Projet de Programmation Impérative

K Plus Proches Voisins

Mathevon B., Palisse E.

UJM - L2 Info

Semestre 4

Sommaire

Protocole Expérimental

Mesure du temps de calcul des KPPV en fonction de deux critères :
- Un nombre de K fixe
- Un nombre de points fixe

K prend la valeur suivante : $k = (2^n) + 1$
K allant de 1 a n-1 , max $= 10000$

N prend les valeurs suivantes :
10-25-50-100-250-500-1000-5000-10000-50000-100000-500000-1000000

Pour chaque valeur de N : 10 fichiers aléatoires générés à partir du script data_gen.sh(nb_n, nb_dim, nb_classe)

Protocole Expérimental

Pour chaque valeur de n et chaque valeur de k : calcul du temps x10

Deux méthodes de tris utilisés pour le tableau :
-Tri a bulles ($O(n^2)$)
-Tri fusion ($O(n \log n)$)

Pour le tri a bulles : n max = 10 000
Pour le tri fusion : n max = 1 000 000

Arbre KD ($O(\log n)$) : n max = 1 000 000

Calcul de la moyenne pour chaque méthode de tri et chaque valeur de k/n
Calcul de l'écart-type (affichage sur les courbes)

Scripts utilisés

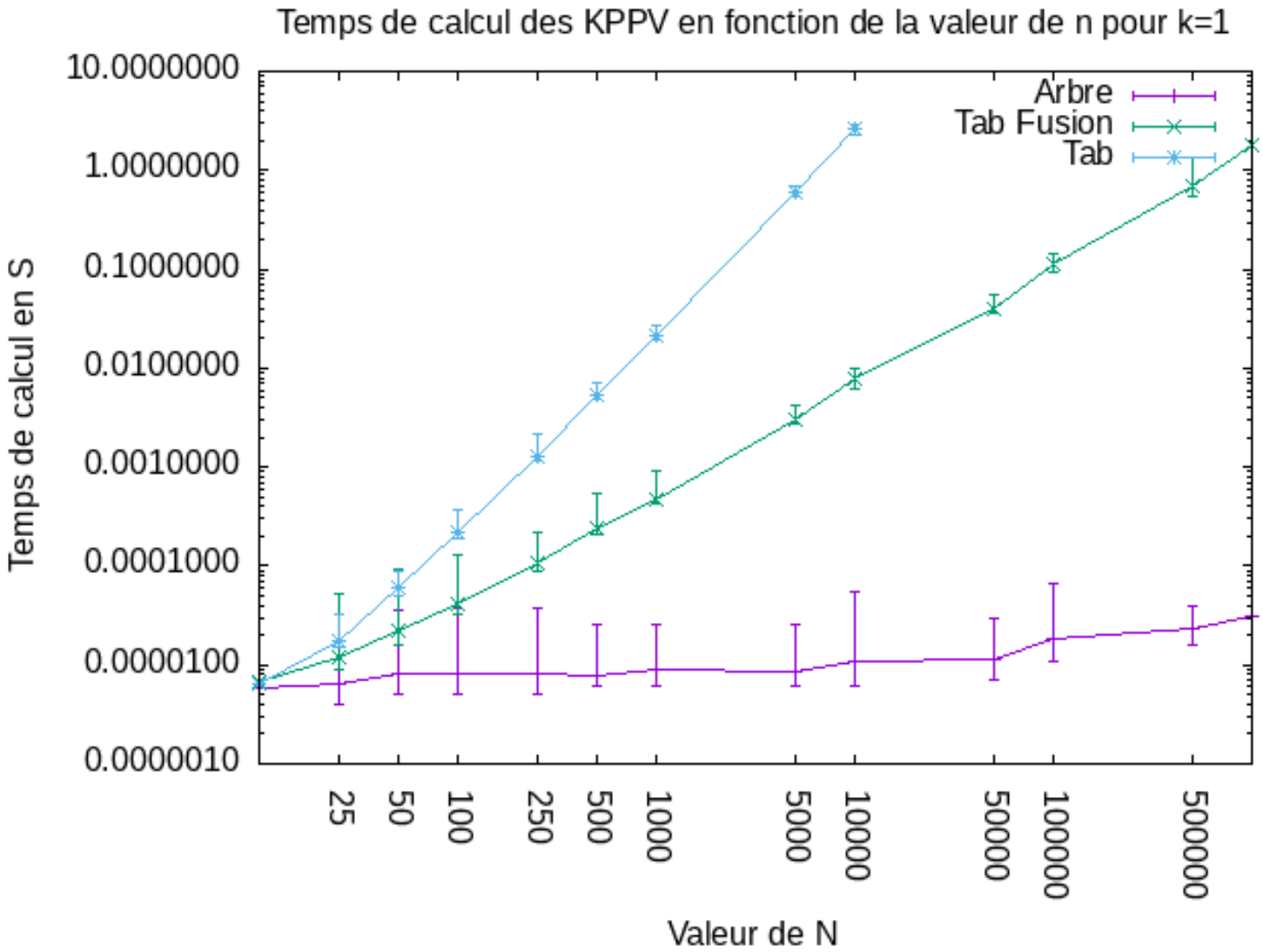
Temps de calcul :
Usage : script_time.sh
⟨nb_points⟩
⟨valeur_de_k⟩
⟨mode(tab/tab_fusion/arbre)⟩

Moyenne des temps de calcul :
Usage: script_moyenne.sh
⟨type_moy (en fct de k/n)⟩
⟨nb_n si n/nb_k si k⟩
⟨mode⟩

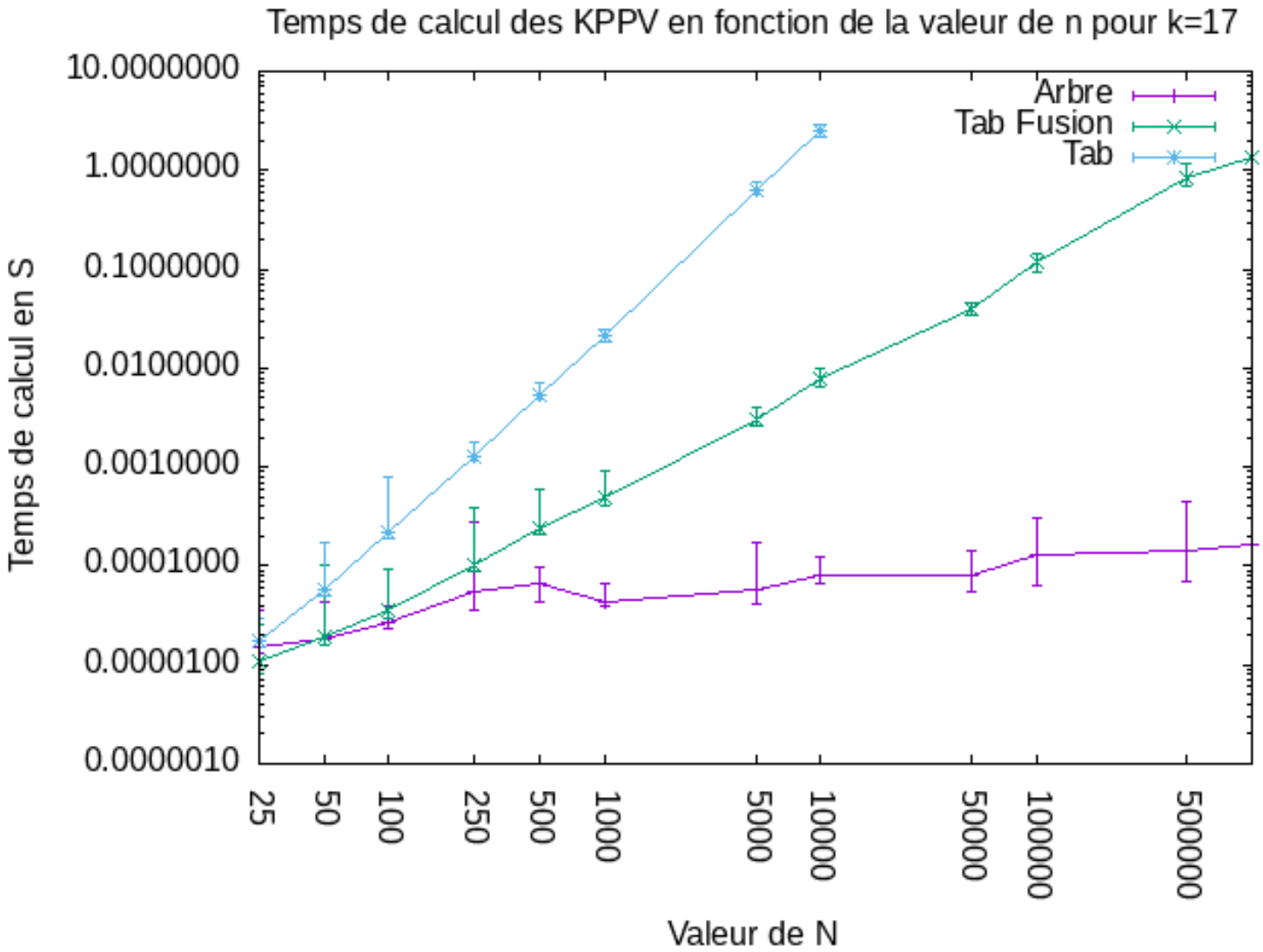
Génération d'un graphique :
Usage: script_graph.sh
⟨en fct de n ou k⟩
⟨valeur de n/k⟩

Table des Matières

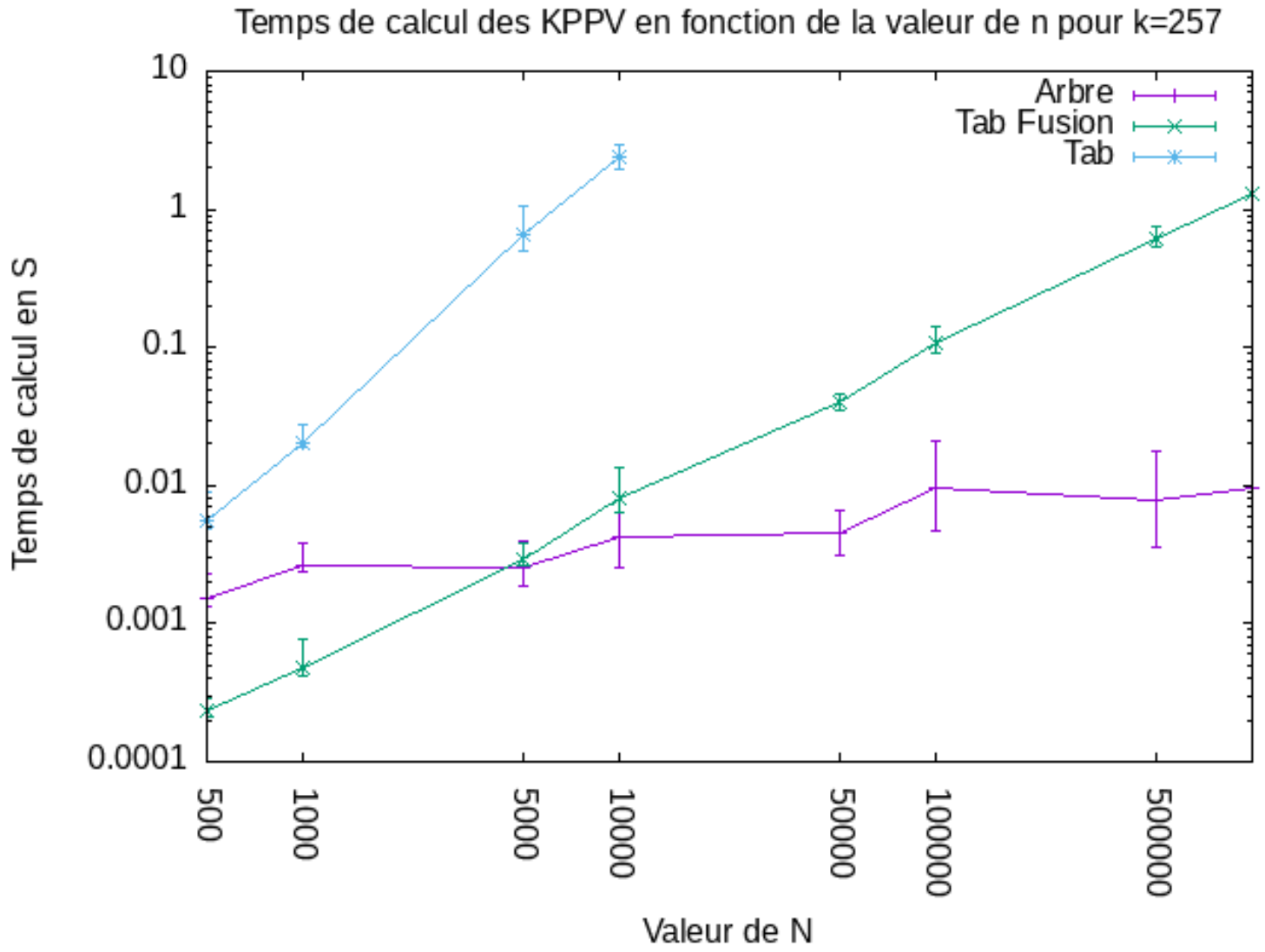
K fixe


Temps de calcul des KPPV en fonction de la valeur de n pour k=1
Temps de calcul en S
10.0000000
1.0000000
0.1000000
0.0100000
0.0010000
0.0001000
0.0000100
0.0000010
Arbre
Tab Fusion
Tab
25
50
100
250
500
1000
5000
10000
50000
100000
500000
Valeur de N

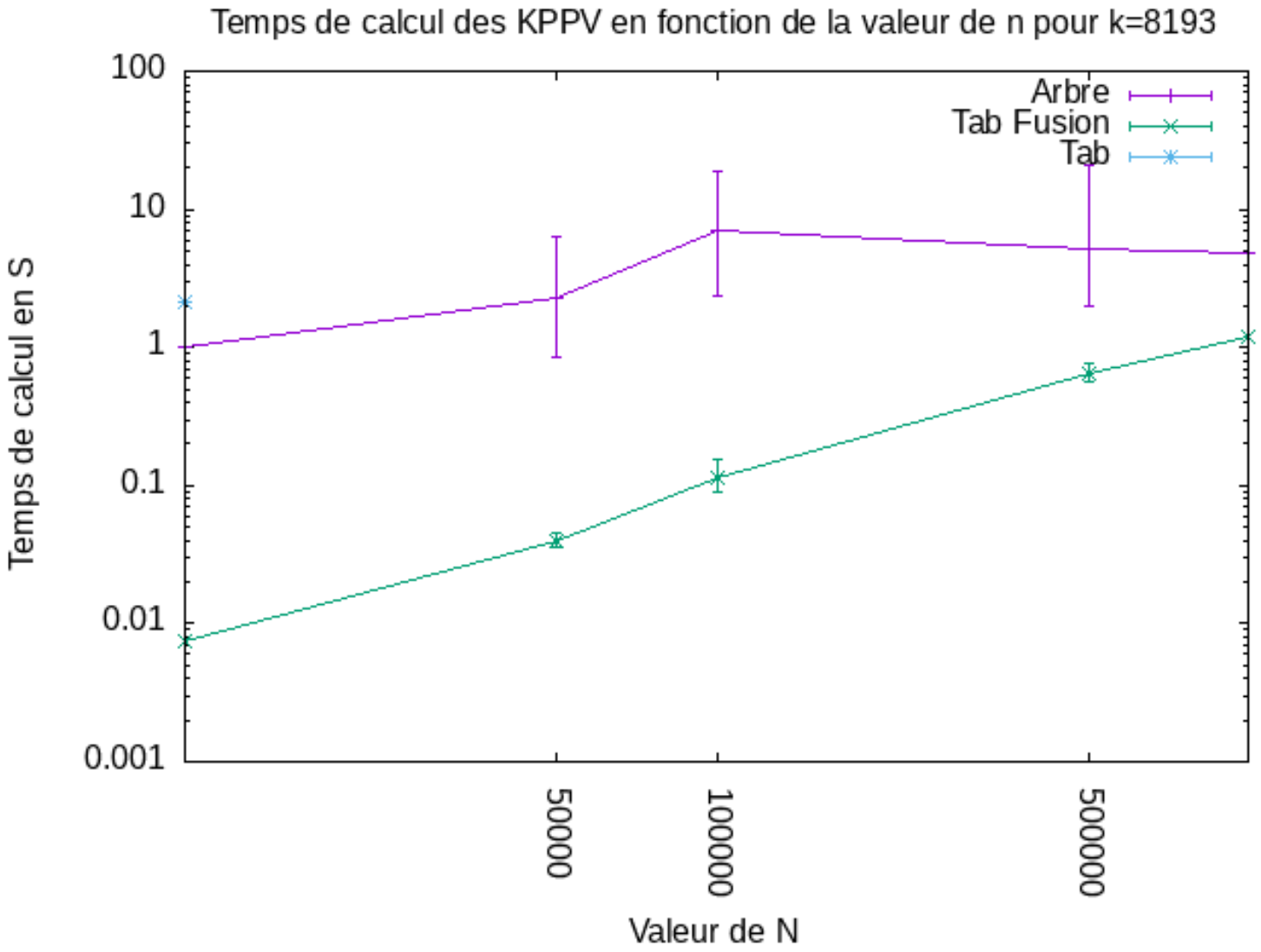
K fixe


Temps de calcul des KPPV en fonction de la valeur de n pour k=17
Temps de calcul en S
10.0000000
1.0000000
0.1000000
0.0100000
0.0010000
0.0001000
0.0000100
0.0000010
Arbre
Tab Fusion
Tab
25
50
100
250
500
1000
5000
10000
50000
100000
500000
Valeur de N

K fixe


Temps de calcul des KPPV en fonction de la valeur de n pour k=257
Temps de calcul en S
10
1
0.1
0.01
0.001
0.0001
Arbre
Tab Fusion
Tab
500
1000
5000
10000
50000
100000
500000
Valeur de N

K fixe

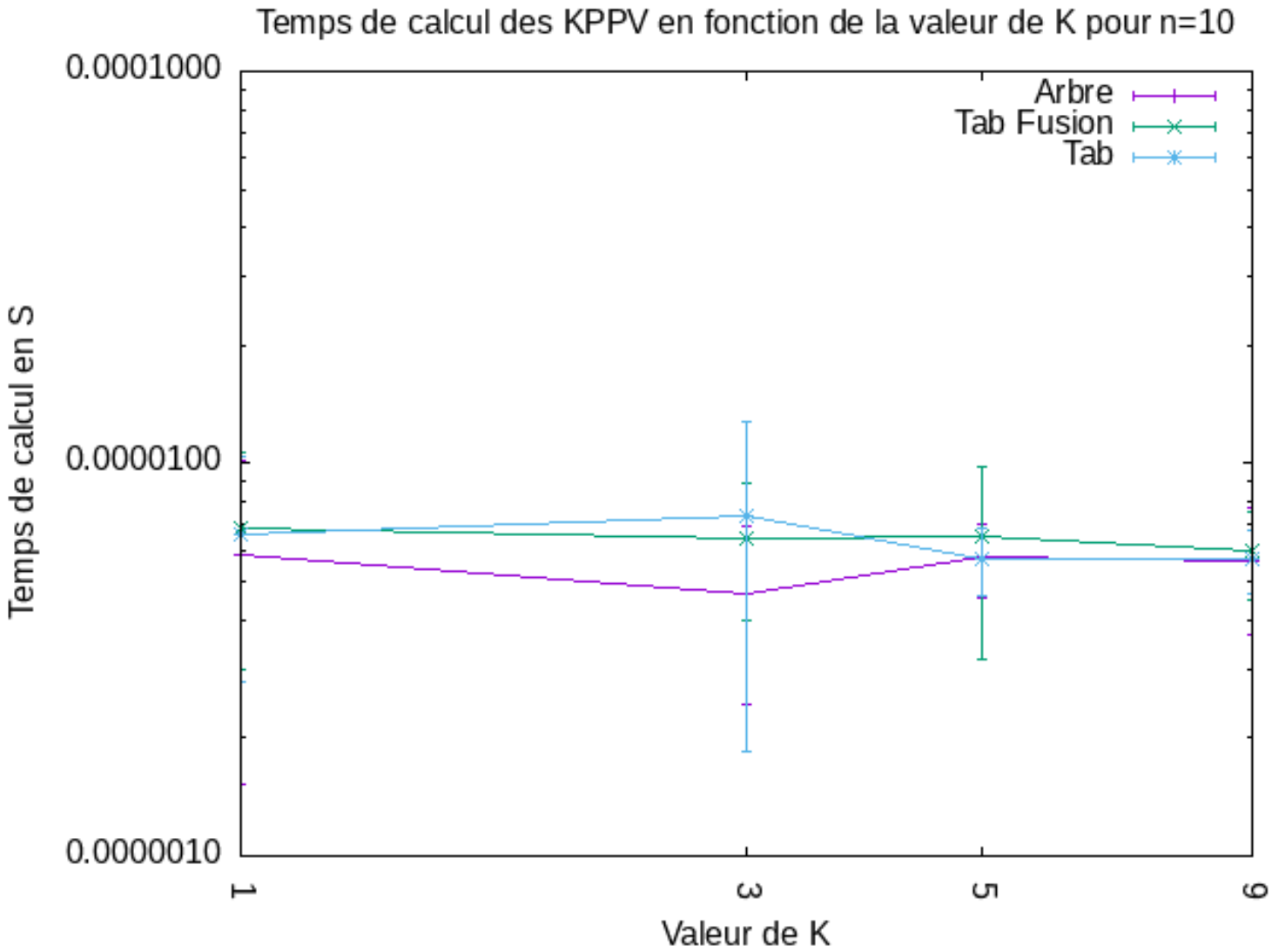

Temps de calcul des KPPV en fonction de la valeur de n pour k=8193
Temps de calcul en S
100
10
1
0.1
0.01
0.001
Arbre
Tab Fusion
Tab
50000
100000
500000
Valeur de N

Conclusion K fixe

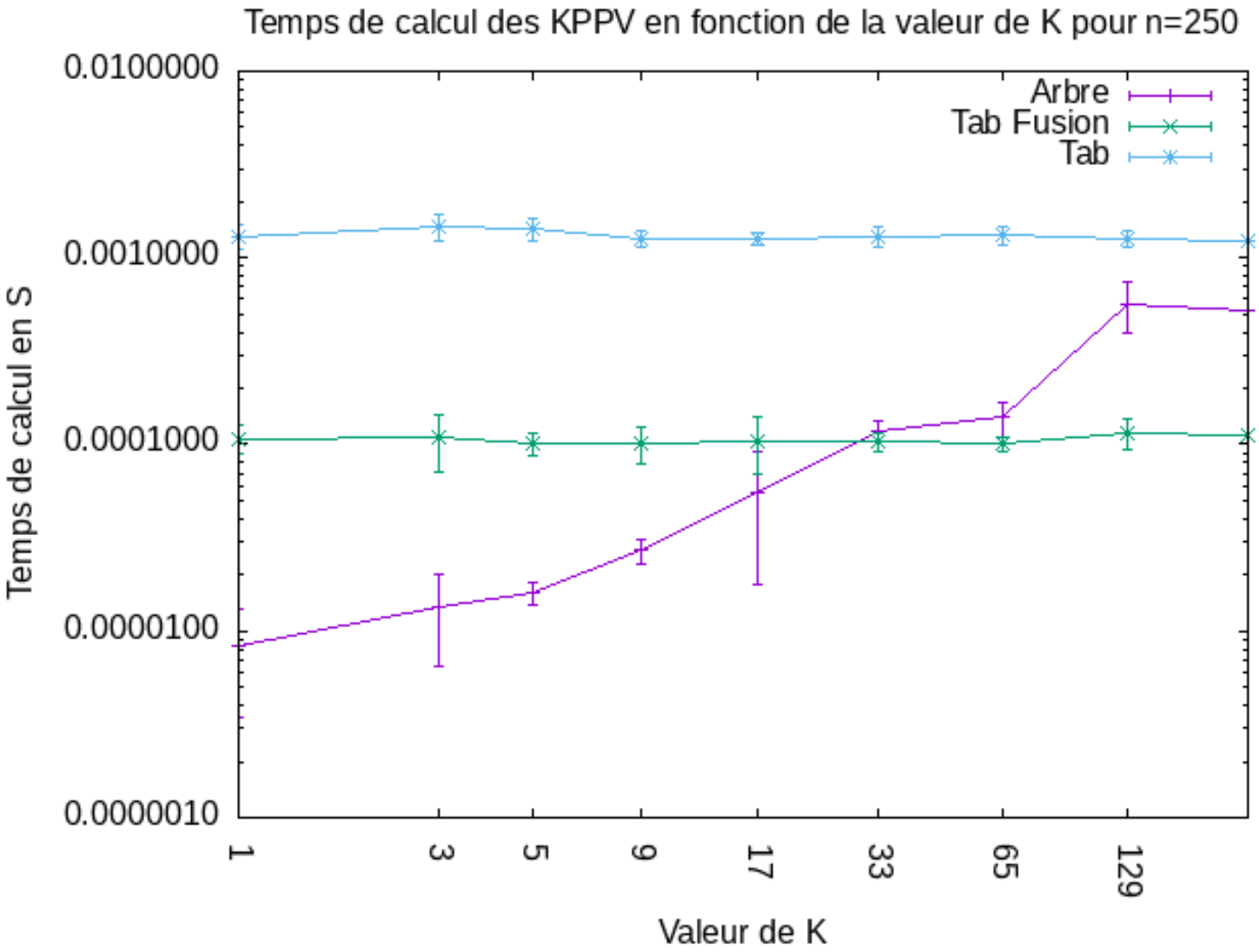
Tri par tableau fusion : plus le nombre de points (n) est élevé, plus le temps de calcul augmente.
Le temps nécessaire pour trier les données augmente de manière significative lorsque le nombre de points augmente.
Méthode moins efficace pour de grands ensembles de données.
Méthode par arbre kd : le temps de calcul augmente légèrement avec le nombre de points (n).
Temps de calcul toujours plus faible pour un k$\langle$10. Donc méthode plus efficace pour un k faible peu importe le nombre de points.
Méthode plus efficace pour des ensembles de données plus importants.
Mais inversion de la rapidité lorsque le k est très grand.

Table des Matières

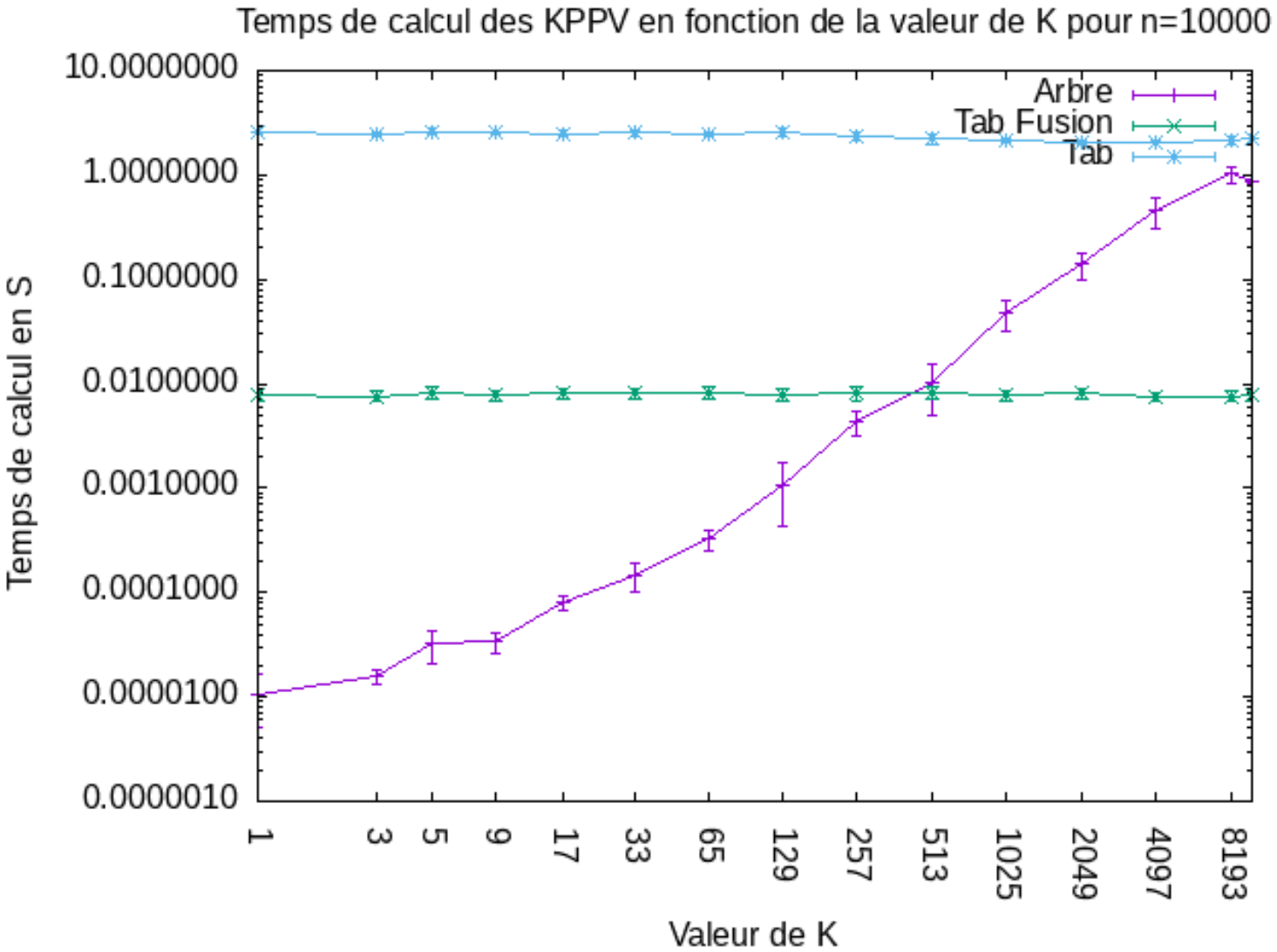
Nombre de points fixe



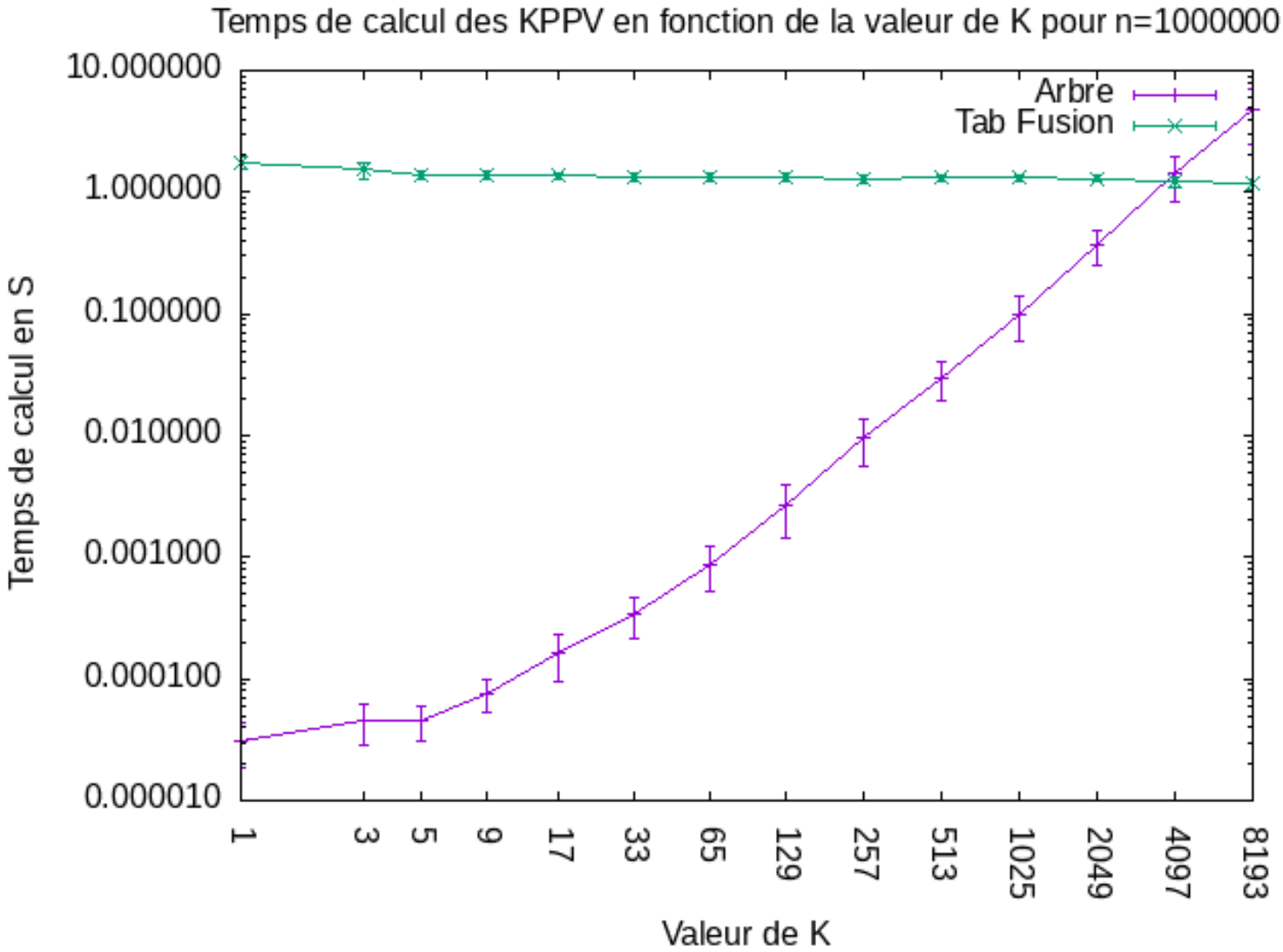
Nombre de points fixe



Nombre de points fixe



Nombre de points fixe



Conclusion pour nombre de points fixe

On voit que le temps de calcul par "tableau" reste constant, peu importe la valeur de K.
La méthode par arbre kd est plus efficace pour un k relativement faible par rapport au nombre de points.
Plus n augmente, plus la valeur de x de l'intersection entre les deux courbes augmente (valeur de k pivot de l'efficacité entre les deux méthodes) car le temps de calcul du tableau augmente.

Table des Matières

Conclusion

- Pour un k relativement faible par rapport au nombre de points, la méthode par "arbre kd" est plus efficace que la méthode par "tableau".
- Grande valeur de n, faible valeur de k : arbre kd.
- Grande valeur de n, grande valeur de k : tableau.
- Faible valeur de n, faible valeur de k : arbre kd.
- Faible valeur de n, grande valeur de k : tableau.
- On voit également sur les courbes que l'écart-type est plus élévé pour la méthode "arbre kd", car fin de parcours de l'arbre plus aléatoire en fonction du placement des points. Donc temps de calcul plus variable que le tableau.