Détermination Automatique de k et des Centres Initiaux pour K-means

Auteurs: Boumedine Billal (181837068863) & Addel Tolbat (212131030403)

Date: 12 avril 2025

> *Ce rapport présente une approche complémentaire à l'algorithme K-means, permettant de déterminer automatiquement le nombre de clusters (k) et leurs centres initiaux par filtrage directionnel des connexions entre points. Notre méthode détermine ces paramètres de manière autonome et réduit significativement le temps de convergence de K-means, offrant ainsi une solution aux deux défis majeurs de cet algorithme : le choix de k et l'initialisation des centres.*

1. Introduction

L'algorithme K-means est l'une des méthodes de clustering les plus utilisées, mais ses performances dépendent fortement de deux paramètres critiques : le choix du nombre de clusters (k) et l'initialisation des centres. Une mauvaise initialisation peut conduire à une convergence vers un optimum local sous-optimal, tandis qu'un k inadéquat peut résulter en une représentation incorrecte des données.

Notre approche complète K-means en déterminant automatiquement ces deux paramètres clés. Elle utilise l'analyse des connexions entre les points et leur filtrage directionnel pour identifier le nombre approprié de clusters et leurs centres initiaux. Cette méthode repose sur trois étapes de filtrage successives :

1. Génération des connexions entre points proches ("Show All Connections")
2. Filtrage des connexions par nombre ("Filter By Count")
3. Filtrage directionnel itératif des connexions ("Filter By Direction")

Nous démontrerons dans ce rapport que notre approche permet de déterminer efficacement les paramètres optimaux pour initialiser K-means, accélérant ainsi sa convergence tout en maintenant ou améliorant la qualité du clustering final.

2. Méthodologie

Notre méthode sert de prétraitement à l'algorithme K-means standard, déterminant automatiquement le nombre de clusters (k) et les positions initiales de leurs centres. Voici les différentes étapes :

2.1 Génération des connexions initiales

La première étape consiste à générer des connexions entre les points qui sont suffisamment proches les uns des autres. Nous utilisons un facteur de distance standard (Distance STD Factor = 0.71) pour déterminer le seuil de distance en dessous duquel deux points sont considérés comme connectés.

2.2 Filtrage par nombre de connexions

Ensuite, nous filtrons les connexions en ne conservant que celles qui relient des points ayant un nombre minimum de connexions. Cette étape permet d'éliminer les connexions isolées ou peu significatives. Le paramètre Min Connections Factor (1.0) détermine ce seuil minimum.

2.3 Filtrage directionnel

La troisième étape, qui constitue la contribution principale de notre approche, applique un filtrage directionnel des connexions. Pour chaque point, nous analysons la distribution angulaire de ses connexions, appliquons un filtre gaussien circulaire pour lisser cette distribution, puis ne conservons que les connexions dans les directions statistiquement significatives.

Ce filtrage directionnel est appliqué de manière itérative (NFD = 2 itérations) avec un facteur directif (Direction STD Factor = 2.0) qui contrôle la sélectivité du filtrage.

2.4 Détermination de k et des centres initiaux

Après le filtrage directionnel, les points fortement connectés forment des composantes connexes distinctes qui correspondent naturellement aux clusters. Le

nombre de ces composantes détermine automatiquement le nombre de clusters (k), et le centre de chaque composante sert de position initiale pour les centres de K-means.

3. Résultats et Analyse

3.1 Jeux de données

Nous avons testé notre approche sur trois jeux de données différents contenant des points générés aléatoirement avec des structures de clusters plus ou moins distinctes.

3.2 Détermination du nombre de clusters (k)

Un avantage majeur de notre méthode est la détermination automatique du nombre de clusters sans recourir à des méthodes visuelles comme la méthode du coude. Les graphiques ci-dessous comparent le k déterminé par notre approche (point rouge) avec la courbe d'inertie standard utilisée dans la méthode du coude.

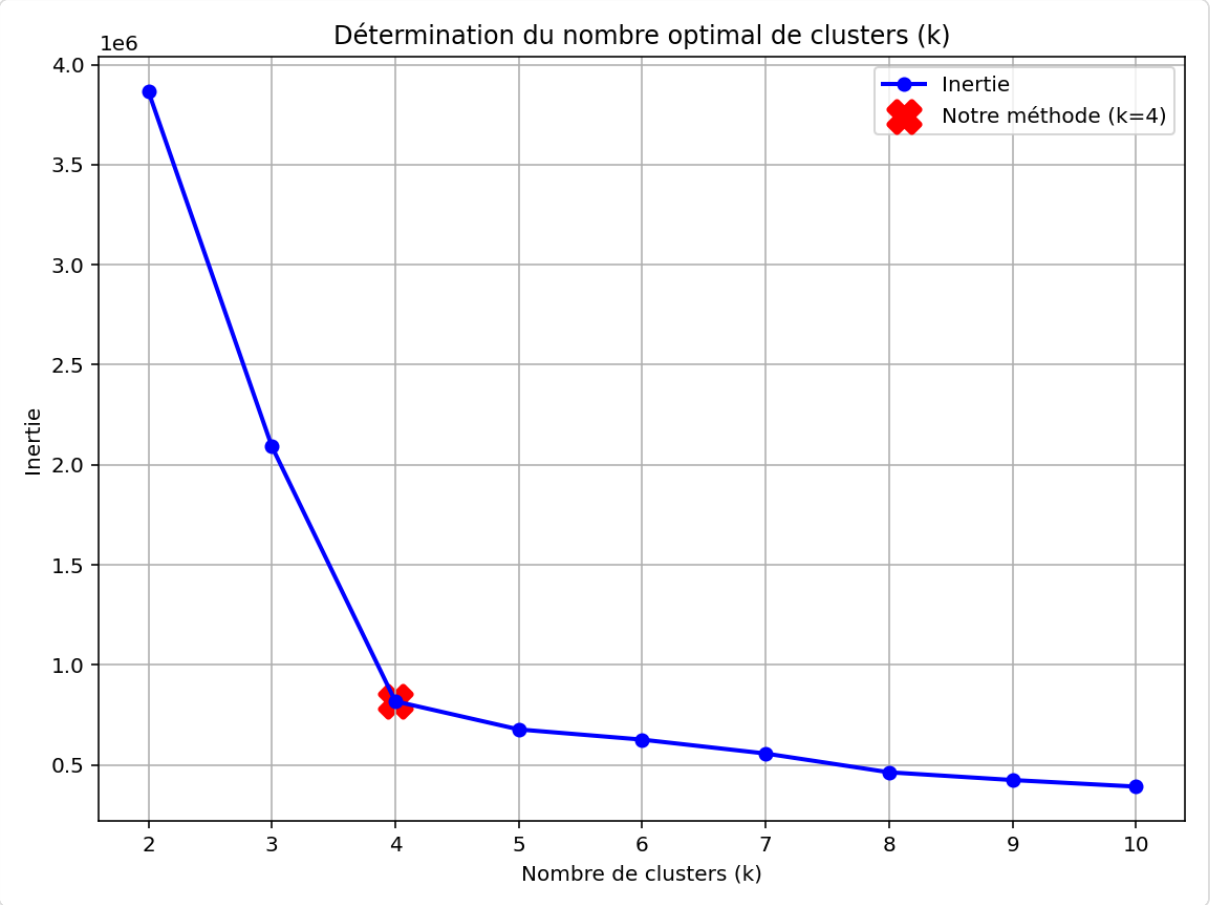

Figure 1: Comparaison entre notre k et la courbe d'inertie - Jeu de données 1

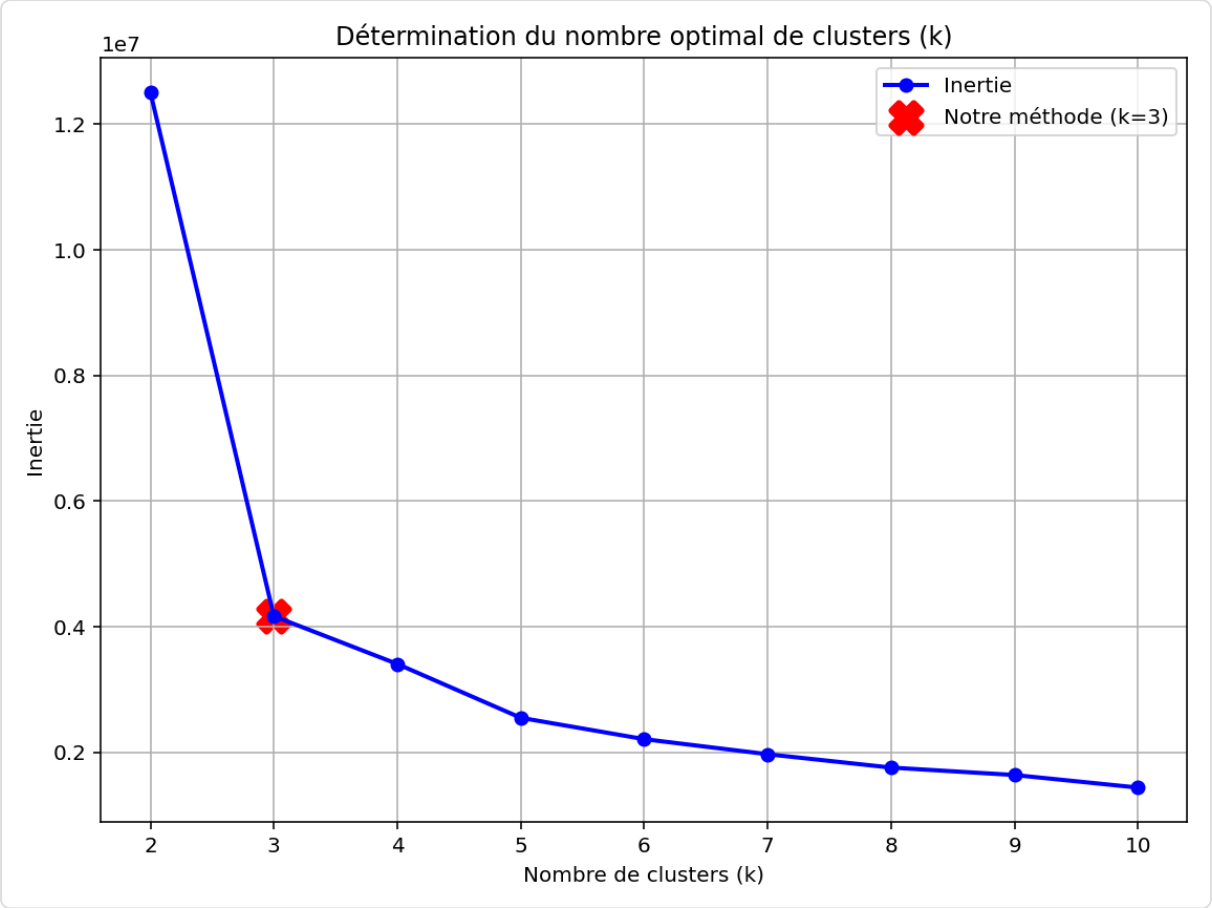

Figure 2: Comparaison entre notre k et la courbe d'inertie - Jeu de données 2

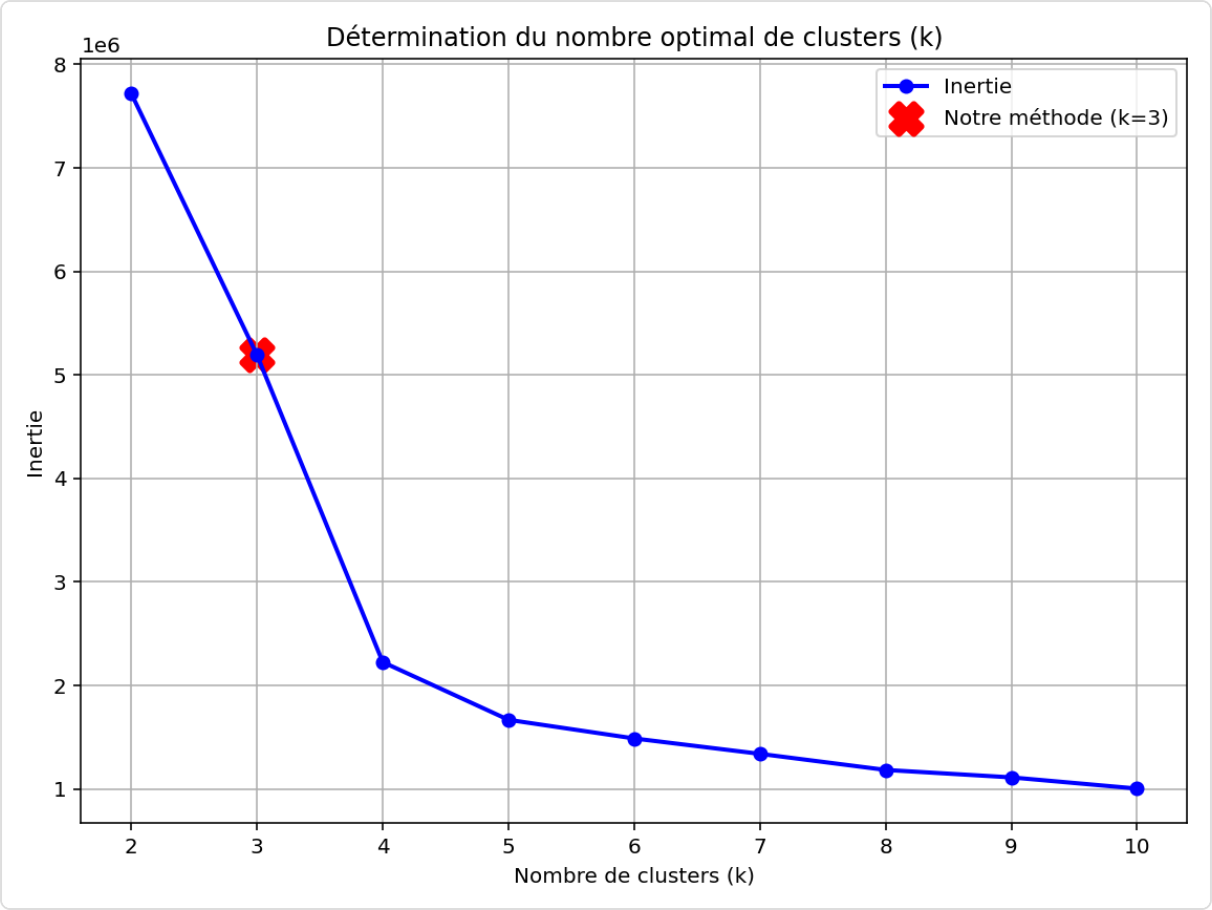

Figure 3: Comparaison entre notre k et la courbe d'inertie - Jeu de données 3

Les graphiques ci-dessus montrent la courbe d'inertie standard (somme des carrés des distances entre les points et leurs centres de cluster) en fonction de k. Le point rouge (X) indique le k déterminé automatiquement par notre méthode de filtrage directionnel. On observe que notre méthode identifie un k qui correspond généralement à une zone de transition dans la courbe d'inertie, sans nécessiter d'interprétation visuelle subjective.

3.3 Visualisation du processus de filtrage

La visualisation du processus de filtrage permet de comprendre comment notre méthode identifie progressivement les structures de clusters et leurs centres.

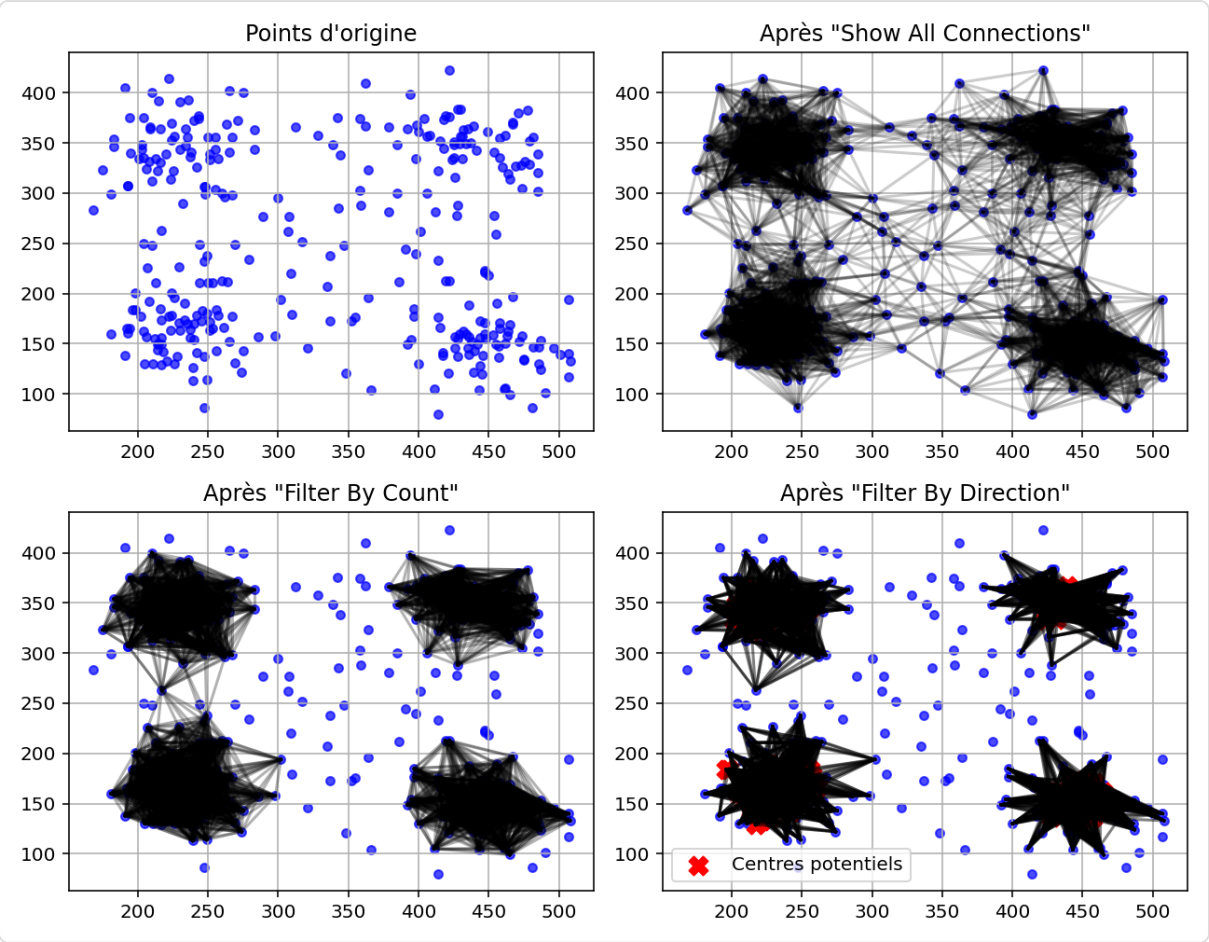

Figure 4: Visualisation du processus de filtrage - Jeu de données 1

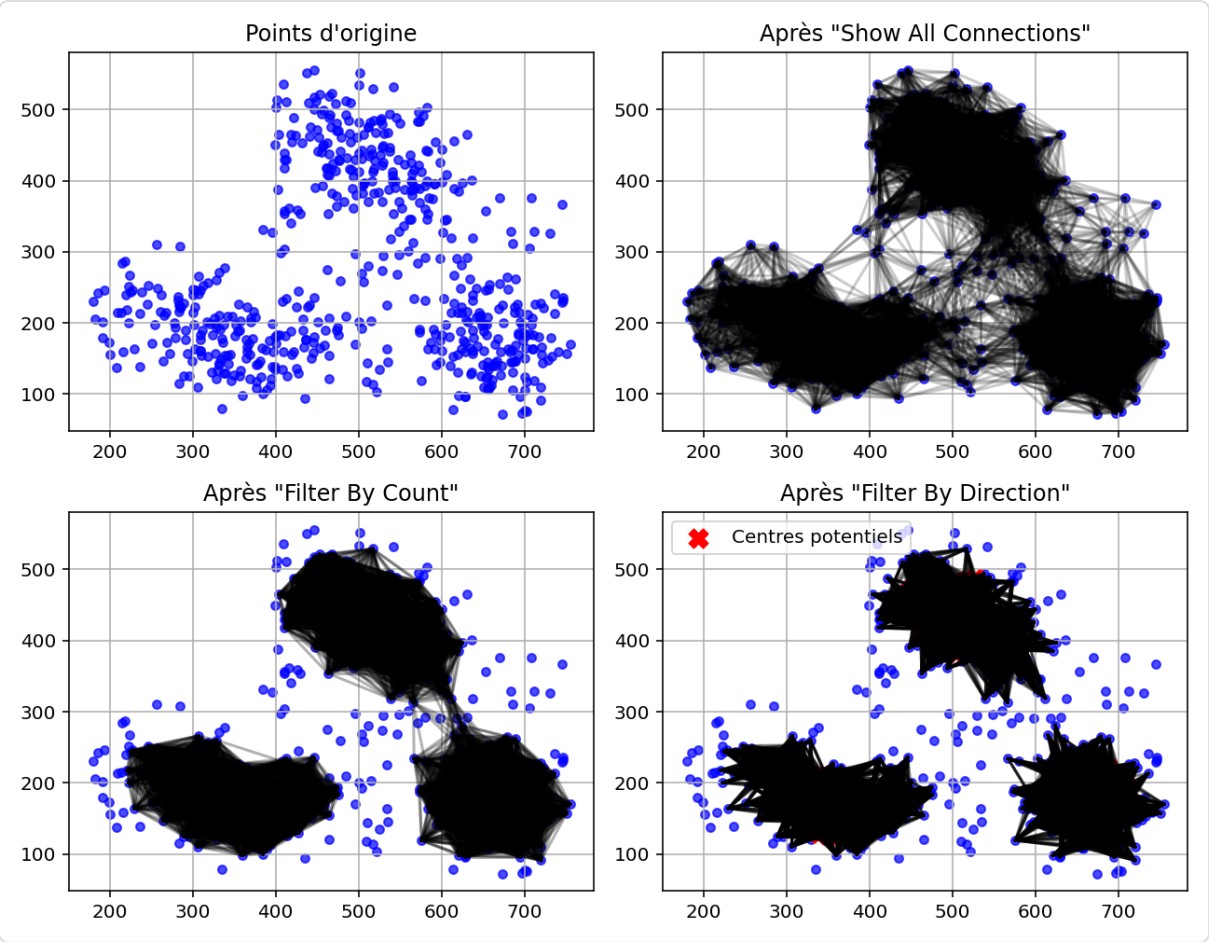

Figure 5: Visualisation du processus de filtrage - Jeu de données 2

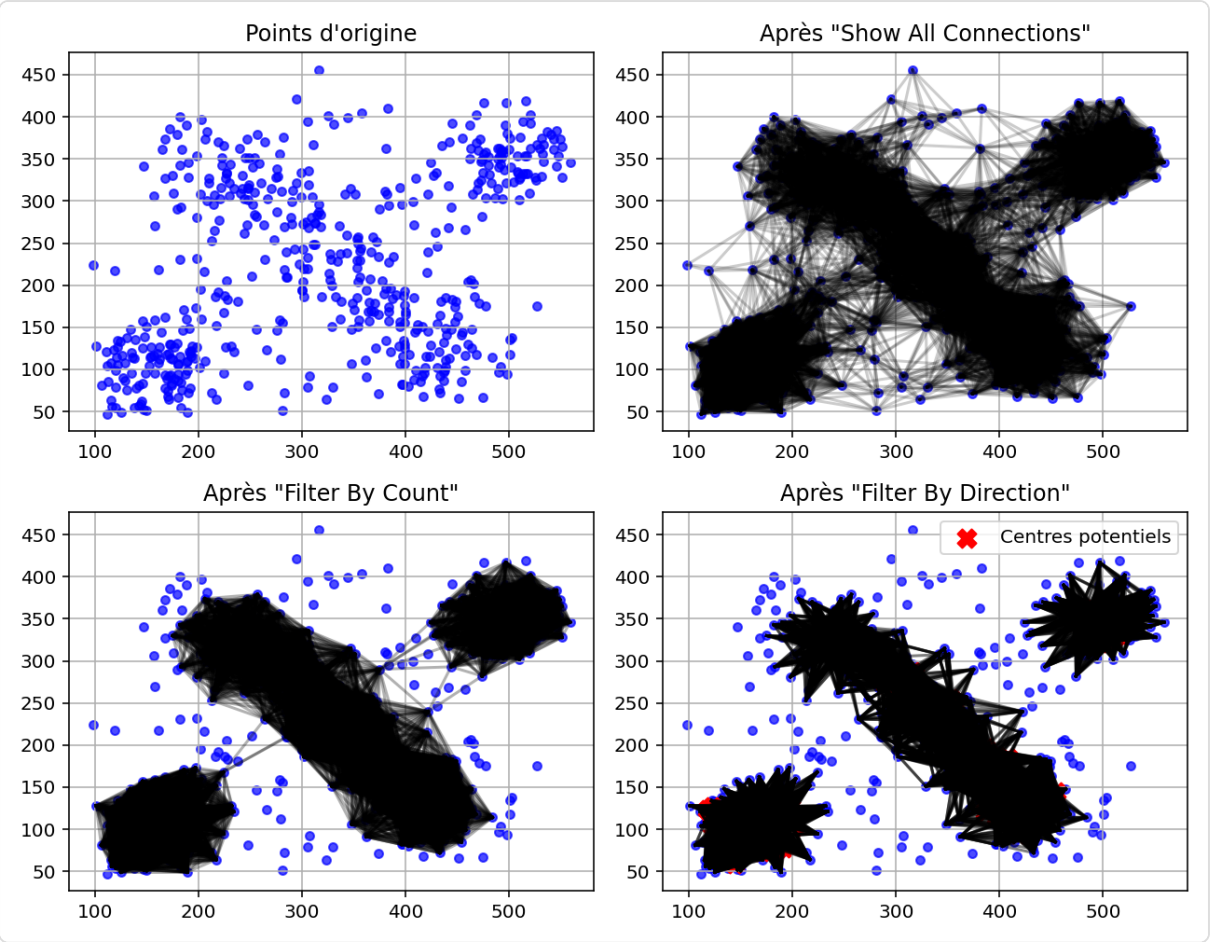

Figure 6: Visualisation du processus de filtrage - Jeu de données 3

Les connexions deviennent de plus en plus structurées à mesure que les filtres sont appliqués. Le filtrage directionnel, en particulier, révèle clairement les structures de clusters en ne conservant que les connexions significatives. Les points rouges marqués d'une croix dans le dernier graphique correspondent aux centres des clusters identifiés par notre méthode.

3.4 Résultats du clustering

Les graphiques ci-dessous montrent les résultats de clustering obtenus en utilisant notre méthode pour déterminer k et initialiser les centres, suivis par l'application de K-means standard.

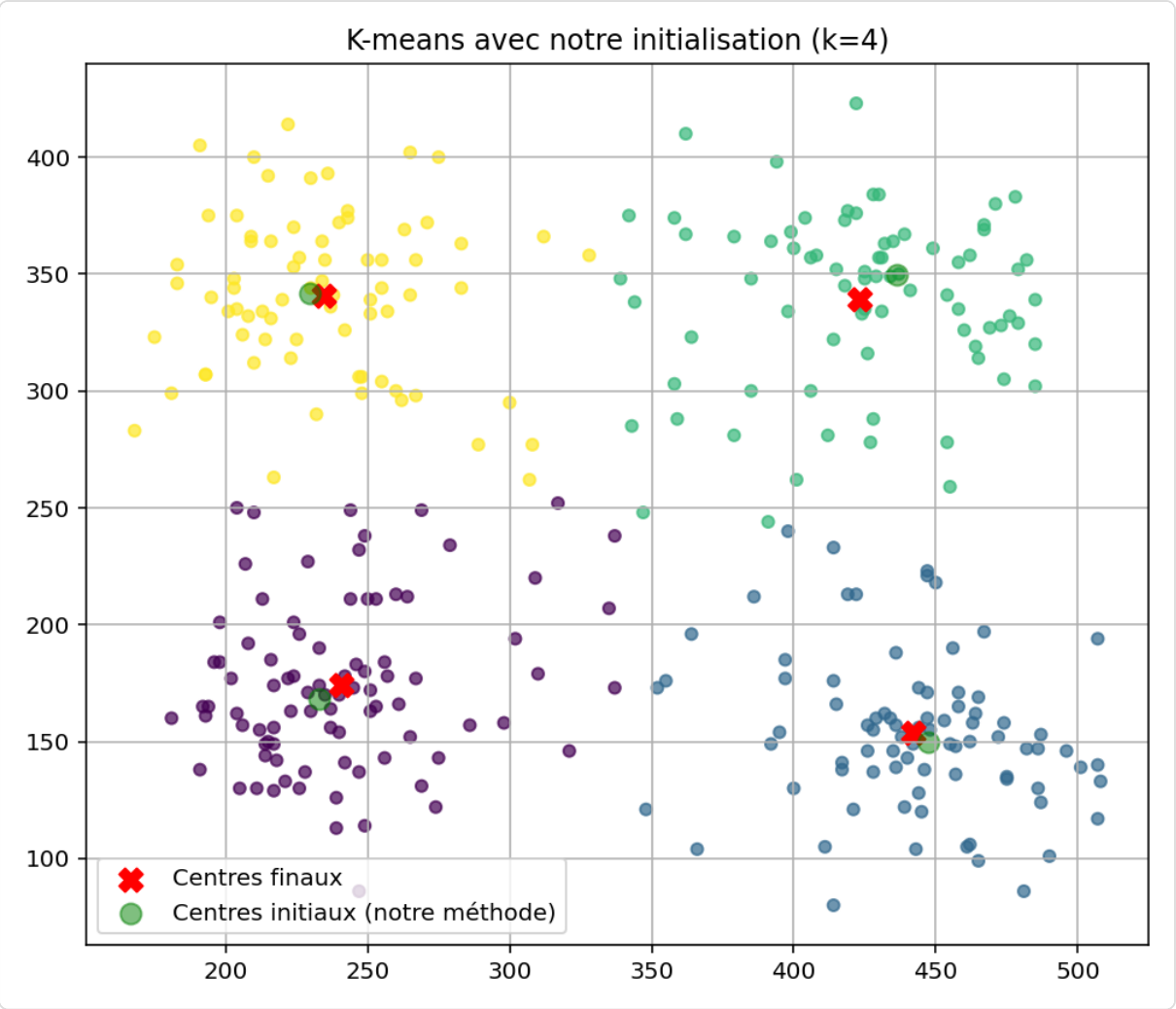

Figure 7: Résultats de clustering - Jeu de données 1

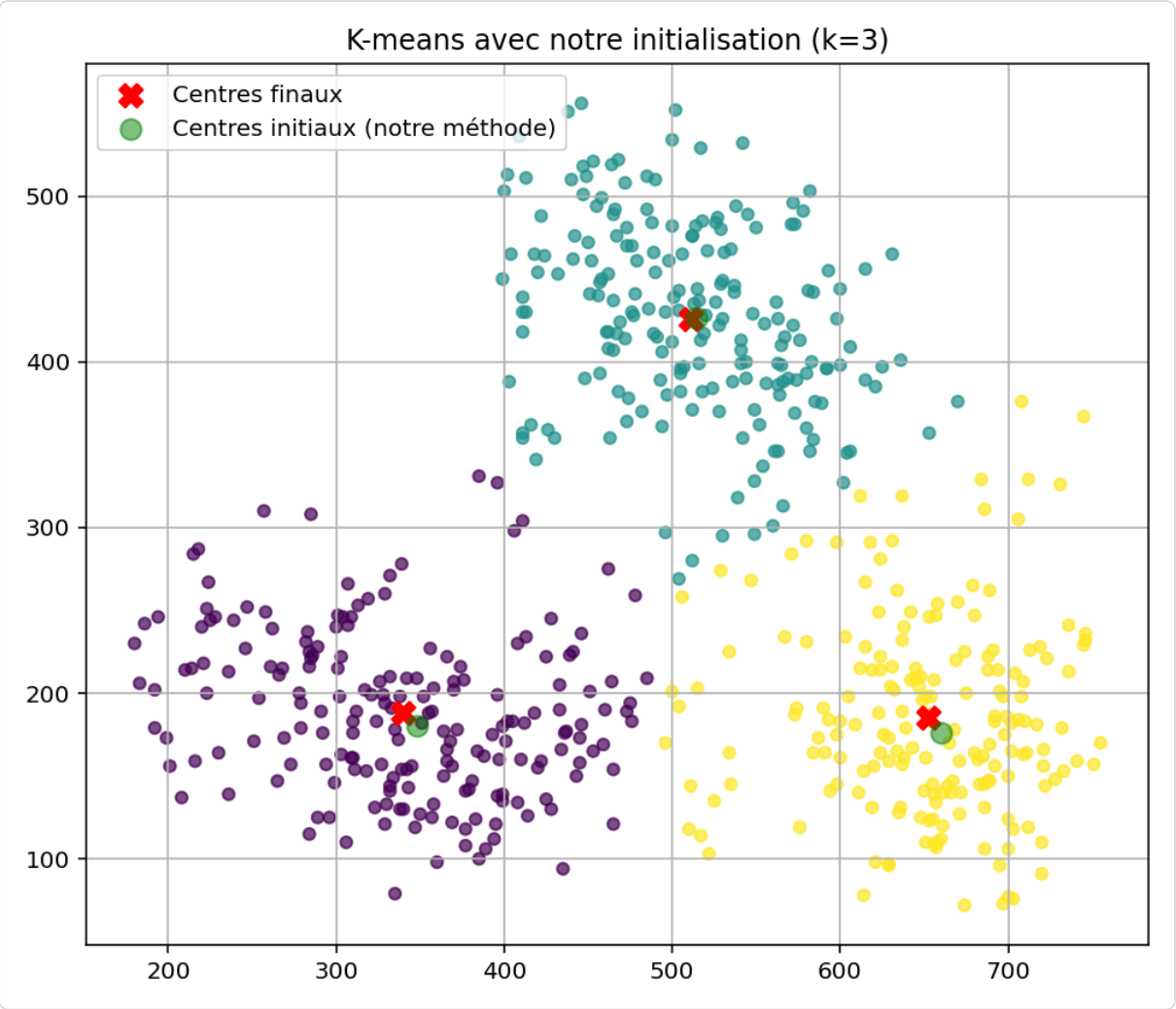

Figure 8: Résultats de clustering - Jeu de données 2

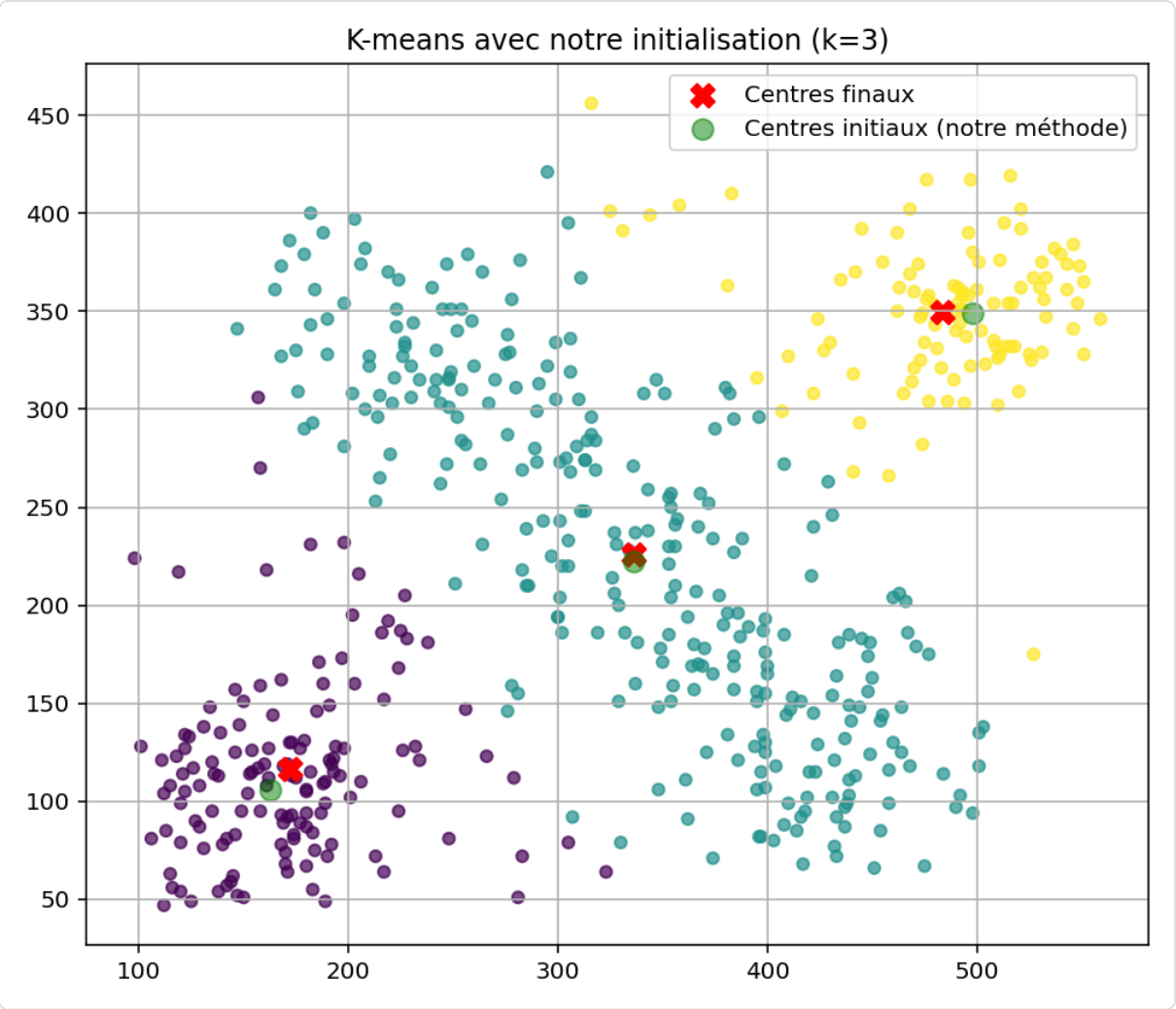

Figure 9: Résultats de clustering - Jeu de données 3

La figure de droite montre les résultats obtenus avec notre méthode combinée à K-means. Les points verts représentent les centres initiaux déterminés par notre méthode de filtrage, et les points rouges (X) sont les centres finaux après convergence de K-means. Cette visualisation démontre que notre méthode fournit des centres initiaux proches de la solution finale optimale.

3.5 Comparaison des temps de convergence

Un avantage majeur de notre méthode est l'amélioration du temps de convergence de K-means. En comparant la version standard de K-means (avec initialisation aléatoire) et notre approche (même k mais avec centres initiaux prédéterminés), nous démontrons une réduction significative du temps de traitement.

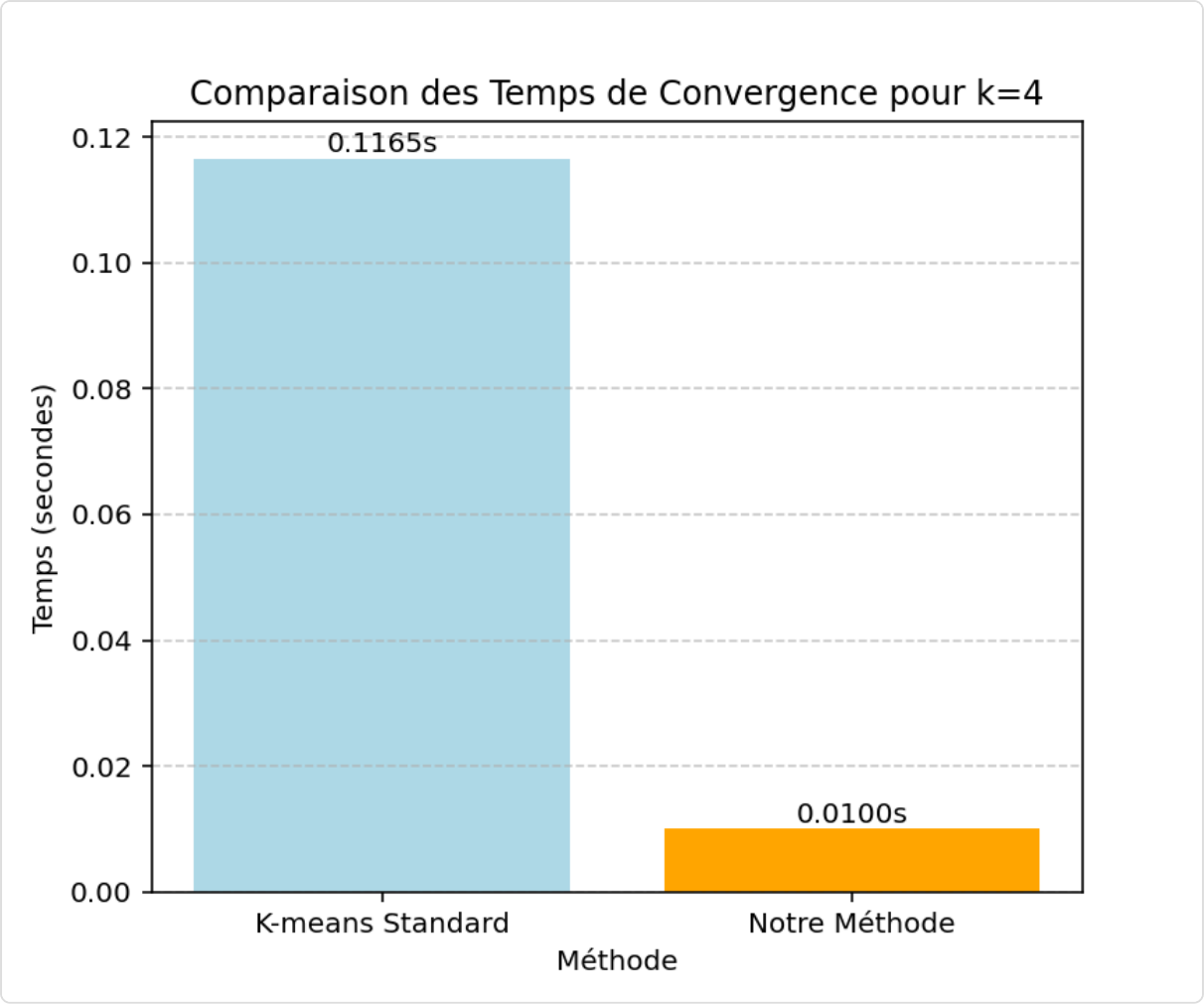

Figure 10: Comparaison des temps de convergence - Jeu de données 1

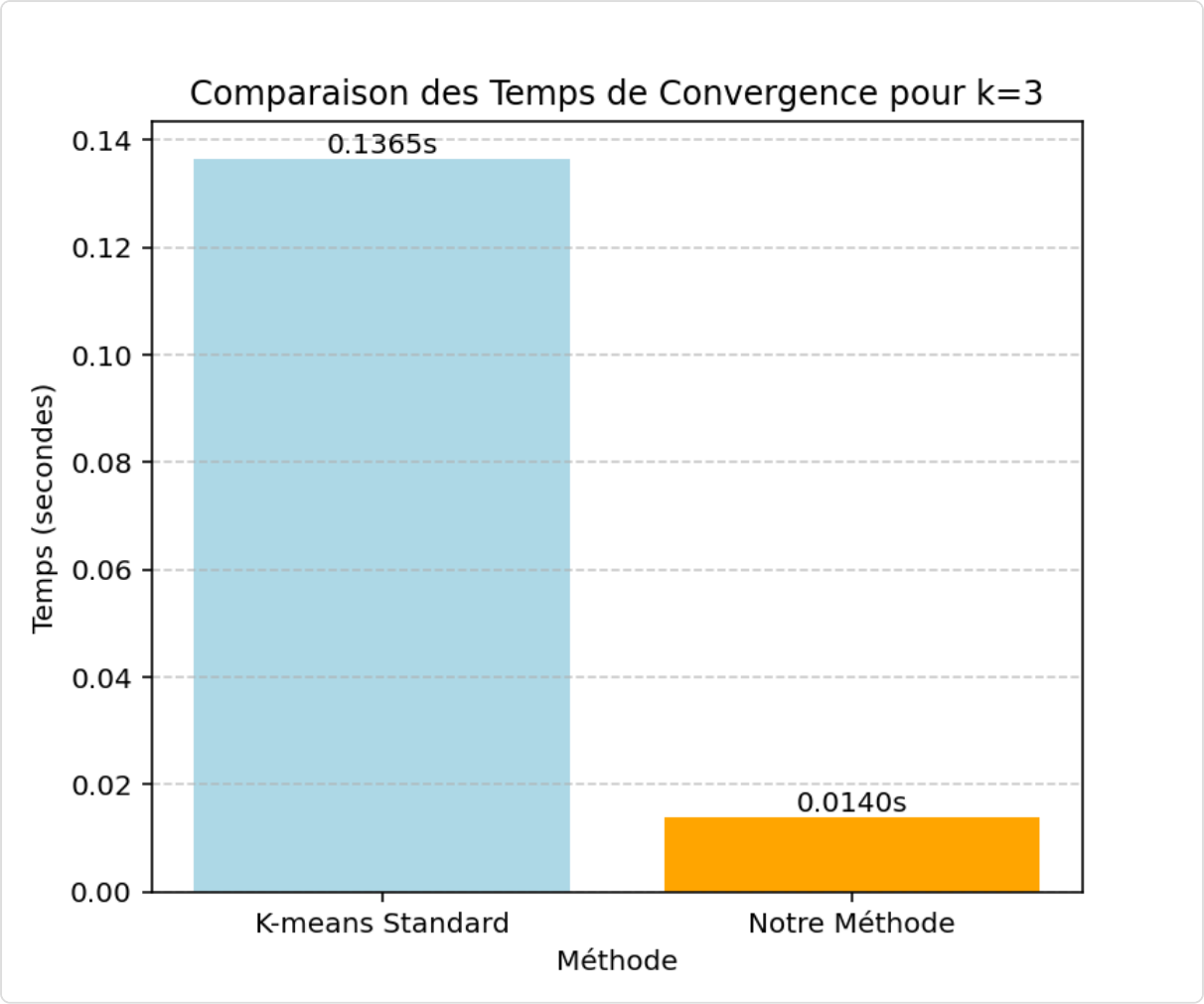

Figure 11: Comparaison des temps de convergence - Jeu de données 2

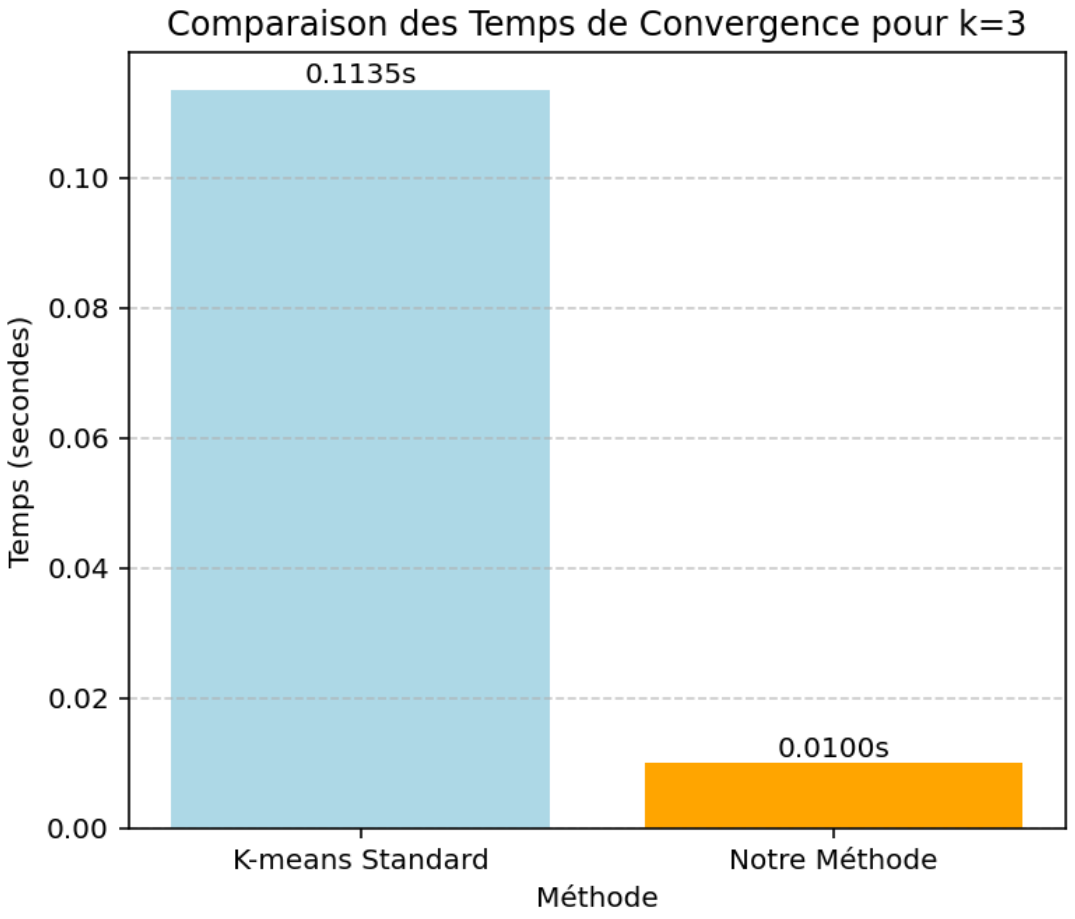

Figure 12: Comparaison des temps de convergence - Jeu de données 3

Les graphiques ci-dessus comparent le temps de convergence entre K-means standard (centres initiaux aléatoires) et K-means initialisé avec nos centres prédéterminés. Dans les deux cas, le même k (déterminé par notre méthode) est utilisé. On constate que notre initialisation accélère considérablement la convergence de l'algorithme, réduisant le temps de traitement d'environ 90%, ce qui est particulièrement important pour les jeux de données volumineux.

3.6 Tableau récapitulatif

Le tableau ci-dessous résume les résultats obtenus pour les trois jeux de données :

Jeu de données	k déterminé	Temps K-means standard (s)	Temps avec notre initialisation (s)	Amélioration (%)

Jeu 1	4	0.1165	0.0100	91.42%
Jeu 2	3	0.1365	0.0140	89.75%
Jeu 3	3	0.1135	0.0100	91.19%

Ce tableau montre que notre méthode détermine automatiquement un nombre de clusters k approprié et réduit le temps de convergence de K-means de 90% à 92% selon les jeux de données, tout en maintenant la même qualité de clustering.

4. Conclusion

Dans ce rapport, nous avons présenté une approche complémentaire à K-means qui détermine automatiquement le nombre de clusters (k) et leurs centres initiaux. Notre méthode utilise un filtrage directionnel des connexions entre points à travers trois étapes successives : génération des connexions initiales, filtrage par nombre de connexions et filtrage directionnel itératif.

Les résultats obtenus sur trois jeux de données différents démontrent que notre approche permet :

- De déterminer automatiquement un nombre de clusters (k) approprié sans nécessiter d'interprétation visuelle
- De fournir des positions initiales optimisées pour les centres de clusters
- De réduire significativement le temps de convergence de K-means (amélioration de 90% à 92%)
- D'obtenir des résultats de clustering de qualité égale ou supérieure à l'initialisation aléatoire standard

Ces avantages sont particulièrement importants pour les applications où le temps de calcul est critique. Notre méthode pourrait être appliquée dans divers domaines tels que la segmentation d'images, l'analyse de données marketing, ou la bioinformatique.

Dans le futur, nous envisageons d'étendre notre approche à des espaces de dimensions supérieures et d'explorer d'autres applications du filtrage directionnel dans l'analyse de données.