Master MLDS / AMSD

Machine Learning for Data Science

Projet de classification non supervisée

Réalisé par :

SALHI Youssef | AMSD

LABRIJI Saad | MLSD

I. Introduction :

La gestion efficace de la consommation d'énergie constitue un enjeu majeur dans le contexte actuel de préoccupations croissantes en matière de durabilité et de gestion des ressources. Dans cette optique, la disparité frappante de la consommation d'électricité entre les différents ménages est le résultat d'une interconnexion complexe de facteurs, notamment le nombre et l'utilisation spécifique des appareils ménagers, répartis entre le chauffage, l'eau chaude, la cuisson et divers autres équipements électriques. Cette diversité est également influencée par la fréquence et la durée d'utilisation de ces appareils. La nature non stockable de l'électricité à grande échelle souligne l'importance cruciale de maintenir un équilibre constant entre la production et la consommation pour assurer la stabilité du système électrique et prévenir d'éventuelles pannes.

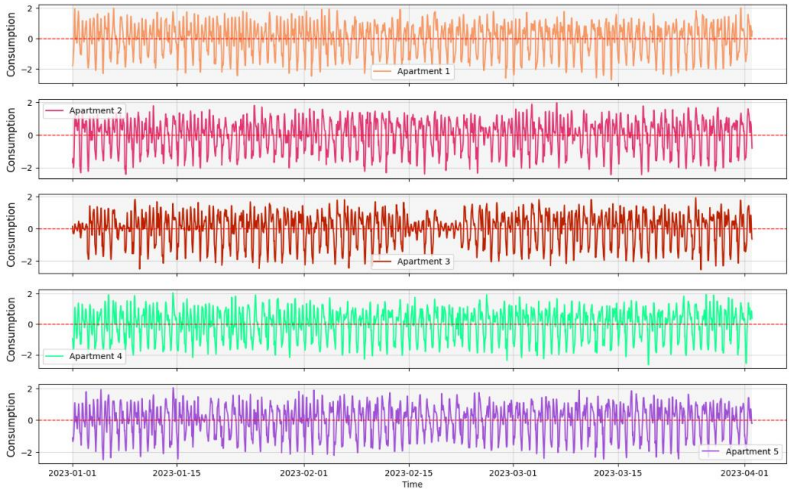
Compte tenu de la variabilité des comportements des consommateurs, le regroupement de ces comportements en catégories homogènes peut offrir des avantages significatifs. Non seulement cela permet de mieux comprendre les habitudes de consommation, mais cela ouvre également la voie à des offres personnalisées et à des prévisions plus précises. En effet, la prévision de la consommation de groupes homogènes facilite la compréhension globale en agrégeant ces prévisions.

Ce projet aborde ces défis en appliquant une classification non supervisée pour segmenter un ensemble de séries temporelles représentant la consommation d'électricité de 100 appartements sur une période de 91 jours consécutifs. L'objectif spécifique est de détecter automatiquement les ruptures dans les séries, symbolisant des changements significatifs dans la consommation d'électricité des appartements. L'objectif particulier est de détecter de manière automatique des ruptures, symbolisant des changements notables dans les habitudes de consommation, et ce, de manière uniforme à travers toutes les séries temporelles. En établissant ces segments homogènes, le projet aspire à apporter des éclairages significatifs sur les dynamiques complexes de la consommation énergétique résidentielle, ouvrant la voie à des stratégies de gestion plus précises et durables.

II. Exploration des données (EDA) :

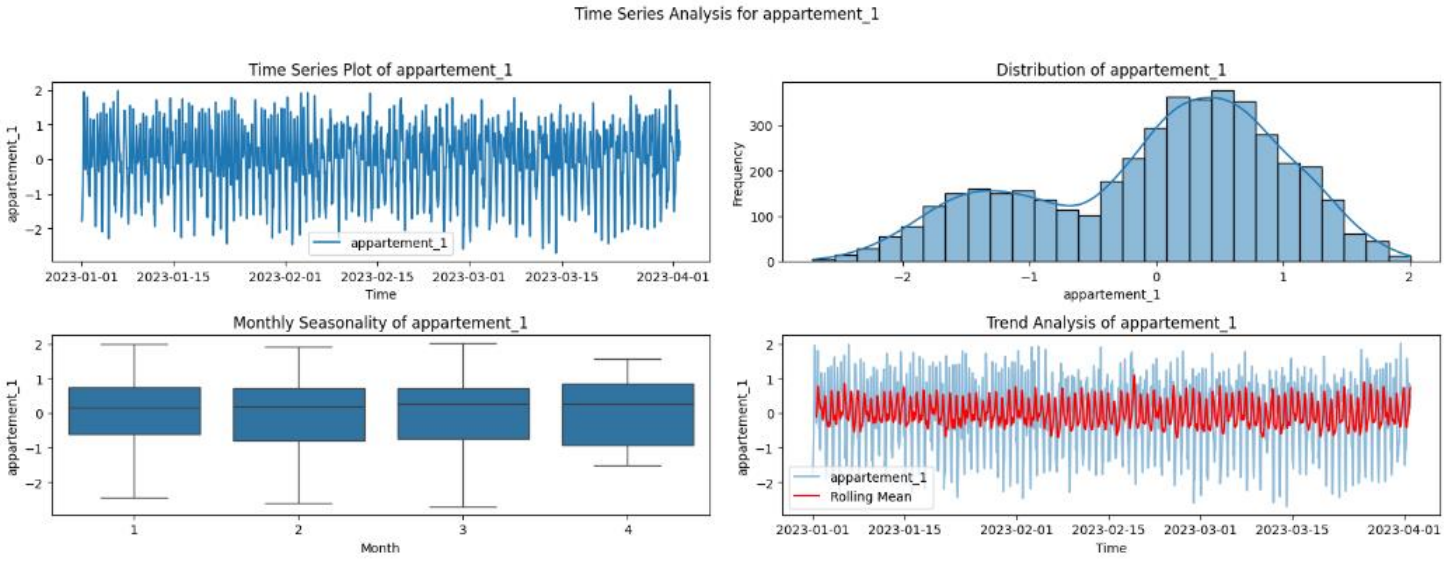
Dans notre analyse de données en Data Science, nous importons les données collectées toutes les 30 minutes sur 91 jours auprès de 100 appartements. Ces données sont structurées en trois tableaux - X, APPART, et JOUR - représentant la consommation énergétique, le numéro de l'appartement, et le numéro du jour. Après l'importation, nous utilisons des graphiques classiques pour explorer les séries temporelles.

1. **Consommation d'Énergie des Appartements 1 à 5 au Fil du Temps :**



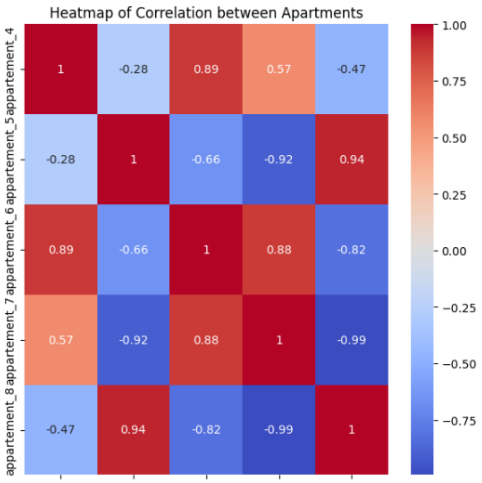
2. **Tendances Temporelles de l'Appartement 1 :**

Pour l'appartement 1, nous examinons la distribution des valeurs de la série temporelle, la saisonnalité avec un boxplot, et l'analyse de tendance avec la moyenne mobile.

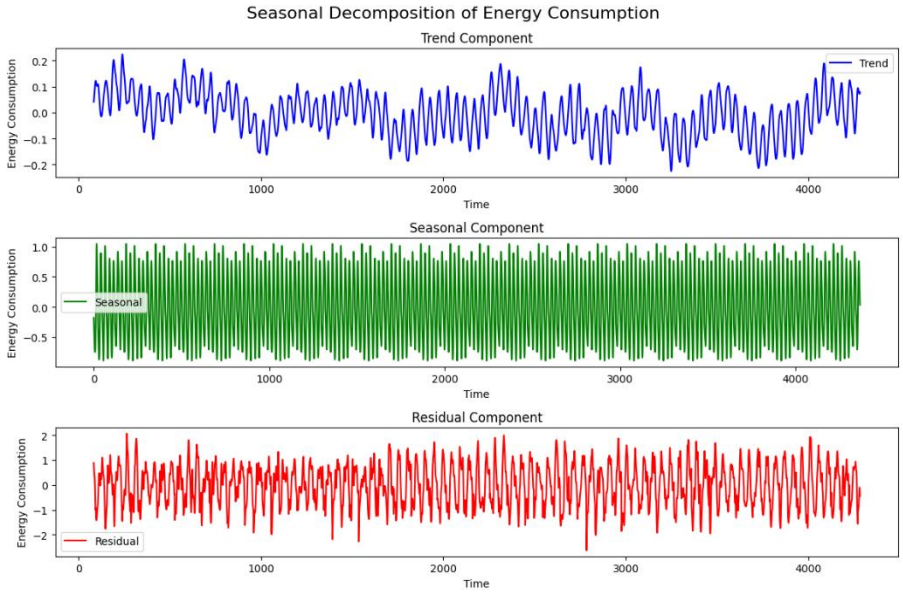


3. **Corrélations entre les Appartements 4 à 8 :**

La matrice de corrélation entre les appartements 4 à 8 permet d'évaluer les corrélations linéaires entre les séries temporelles.



4. Analyse Saisonnière de la Consommation d'Énergie pour l'Appartement 1 :



Ces visualisations fournissent un aperçu des tendances et des motifs saisonniers de la consommation énergétique, de la matrice de corrélation et de la décomposition saisonnière. L'ensemble de ces étapes représente une phase cruciale de notre exploration des données afin de les comprendre plus efficacement.

III. Réduction des Dimensions :

Afin de simplifier la complexité de nos données temporelles tout en préservant leurs caractéristiques essentielles, nous avons élaboré trois approches distinctes :

1. **Analyse en Composantes Principales (ACP) :**

L'intégration de l'Analyse en Composantes Principales (ACP) est cruciale dans notre exploration des données, visant à optimiser la réduction de la variance des séries temporelles de consommation. Malgré une perte partielle d'interprétabilité, cette approche garantit la préservation d'au moins 90 % de la variabilité des données. Après normalisation et consolidation, l'ACP est appliquée avec le paramètre n_components fixé à 0.95, stockant les composantes principales dans un nouveau DataFrame. Cette étape de réduction dimensionnelle, essentielle pour concentrer l'information, facilite la suite de notre analyse. L'objectif final est de réduire les données de chaque foyer à un nombre optimal de caractéristiques expliquant 90 % de la variabilité, en assurant une application efficace de l'ACP grâce à une normalisation caractéristique par caractéristique. Cela garantit que l'analyse ne se limite pas aux moments de grandes différences de consommation brute, mais vise à identifier les variations comportementales à tout moment, évitant également un poids disproportionné sur certaines dates.

2. **Les Caractéristiques de la série temporelle :**

En adoptant cette approche, nous avons la liberté de créer nos propres caractéristiques en utilisant des techniques d'analyse propres aux séries temporelles. Ces caractéristiques, bien qu'interprétables, ne sont pas nécessairement optimisées pour le clustering. Nous pouvons extraire des statistiques descriptives, comme la moyenne, l'écart-type, et d'autres mesures de tendance centrale et de dispersion, pour chaque série temporelle afin de saisir les aspects les plus significatifs de nos données.

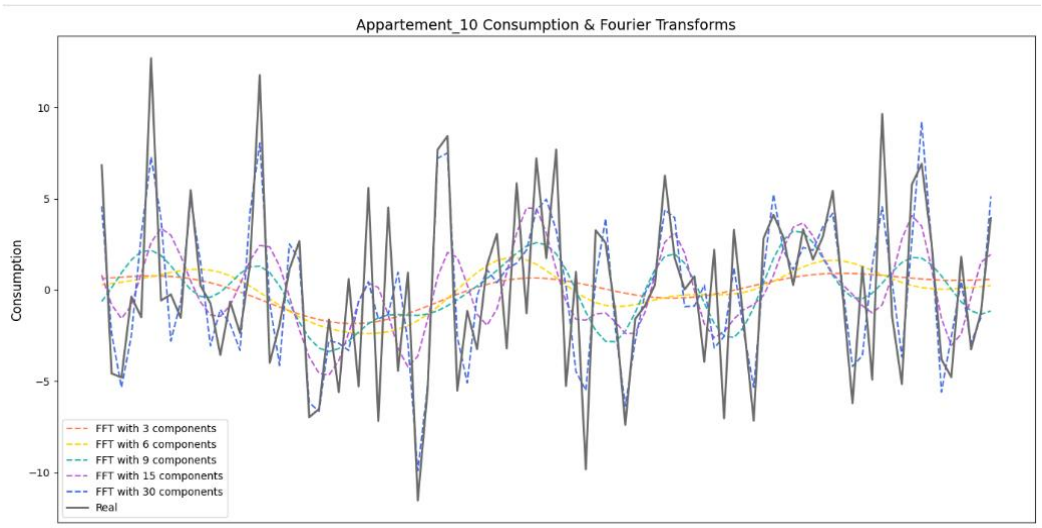
	Mean	Minimum	Maximum	Month_Max_Consumption	Month_Min_Consumption
appartement_1	-0.00253	-2.712846	2.007661	1	3
appartement_2	0.004413	-2.447783	1.964942	2	1
appartement_3	0.0155	-2.551442	1.927615	1	2
appartement_4	0.006204	-2.653337	2.054251	3	1
appartement_5	-0.01502	-2.467466	2.030829	3	1
...	...	...	...	...	...
appartement_96	0.012919	-2.379652	2.264872	3	4
appartement_97	-0.003102	-2.649557	2.107669	3	1
appartement_98	-0.006116	-2.731046	2.094527	1	3
appartement_99	0.005128	-2.532689	1.971988	3	2
appartement_100	0.005883	-2.87689	1.961424	3	1

Dans cette approche, nous nous écartons de l'Analyse en Composantes Principales (ACP) en créant des variables interprétables pour chaque individu, telles que la moyenne, le minimum ou le maximum de sa consommation. Nous travaillons avec des données non normalisées par individu, estimant pertinent de conserver des valeurs brutes pour mieux mettre en évidence les différences

entre les consommateurs. De plus, la plupart des caractéristiques que nous allons reconstituer sont basées sur des variations brutes, notamment la consommation moyenne de chaque individu.

3. Transformation de Fourier :

En appliquant la Transformation de Fourier, nous extrayons les caractéristiques fréquentielles de nos séries temporelles, représentant ainsi nos données par des coefficients fréquentiels. Cette approche permet la détection de motifs ou de tendances périodiques, avec la possibilité de sélectionner les coefficients les plus pertinents pour représenter les informations significatives. L'objectif de ces méthodes est de réduire la dimensionnalité des données, préservant un haut niveau d'interprétabilité. Cette approche vise à faciliter un clustering efficace tout en fournissant une compréhension approfondie des disparités de comportement au sein de nos données temporelles. Dans la figure ci-dessous, nous pouvons visualiser que la composante spectrale de 30 est la plus proche de la consommation réelle : c'est ainsi que nous avons considéré d'avancer avec 30



.

IV. Clustering :

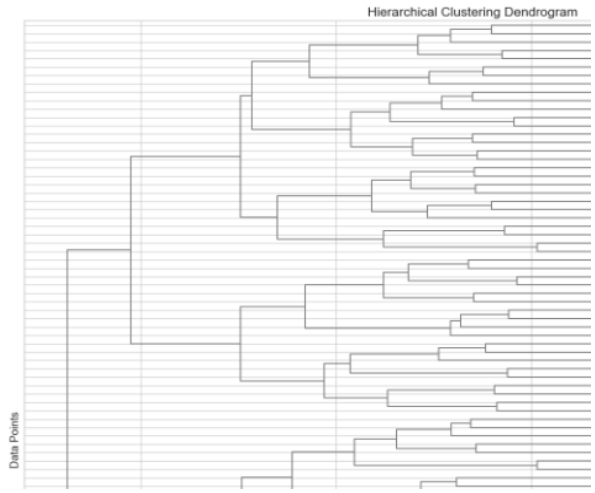
Dans la phase de Clustering, nous avons appliqué la méthode de regroupement hiérarchique, une approche robuste visant à découvrir des structures intrinsèques au dataset. Le regroupement hiérarchique offre la souplesse de déterminer le nombre de clusters en fonction des besoins, offrant ainsi des indications sur la configuration sous-jacente des données.

4.1 Regroupement Hiérarchique :

Nous avons adopté deux approches principales dans le cadre du regroupement hiérarchique : les méthodes basées sur l'extraction de caractéristiques et une méthode directe.

4.1.1 Méthode basée sur l'extraction de caractéristiques :

Pour évaluer la qualité de nos résultats, nous avons utilisé l'indice de Calinski-Harabasz, une métrique permettant de mesurer l'efficacité du regroupement. La fonction `**CAH**` a été mise en œuvre pour effectuer une analyse de regroupement hiérarchique sur des données réduites, fournissant un dendrogramme pour visualiser la structure des clusters.

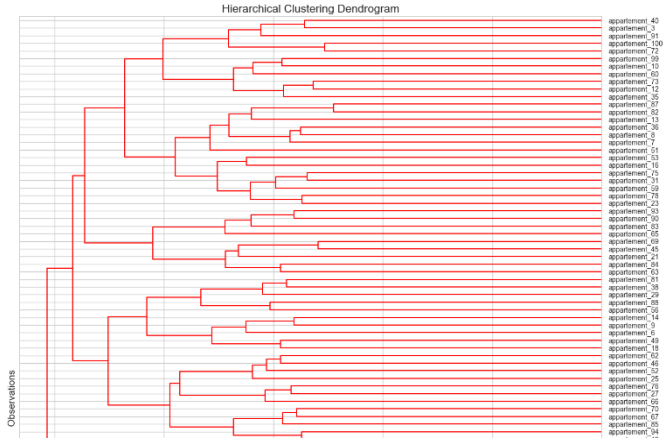


Cette Figure présente le dendrogramme résultant de l'application de la méthode hiérarchique. Ce dendrogramme a guidé le choix du nombre optimal de clusters.

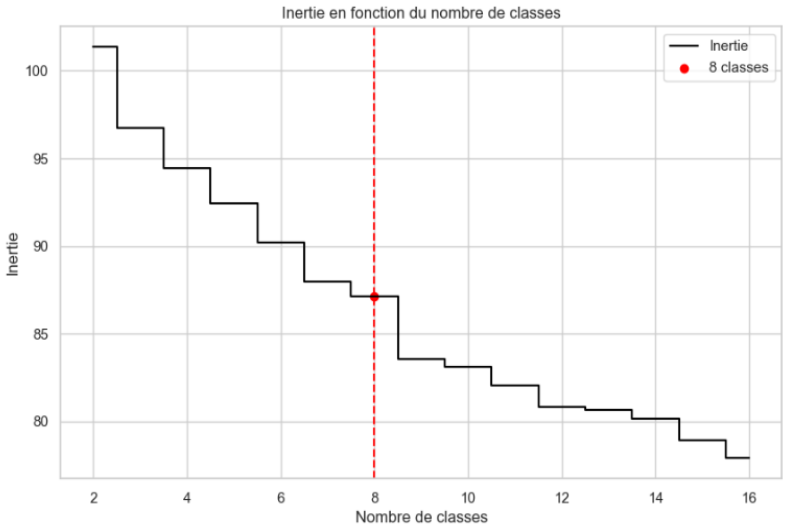
Le dendrogramme est un diagramme en forme d'arbre qui montre comment les éléments sont regroupés ou divisés en fonction de leurs similarités. C'est une représentation visuelle simple des relations hiérarchiques dans un ensemble de données.

4.1.2 Méthode Directe :

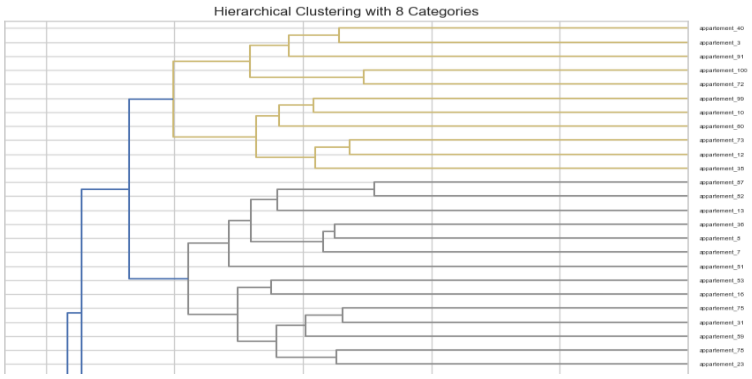
La méthode directe implique la génération d'une matrice de liaison et la visualisation du dendrogramme de regroupement hiérarchique. Nous avons défini un seuil de différenciation des couleurs pour faciliter l'identification de clusters distincts.



Cette figure illustre le dendrogramme avec un seuil de différenciation des couleurs. La ligne verticale indique le point de coupe optimal pour définir le nombre de clusters. L'inspection visuelle des dendrogrammes a révélé des sauts significatifs d'inertie pour 2 et 8 clusters, suggérant des choix de segmentation pertinents.



L'inspection visuelle des dendrogrammes a révélé des sauts significatifs d'inertie pour 2 et 8 clusters, suggérant des choix de segmentation pertinents.



Cette figure présente le graphique d'inertie en fonction du nombre de clusters. Deux sauts notables sont observés pour 2 et 8 clusters, suggérant une pertinence particulière pour ces configurations.

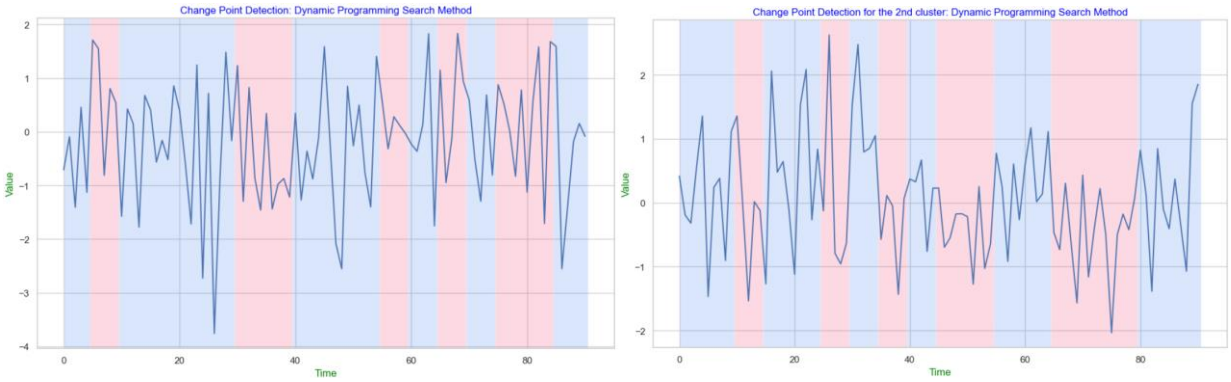
En conclusion, le regroupement hiérarchique a fourni des indications cruciales sur la structure des données, orientant le choix du nombre optimal de clusters. Les visualisations, notamment les dendrogrammes et les graphiques d'inertie, ont constitué des outils essentiels pour interpréter les résultats et guider les étapes ultérieures de l'analyse.

V. Segmentation :

Dans cette section, nous aborderons le post-traitement des clusters obtenus précédemment. Nous introduirons un signal bruyant constant par morceaux, puis nous effectuerons une détection des points de changement en utilisant un noyau pénalisé. Les résultats seront présentés avec des couleurs alternées indiquant les régimes réels et des lignes pointillées marquant les points de changement estimés.

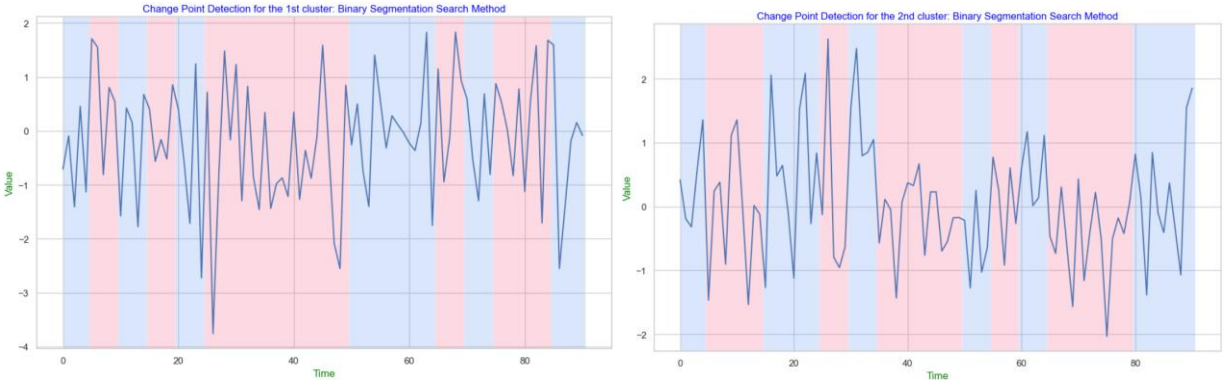
1. Algorithme de Programmation Dynamique :

Nous utiliserons une méthode exacte basée sur la programmation dynamique, avec un coût de calcul considérable de O(Qn^2), où Q est le nombre maximal de points de changement et "n" est le nombre de points de données.



2. Méthodes de Recherche par Segmentation Binaire :

Une approche bien établie dans la littérature est la segmentation binaire, une méthode approximative avec un coût de calcul efficace de O(n log n), où "n" est le nombre de points de données. L'algorithme fonctionne en appliquant itérativement une méthode de point de changement unique à l'ensemble de la séquence pour détecter les divisions. En cas de détection, la séquence est divisée en deux sous-séquences, et le processus est répété pour chaque sous-séquence.



Les graphiques ci-dessus illustrent que les points de changement détectés dans la séquence varient en fonction de la méthode de recherche utilisée.

VI. Conclusion :

En conclusion, notre exploration des techniques d'apprentissage non supervisé a considérablement enrichi notre compréhension des données de consommation électrique des ménages. À travers l'application de méthodes telles que le clustering, la segmentation, et le Binary Segmentation Search, nous avons réussi à mettre en lumière des schémas et des comportements jusque-là imperceptibles.

Les résultats obtenus grâce à l'apprentissage non supervisé ont permis de regrouper les consommateurs en fonction de leurs habitudes de consommation, de segmenter les profils électriques pour identifier des tendances spécifiques, et de détecter des points de changement significatifs. Ces informations revêtent une importance cruciale pour appréhender les dynamiques de consommation, anticiper les variations de la demande, et concevoir des stratégies d'efficacité énergétique ciblées.

L'utilité démontrée par l'apprentissage non supervisé s'étend au-delà de la simple révélation d'insights précieux. En effet, ces techniques facilitent une prise de décision éclairée dans le domaine de l'énergie, offrant ainsi une opportunité d'optimiser les pratiques durables, de réduire les coûts énergétiques, et de contribuer activement à la préservation de l'environnement.

Notre projet a également souligné l'importance de l'application de méthodes de classification non supervisée sur des données temporelles. En manipulant les séries temporelles à différentes échelles, en recourant à des techniques de réduction de dimension et d'extraction de caractéristiques, et en mettant en œuvre des méthodes de clustering, nous avons réussi à obtenir des résultats significatifs. De plus, la clarification des objectifs de recherche s'est avérée être une étape cruciale, permettant une approche méthodique et efficace dans l'analyse des données temporelles complexes. En somme, cette exploration a ouvert la voie à une utilisation plus judicieuse des données de consommation électrique, favorisant des initiatives durables et une gestion éclairée de l'énergie.