme en « w » du graphique généré. Et aucune autre solution n'a réussi à égaler ces résultats.

10.2.2 Analyse du dataset Blob_MRUA

Le tableau ci- présente 9 time steps représentant la dynamique des trois nuages du dataset Blob_MRUA.



En ce qui concerne le dataset Blob_MRUA, son comportement est différent de Blob_MRU, chaque cluster part dans la direction de la position initiale d'un autre cluster mais en accélérant, pouvant ainsi dépasser le point de destination prévu. Du time step 5 au time step 10, un frottement entre le cluster A et le cluster B est enregistré faisant varier les voisinages. Ensuite, du time step 15 au 26, le cluster A entre en légère collision avec le cluster C, environs 10% du volume de chacun des deux clusters se mélangent avec l'autre. Dès le time step 11, le cluster B reste isolé et très éloigné des deux autres. Le comportement attendu de la courbe « parfaite » du voisinage est plus complexe à envisager qu'avec Blob_MRU, mais il est intéressant d'observer la courbe générée par le temporal AUClogRNX sur Blob_MRUA avec TCP+ utilisant comme time step de référence le premier de la temporalité.

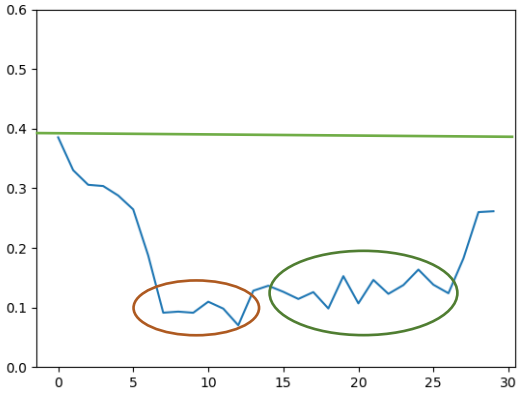
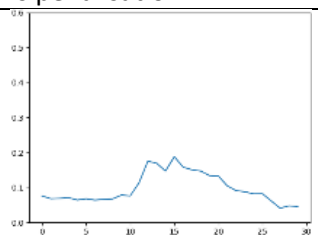
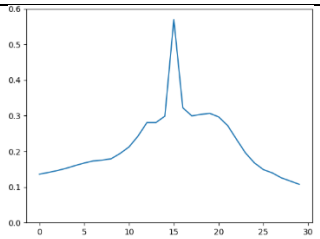
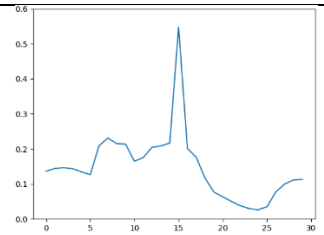
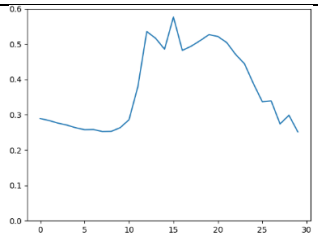
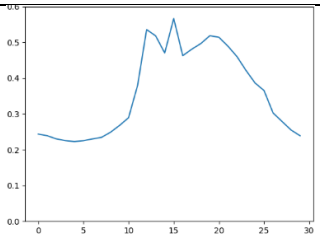
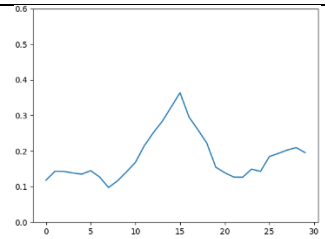


Figure 266 Détection des deux collisions inter-cluster du dataset Blob_MRUA à l'aide du temporal AUClogRNX avec pour time step de référence : 0/29 (TCP+)

Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations
 *Figure 267 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Blob_MRUA sur le time step de référence 15/29*	 *Figure 268 Temporal AUClogRNX du TDP avec pénalisation des mouvements sur le dataset Blob_MRUA sur le time step de référence 15/29*	 *Figure 269 Temporal AUClogRNX du TDP avec pénalisation des accélérations sur le dataset Blob_MRUA sur le time step de référence 15/29*
TCP	**Hybride**	**TCP+**
 *Figure 270 Temporal AUClogRNX du TDP avec TCP sur le dataset Blob_MRUA sur le time step de référence 15/29*	 *Figure 271 Temporal AUClogRNX du TDP avec la version hybride sur le dataset Blob_MRUA sur le time step de référence 15/29*	 *Figure 272 Temporal AUClogRNX du TDP avec TCP+ sur le dataset Blob_MRUA sur le time step de référence 15/29*

En Figure 266, le point de référence est positionné sur un time step où chaque cluster est clairement séparé. L'encadré orange représente la première chute de similarité de voisinage se produisant lorsque la première collision entre clusters arrive. L'encadré vert indique la chute de similarité de la deuxième collision. Les deux derniers time steps voient leur voisinage augmenter car les clusters ont fini de se mélanger et sont à nouveau distincts. La droite horizontale verte met en évidence la remontée de voisinage qui, si le dataset était composé de quelques time steps supplémentaires, aurait plafonné à la hauteur du time step de référence.

TCP+ est la solution à l'origine de ce graphique, le défaut des autres solutions provient de leur difficulté à refaire le lien entre le premier et le dernier time step, résultant en une chute du voisinage suivant les deux collisions ne remontant plus ensuite. Les quelques figures suivantes témoignent des difficultés de certaines solutions à représenter la situation.

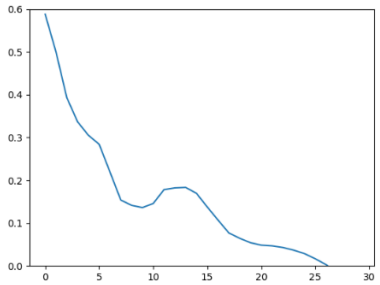


Figure 273 Temporal AUClogRNX des instances du dataset Blob_MRUA projetées par pénalisation des mouvements avec pour time step de référence : 0/29

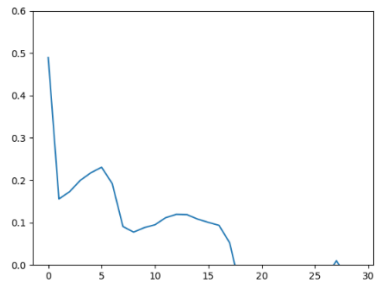


Figure 274 Temporal AUClogRNX des instances du dataset Blob_MRUA projetées par pénalisation des accélérations avec pour time step de référence : 0/29

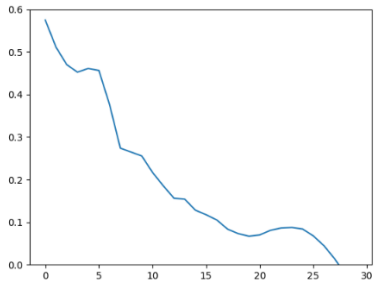
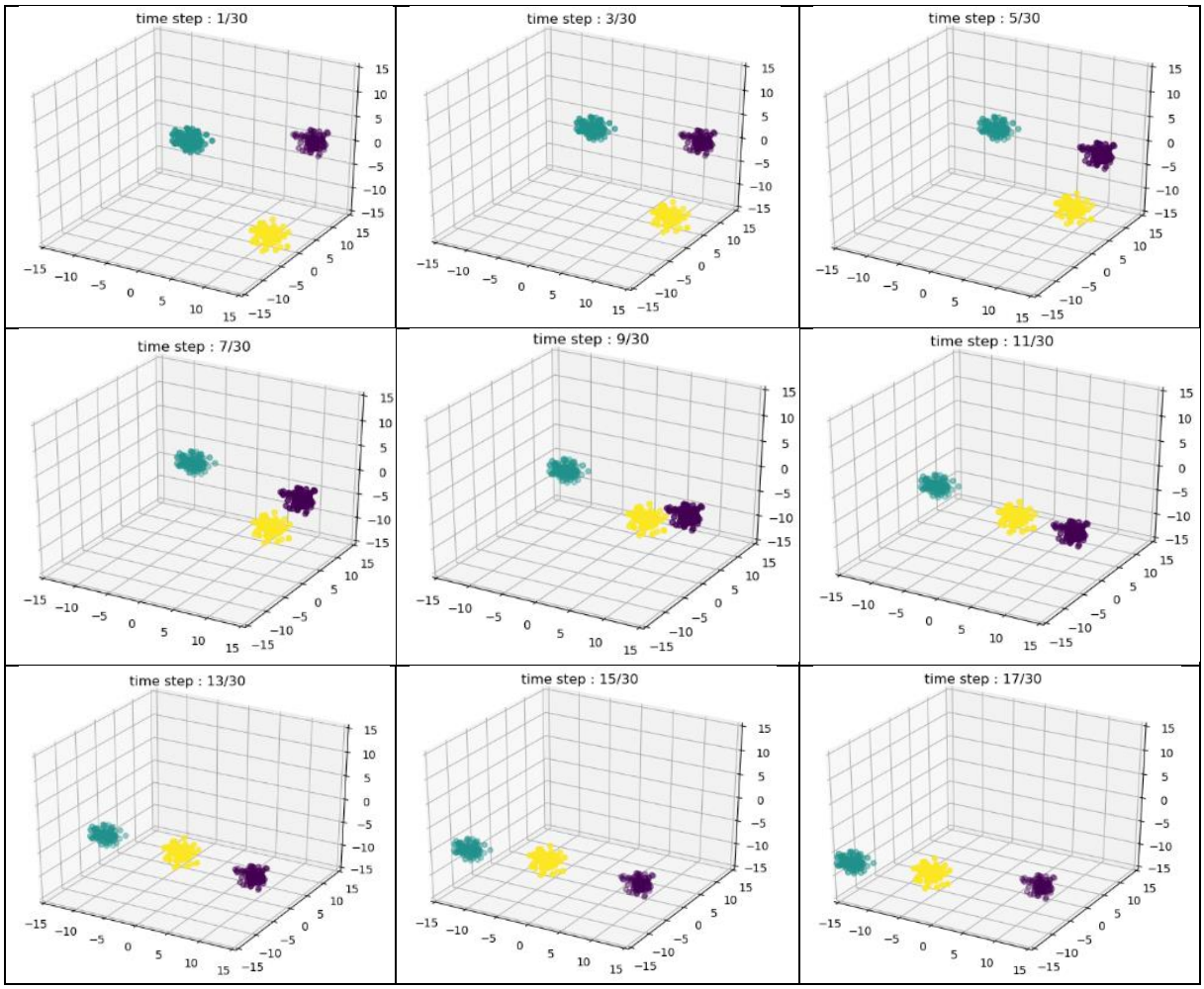


Figure 275 Temporal AUClogRNX des instances du dataset Blob_MRUA projetées par TCP avec pour time step de référence : 0/29

10.2.3 Analyse du dataset Blob_MCu

Le tableau ci-dessous présente 9 time steps de la dynamique des trois nuages du dataset Blob_Mcu.



Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations

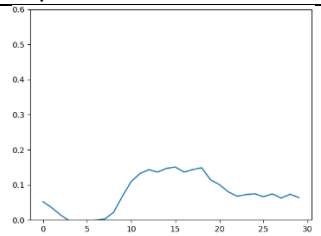
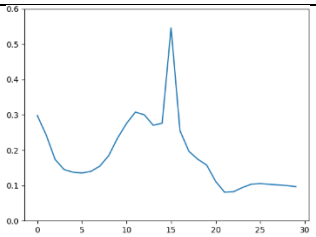
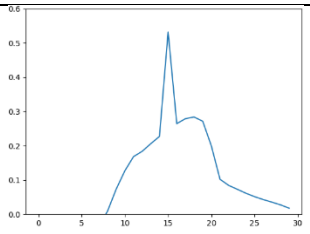

		

Figure 276 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Blob_MCu sur le time step de référence 15/29	*Figure 277 Temporal AUClogRNX du TDP avec pénalisation des mouvements sur le dataset Blob_MCu sur le time step de référence 15/29*	*Figure 278 Temporal AUClogRNX du TDP avec pénalisation des accélérations sur le dataset Blob_MCu sur le time step de référence 15/29*
TCP	**Hybride**	**TCP+**

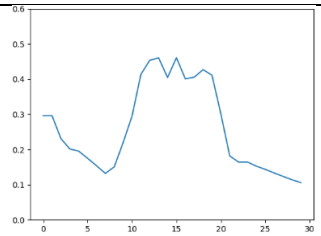
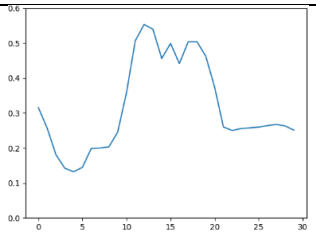
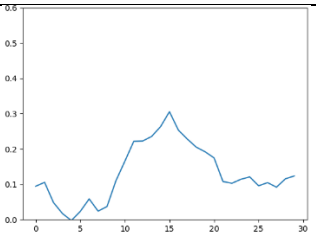
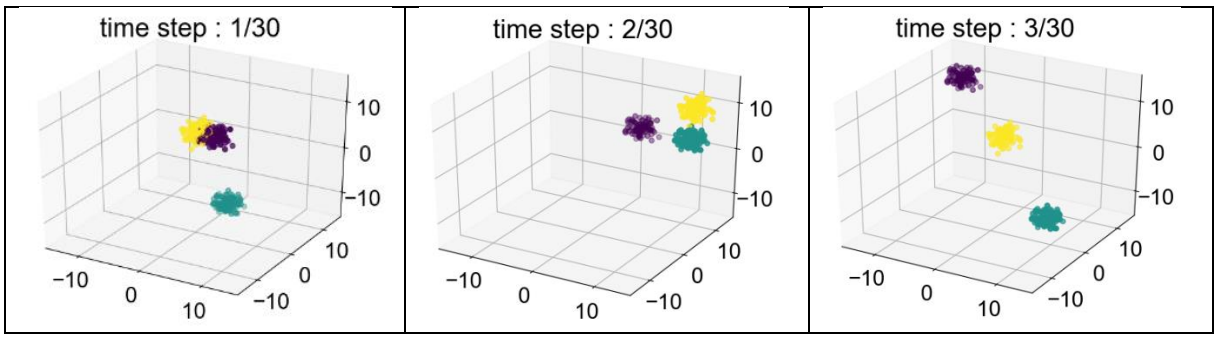

		

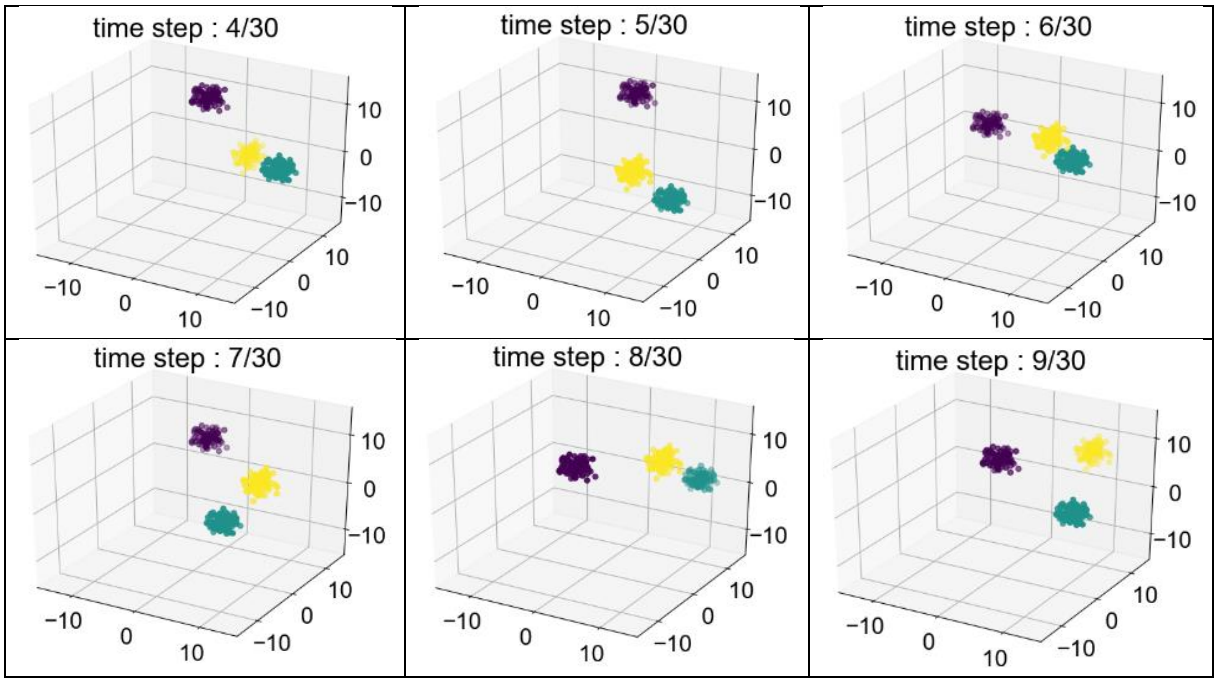
Figure 279 Temporal AUClogRNX du TDP avec TCP sur le dataset Blob_MCu sur le time step de référence 15/29	*Figure 280 Temporal AUClogRNX du TDP avec la version hybride sur le dataset Blob_MCu sur le time step de référence 15/29*	*Figure 281 Temporal AUClogRNX du TDP avec TCP+ sur le dataset Blob_MCu sur le time step de référence 15/29*

Le Blob_MCu, malgré les mouvements circulaires des clusters, n'enregistre qu'un léger frôlement de deux clusters dans le deuxième quart de la temporalité. Ce peu d'événements induit que la majorité des techniques sont capables de retransmettre l'évolution des voisinages sans trop de problème. L'élément principal qui doit être présent sur le temporal AUClogRNX est la chute de similarité aux alentours du time step 8. Cette chute représente le moment où deux clusters se sont suffisamment rapprochés pour altérer suffisamment le voisinage de leurs instances. Les solutions issues de TCP ont plus de facilités pour représenter cette chute. Pour toutes les solutions, une chute de voisinage apparaît à la fin de la temporalité pour une raison qui reste encore inconnue.

10.2.4 Analyse du dataset Blob_Rand

Le tableau ci-dessous présente les 9 premiers time steps du déplacement des trois nuages du dataset Blob_Rand.





Sans pénalisation	Pénalisation des mouvements	Pénalisation des accélérations

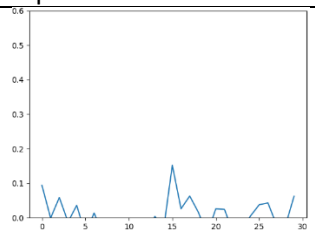
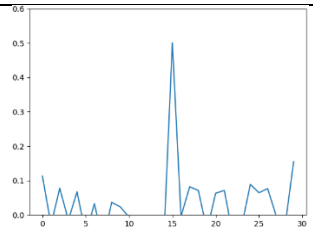
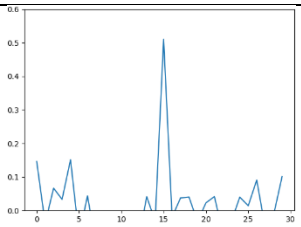

		

Figure 282 Temporal AUClogRNX du TDP sans pénalisation sur le dataset Blob_Rand sur le time step de référence 15/29	*Figure 283 Temporal AUClogRNX du TDP avec pénalisation des mouvements sur le dataset Blob_Rand sur le time step de référence 15/29*	*Figure 284 Temporal AUClogRNX du TDP avec pénalisation des accélérations sur le dataset Blob_Rand sur le time step de référence 15/29*
TCP	**Hybride**	**TCP+**

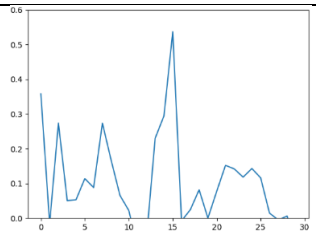
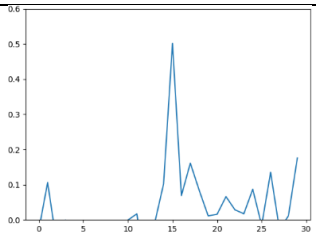
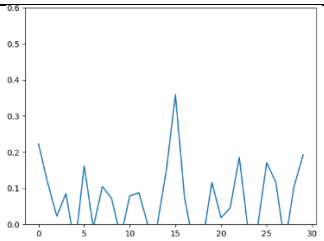

		

Figure 285 Temporal AUClogRNX du TDP avec TCP sur le dataset Blob_Rand sur le time step de référence 15/29	*Figure 286 Temporal AUClogRNX du TDP avec la version hybride sur le dataset Blob_Rand sur le time step de référence 15/29*	*Figure 287 Temporal AUClogRNX du TDP avec TCP+ sur le dataset Blob_Rand sur le time step de référence 15/29*

Le dernier dataset est particulier ; les mouvements des clusters sont aléatoires de time step en time step, mais les positions relatives entre chaque instance d'un même cluster restent identiques. Ce comportement induit que la similarité de voisinage reste très élevée lorsqu'aucune collision n'est détectée. Deux time steps ne percevant aucune collision de cluster ont le même voisinage.

Ce qui est donc attendu des expériences des différentes solutions, c'est une visualisation qui représente bien l'agitation des instances tout en conservant les voisinages des time steps n'étant pas dans la situation où une collision entre clusters intervient. En regardant les positions des clusters en haute dimension, les time steps sans collision sont les suivants : 2, 4, 6,8-11, 14, 17, 20, 22, 24, 26, 27 et le 29. En théorie, positionner le time step de référence sur un des time steps listés devrait générer un graphique temporal AUClogRNX possédant un pic de similarité à chacun des autres time steps de la liste. Malheureusement, aucune solution n'a réussi à être aussi précises que la théorie, certaines similitudes apparaissent, des pics de différentes intensités sont perçus parmi certaines solutions. Le tableau ci-dessous reprend chaque solution et présente une croix à chaque fois qu'un pic est détecté sur le temporal AUClogRNX au time step de la colonne.

Time step ->	4	6	8-11	14	17	20	22	24	26-27	29	Total
Pas de pénalisation		x	x					x			3/10
Pénalisation des mouvements	x	x	x			x				x	5/10
Pénalisation des accélérations	x	x	x		x		x	x	x	x	8/10
TCP		x	x		x	x	x	x	x		7/10
Hybride		x	x			x		x	x		5/10
TCP+		x	x		x	x		x	x	x	7/10

Les résultats récoltés sont étonnants, la solution pénalisant les accélérations se trouve être la solution performant au mieux sur ce dataset. En effet, les observations sur le graphique du temporal AUClogRNX montrent de grands pics de similarité aux bons endroits, TCP et TCP+ ont également des résultats similaires mais ces solutions ne sont pas aussi déterminées dans leur graphique affichant des pics plus faibles.

Partir du principe que la pénalisation des accélérations est la meilleure technique pour ce dataset est une erreur car au visionnage de l'animation, le problème inhérent à la solution apparaît : le lissage des perturbations. On observe que les clusters se déplacent de façon fluide durant la temporalité, ce qui ne reflète absolument pas la réalité. TCP et TCP+, cependant, ne sont pas du tout atteintes par ce phénomène (par la non-présence du facteur de pénalisation) et présentent des instances aux mouvements complètement imprédictibles, ce qui est attendu d'une technique de projection agissant sur Blob_Rand.

Une interrogation s'est présentée : pourquoi TCP+ fournit-il des résultats aussi mauvais lors des expériences sur le dataset du Covid alors qu'avec Blob_Rand, tout semble bien se dérouler ? La réponse provient du fait que même si le dataset Blob_Rand est considéré comme chaotique, il n'en reste pas moins ordonné. Toute instance du dataset, lorsqu'aucune collision n'est détectée, conserve toujours le même voisinage du début à la fin de la temporalité. Cette caractéristique explique complètement que TCP+ soit aussi performant sur Blob_Rand. Pourquoi l'ancrage du dernier time step de la temporalité de Blob_Rand lors de l'entraînement du TCP+ apporte-t-il d'aussi bons résultats ? Car le dataset est très stable dans ses voisinages. Pour le dataset du Covid, le voisinage ne converge pas, aucun cluster n'est sûr d'être conservé et les instances se mélangent. La conclusion prise lors des expériences sur le TCP+ peut être reformulée. TCP+ est capable de bien fonctionner sur les datasets

possédant un minimum d'ordre même si la dynamique est chaotique, si les voisinages se conservent, TCP+ restera la bonne solution à utiliser pour ce dataset.

11 Discussion des expériences

Toutes ces expériences sur les datasets ont fourni de nombreux résultats mais quelle solution est la meilleure finalement ? Dans cette section, l'objectif va être d'essayer de relever les avantages et inconvénients de chaque solution dans un tableau de comparaisons les classifiant selon les critères suivants :

- La qualité de retranscription de la dynamique
- La qualité de retranscription des changements de direction
- La moyenne du score de trustworthiness dans les meilleures conditions sur l'ensemble des datasets
- La conservation des voisinages sur la temporalité
- Le temps de calcul lors de l'entraînement
- La difficulté à trouver les bons paramètres (λ, perplexité)

Pour chaque critère, les solutions sont classifiées de la meilleure à la pire dans l'ordre croissant. Un tableau pour le Covid et un pour le SVHN sont créés séparément car les solutions n'ont pas performé de façon univoque entre les datasets.

Covid	Pas de pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	TCP	Hybride	TCP+ (compatible avec son dataset)
Dynamique	5	4	6	2	3	1
Direction	6	4	3	2	1	5
Trustworthiness	4	2	6	3	1	5
Temporal AUClogRNX	6	5	4	2	1	3
Temps de calcul	1	2	2	3	4	5
Paramétrage	1	2	3	1	3	1
Rang moyen :	3.8	3.1	4	2.1	2.1	3.3

SVHN	Pas de pénalisation	Pénalisation des mouvements	Pénalisation des accélérations	TCP	Hybride	TCP+ (compatible avec son dataset)
Dynamique	6	5	4	2	3	1
Direction	6	4	5	1	3	2
Trustworthiness	4	4	5	1	3	2
Temporal AUClogRNX	6	5	4	2	3	1
Temps de calcul	1	2	2	3	4	5
Paramétrage	1	2	3	1	3	1
Rang moyen :	4	3.6	3.8	1.6	3.1	2

Ces tableaux sont très utiles pour différencier les avantages et inconvénients de chacune des solutions sur les deux dataset, ils peuvent agir comme une matrice de décision. Dans les faits, il apparaît clairement que TCP et TCP+ sont les techniques sortant du lot pour SVHN. TCP+ est cependant privilégié car c'est cette technique performe au mieux sur la métrique temporal AUClogRNX qui est une métrique très importante pour les datasets convergents bien qu'elle soit pénalisée par son temps de calcul. Cette conclusion ne veut pas dire que TCP+ est la meilleure solution pour le Covid, loin de là. Tout est une question de contexte. Si les instances ne convergent pas, la solution TCP+ devient une mauvaise solution. Le Covid, n'ayant pas de données convergentes, trouvera ses meilleures projections du côté de la version hybride ou du TCP classique. Là où la version hybride pose problème, c'est au niveau de son temps de calcul et de la nécessité de le paramétré. Mais si ces paramètres ne sont pas importants pour l'utilisateur (grâce à un petit dataset) la version hybride devient le meilleur choix.

La solution pénalisant les accélérations peut être considérée comme légèrement plus performante sur ses projections que la solution de l'état de l'art mais elle ne peut pas rivaliser avec les performances des techniques issues de TCP.

12 Conclusion

Pour rappel, le but de la recherche est : « Quelles solutions nouvelles peuvent-être développées afin d'améliorer les techniques existantes de visualisation de données temporelle et de quelles manières est-il possible d'évaluer la qualité de projection de telles techniques ? ». La réponse à cette question se formule en deux étapes.

Première étape, il a été démontré dans ce mémoire qu'une technique TDP, le dynamic t-SNE, pouvait être dérivée en de nombreuses solutions, chacune apportant un lot d'avantages et d'inconvénients. Il a également été démontré que la présence d'un facteur de pénalisation dans la fonction de coût n'est en réalité pas nécessaire au bon fonctionnement de l'algorithme et a tendance à être la source de diverses baisses de qualité de la projection. C'est pourquoi la fonction de coût a été modifiée de manière à rendre la technique plus performante et non dépendante d'un facteur de pénalisation. Le TCP est une solution généralisée à tout type de dataset. Elle est polyvalente et dérivable en deux solutions, le TCP+ étant beaucoup plus performant sur les datasets considérés comme convergents et le TCP hybride capable de traiter les datasets non convergents fournissant des projections plus fiables et avec une conservation des similarités des voisinages plus efficace (temporal AUClogRNX) que le TCP classique.

En deuxième étape, il a été démontré que les métriques existantes applicables aux DP sont également adaptables pour convenir aux TDP. Cette manipulation a permis de constituer une bonne base d'évaluation des TDP. La grosse contribution au domaine des métriques aura été la création de trois métriques originales et mesurant des facteurs critiques à la bonne qualité d'une projection dimensionnelle temporelle. Ces trois métriques constituent, avec l'ajout des métriques adaptées aux TDP, une base importante de méthodes d'évaluation des TDP.

Des améliorations dans le domaine des techniques de TDP et dans le domaine des métriques sont envisagées. Une nette optimisation des solutions basée sur TCP est encore à effectuer. Les résultats que ces techniques ont fournis sont très satisfaisants mais sont loin d'être parfaits. En plus de l'augmentation de leur performance, il a été montré que ces techniques se retrouveront fortement handicapées sur des datasets plus volumineux que SVHN ou celui du Covid car leur grand inconvénient provient de la quantité de calculs supplémentaires nécessaires à leur fonctionnement. Une solution aussi efficace mais moins complexe serait un grand atout pour une utilisation évolutive des techniques de TDP.

Au niveau des méthodes d'évaluation, il reste encore à développer des métriques pour capturer certains aspects non pris en compte lors de ce mémoire. Parmi ces aspects, une métrique capable de regrouper les time steps selon les voisinages de leurs instances détectant les time steps similaires sans se préoccuper de choisir un time step de référence serait une bonne idée. Cette métrique permettrait facilement de mettre en relation des time steps et faciliterait les interprétations et la détection de patterns redondants dans la projection, facteur très utile lorsque des données non convergentes sont projetées.

Pour conclure, ce qu'il faut retenir des données temporelles c'est leur difficulté d'approche. Il est très compliqué de créer une solution générale capable de fournir de bons résultats quelle que soit la nature du dataset. Le domaine des techniques de TDP est très jeune et doit encore se développer, mais il est indéniable que ces techniques seront indispensables lorsque la question de l'exploitation de données aussi complexes que les données temporelles se posera.

Bibliographie

[1] «Nuage de points (statistique),» Wikipédia, 13 10 2021. [En ligne]. Available: https://fr.wikipedia.org/wiki/Nuage_de_points_(statistique)#/media/Fichier:Linear_regression.svg. [Accès le 15 04 2022].

[2] «Google Play: Shadowmatic,» Triada Studio Games, [En ligne]. Available: https://play-lh.googleusercontent.com/jzrn13_YoCTWrKPQ8S1aFmDDbeoU6u22v_U3wWFmwy-g9eHpui0j4UB_B15THB0aVME. [Accès le 24 04 2022].

[3] L. &. H. G. Van der Maaten, Visualizing data using t-SNE, Journal of machine learning research, 2008.

[4] S. M. Guy Shtar, «Clustering and Dimensionality Reduction: Understanding the “Magic” Behind Machine Learning,» 31 07 2017. [En ligne]. Available: https://www.imperva.com/blog/clustering-and-dimensionality-reduction-understanding-the-magic-behind-machine-learning/. [Accès le 20 04 2022].

[5] «Dimensionality Reduction using t-Distributed Stochastic Neighbor Embedding (t-SNE) on the MNIST Dataset,» 16 08 2020. [En ligne]. Available: https://towardsdatascience.com/dimensionality-reduction-using-t-distributed-stochastic-neighbor-embedding-t-sne-on-the-mnist-9d36a3dd4521. [Accès le 13 04 2022].

[6] P. E. F. A. X. &. T. A. C. Rauber, Visualizing Time-Dependant Data Using Dynamic t-SNE, 2016.

[7] «Epistat,» [En ligne]. Available: https://epistat.wiv-isp.be/covid/. [Accès le 01 Mars 2022].

[8] T. W. A. C. A. B. B. W. A. Y. N. Yuval Netzer, «Reading Digits in Natural Images with Unsupervised Feature Learning,» NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011. [En ligne]. Available: http://ufldl.stanford.edu/housenumbers/. [Accès le 01 Mars 2022].

[9] J. A. L. a. M. Verleysen., «Scale-independent quality criteria for dimensionality reduction,» *Pattern Recognition Letters,* vol. 31, pp. 2248-2257, 2010.

[10] A. M. R. M. &. K. A. Chatzimparmpas, «t-visne: Interactive assessment and interpretation of t-sne projection,» *IEEE transactions on visualization and computer graphics,* vol. 26, pp. 2696-2714, 2020.

[11] J. &. K. S. Venna, «Visualizing gene interaction graphs with local multidimensional scaling,» *ESANN,* vol. 6, pp. 557-562, 2006, April.

[12] H. K. H. K. J. J. K. Y. &. S. J. Jeon, «Measuring and explaining the inter-cluster reliability of multidimensional projections,» *IEEE transactions on Visualization and Computer Graphics,* vol. 28, pp. 551-561, 2021.

[13] J. A. P.-O. D. H. &. V. M. Lee, Multiscale stochastic neighbor embedding: Towards parameter-free dimensionality reduction, ESANN, 2014, April.

[14] E. F. G. R. S. I. D. C. J. L. D. &. T. A. C. Vernier, «Quantitative Evaluation of Time-Dependent Multidimensional Projection Techniques,» *Computer Graphics Forum,* vol. 39, n° %13, pp. 241-252, 2020, June.