

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Modélisation de la variabilité d'un réseau de transport collectif

ROMARIO ANDRÉ JUNIOR PHILIBERT

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Octobre 2024

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Modélisation de la variabilité de l'utilisation d'un réseau de transport collectif

présenté par **Romario André Junior PHILIBERT**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Bruno AGARD, président

Martin TRÉPANIÉ, membre et directeur de recherche

Catherine MORENCY, membre et codirectrice de recherche

Samira KEIVANPOUR, membre

DÉDICACE

À ma mère, ma complice de toujours.

À mon père, combien il est dur de te dépasser.

Je deviendrai professeur un autre jour,

Mais ton exemple reste ma plus grande leçon.

À mon amie et partenaire, Frantzlande, qui

supporte ma boulimie d'apprendre.

"Nous aimons parce que c'est la seule véritable

aventure." – Nikki Giovanni

À mes petites sœurs, Pascale et Dominique,

"Les sœurs sont des fleurs différentes du même

jardin." – Auteur inconnu

À la vie, parce que nous venons de loin.

REMERCIEMENTS

À mon directeur de mémoire, Pr Martin Trépanier, qui ne s'est jamais découragé malgré mes longs séjours sur la lune. J'ai finalement accouché de ce travail grâce à votre patience et votre soutien inébranlables.

À ma codirectrice, Pre Catherine Morency, dont la passion et l'enthousiasme lorsqu'elle expose sont une véritable source d'inspiration. J'espère un jour trouver ma voie avec une telle ferveur.

À Polytechnique Montréal, pour sa qualité et son exigence. J'aurais aimé y séjourner plus longtemps.

À Vincent Cadoret, qui a cru en moi et m'a offert l'opportunité d'aller plus loin. Votre confiance a été déterminante dans mon parcours.

À Mitacs et Kéolis, pour leur générosité et leur soutien, qui ont été précieux tout au long de cette aventure.

Aux quelques camarades avec lesquels j'ai eu l'occasion d'échanger et de partager, j'espère rattraper mon silence et ma distance. Merci de m'avoir accueilli et soutenu.

À ces regards bienveillants, ces épaules et ces mains qui m'ont empêché de me noyer, merci du fond du cœur.

"La gratitude est la mémoire du cœur." – Jean-Baptiste Massillon

RÉSUMÉ

Des horaires bien développés et régulièrement entretenus garantissent une prestation de services adéquate en fonction de la demande, réduisant ainsi les temps d'attente des passagers et optimisant l'efficacité opérationnelle. Une planification efficace garantit également une utilisation optimale des véhicules, évitant ainsi les surcharges ou les périodes de sous-utilisation. Cet aspect est d'une importance capitale en termes de coûts opérationnels, notamment en ce qui concerne les dépenses de main d'œuvre sur le terrain, qui représentent une part importante des coûts d'exploitation des entreprises de transport.

Historiquement, les sociétés de transport ont tendance à créer des horaires fixes pour les jours de la semaine, du lundi au vendredi. Cependant, cette approche est-elle toujours pertinente face aux bouleversements récents, tels que ceux apportés par la pandémie et autres changements sociétaux? Ces événements ont modifié les schémas de déplacement et les habitudes des usagers. Par exemple, avec l'adoption croissante du télétravail, les déplacements domicile-travail ont drastiquement diminué, entraînant une baisse de la demande de transport aux heures de pointe et une augmentation de la flexibilité des horaires de déplacement.

Ce projet vise à appliquer des méthodes de segmentation des données pour évaluer si les horaires actuels du transport collectif à Orléans sont en adéquation avec la réalité des déplacements. En exploitant les données issues des cartes à puce des usagers (collectées grâce aux systèmes d'aide à l'exploitation et à l'information aux voyageurs (SAEIV), on peut ajuster les horaires pour qu'ils reflètent plus fidèlement la demande actuelle. Cela pourrait impliquer des modifications telles que l'introduction d'horaires spécifiques pour certains jours de la semaine, l'ajustement des fréquences en fonction des périodes de forte et de faible affluence, et la mise en place de services supplémentaires lors d'événements spéciaux. Un bon horaire, défini par sa capacité à répondre aux besoins des passagers tout en minimisant les coûts, est crucial pour le succès et la durabilité des services de transport en commun.

En se basant sur les montées (nombre de validations) des passagers, l'algorithme des K-moyennes et la classification hiérarchique ont été appliqués directement sur divers objets du réseau d'Orléans. Ensuite, ces algorithmes ont été combinés à la méthode DTW et appliqués à dix transformations des données en utilisant quatre distances différentes. Enfin, par le biais d'indicateurs et étapes successives, un modèle final est retenu à la fin de chaque itération pour classer les données. Enfin,

des statistiques finales sont calculées afin d'affiner les conclusions sur les groupes obtenus. Ces étapes ont permis de confirmer la robustesse de la classification la plus simple « type de jour »: jours de semaine, samedi, dimanche et jours fériés.

Dans certains cas, le volume de transport de passagers selon les jours de la semaine de travail est suffisamment différent pour justifier l'établissement d'horaires pour plusieurs jours de la semaine afin de réaliser une économie d'horaire maximale. Ainsi, il peut y avoir un horaire pour le lundi uniquement, couplé à un autre horaire du mardi au vendredi, un autre pour le samedi et encore un autre pour le dimanche.

Ce travail a permis aussi de démontrer que les transformations par quantile normal et uniforme sont des outils puissants pour améliorer la performance des algorithmes de classification : elles aident à normaliser les données en présence de distributions fortement asymétriques et réduisent les variances extrêmes et valeurs aberrantes. La méthode "statistique de l'écart K" suggère, indépendamment des méthodes de segmentation et des métriques de distance, plus de groupes notamment pour les situations où les groupes ne sont pas bien séparés. De plus, la méthode DTW en général, plus particulièrement combinée à la segmentation hiérarchique, en dépit de sa complexité computationnelle, fournit des résultats robustes même si les séries ne sont pas parfaitement alignées et fournit un dendrogramme qui peut être facilement interprété pour comprendre les relations entre les groupes.

Enfin, l'application de ces méthodes à des objets démontrent aussi qu'il n'existe aucune méthode universelle i-e qu'il faut aller au cas par cas afin de capturer les similitudes et différences au sein d'un mode ou groupe de lignes. Une grande marge peut aussi être laissée au modélisateur qui peut progressivement adapter les outils (techniques de segmentation, métriques) selon l'objectif et la précision poursuivis (généralisation ou particularisation pour un ou plusieurs objets).

En conclusion, en utilisant des algorithmes de segmentation pour analyser les données de cartes à puce, il est possible d'identifier des groupes distincts de jours avec des profils de demande en transport variés. Ces méthodes ont mis en lumière l'importance de personnaliser les horaires en fonction de ces tendances, plutôt que d'appliquer un modèle uniforme. En ajustant les horaires en conséquence, il est possible d'améliorer l'efficacité du réseau de transport collectif, l'expérience des usagers, réduire les coûts, soutenant ainsi les objectifs de durabilité urbaine.

ABSTRACT

Well-developed and regularly maintained schedules ensure an adequate level of service based on demand, thereby reducing passenger wait times and optimizing operational efficiency. Effective planning also guarantees optimal vehicle utilization, preventing overcrowding or periods of underutilization. This aspect is crucial in terms of operational costs, particularly concerning labor expenses, which constitute a significant portion of transport companies' operating costs.

Historically, transport companies have tended to create fixed schedules for weekdays, from Monday to Friday. However, is this approach still relevant considering recent upheavals, such as those brought about by the pandemic and other societal changes? These events have altered travel patterns and user habits. For example, with the increasing adoption of telecommuting, home-to-work travel has drastically decreased, leading to a decline in peak-hour transport demand and increased flexibility in travel schedules.

This project aims to apply data segmentation methods to evaluate whether the current public transport schedules in Orléans align with actual travel patterns. By leveraging data from users' smart cards (collected through passenger information and operations support systems, PISA), schedules can be adjusted to more accurately reflect current demand. This might involve modifications such as introducing specific schedules for certain days of the week, adjusting frequencies according to periods of high and low demand, and implementing additional services during special events. A good schedule, defined by its ability to meet passengers' needs while minimizing costs, is crucial for the success and sustainability of public transport services.

Based on passenger boardings (number of validations), the K-means algorithm and hierarchical classification, applied directly and combined with DTW on various network objects in Orléans, have been utilized. Ten transformations and four different distances were used. Through successive indicators and steps, a final model was selected at the end of each iteration to classify the data. Final statistics were then calculated to refine conclusions about the obtained groups. These steps confirmed the robustness of the simplest classification by "type of day": weekdays, Saturdays, Sundays, and holidays.

In some cases, the volume of passengers varies sufficiently by day of the workweek to justify establishing schedules for several days to achieve maximum schedule economy. Thus, there may

be a schedule for Monday only, coupled with another schedule for Tuesday through Friday, another for Saturday, and yet another for Sunday.

This work also demonstrated that quantile transformations (normal and uniform) are powerful tools for improving the performance of classification algorithms: they help normalize data with highly skewed distributions, reducing extreme variances and outliers. The "K gap statistic" method independently suggests more clusters for situations where clusters are not well separated, regardless of the methods and distance metrics used. Furthermore, the DTW method, particularly when combined with hierarchical clustering despite its computational complexity, provides robust results even if the series are not perfectly aligned and offers a dendrogram that can be easily interpreted to understand the relationships between clusters.

Finally, the application of these methods to various objects also shows that there is no "one size fits all" approach; it is necessary to proceed on a case-by-case basis to capture similarities and differences within a mode or group of lines. A considerable margin can also be left to the modeler, who can progressively adapt the methods (segmentation techniques, metrics) according to the objective and precision pursued (generalization or particularization for one or more objects).

In conclusion, using segmentation algorithms to analyze smart card data makes it possible to identify distinct groups of days with varying transport demand profiles. These methods highlight the importance of customizing schedules based on these trends rather than applying a uniform model. By adjusting schedules accordingly, it is possible to improve the efficiency of the public transport network, enhance the user experience, reduce costs, and thereby support urban sustainability goals.

TABLE DES MATIÈRES

DÉDICACE.....	III
REMERCIEMENTS	IV
RÉSUMÉ.....	V
ABSTRACT	VII
TABLE DES MATIÈRES	IX
LISTE DES TABLEAUX.....	XII
LISTE DES FIGURES.....	XIII
LISTE DES SIGLES ET ABRÉVIATIONS	XVI
LISTE DES ANNEXES.....	XVIII
CHAPITRE 1 INTRODUCTION.....	1
1.1 Mise en contexte.....	3
1.2 Problématique.....	4
1.3 Objectif.....	7
1.4 Structure du mémoire	8
CHAPITRE 2 REVUE DE LITTÉRATURE	10
2.1 Conception des horaires	10
2.2 Carte à puce en transport collectif.....	11
2.3 Séries temporelles en transport	20
2.4 Séries temporelles des données de comptage et de cartes à puce dans les transports....	21
2.5 Classification des données de séries temporelles	26
2.5.1 Approches basées sur les données brutes	27
2.5.2 Approches basées sur l'extraction de caractéristiques	28
2.5.3 Modèles et algorithmes de segmentation	31

2.5.4	Complément : Approche COTE (Collection of Transformation Ensembles)	34
2.6	Synthèse de la revue	36
CHAPITRE 3 MÉTHODOLOGIE ET DESCRIPTION DES DONNÉES		39
3.1	Méthodologie de recherche	39
3.2	Cas d'étude	42
3.3	Données	45
3.3.1	Données de transaction de paiement	45
3.3.2	Données GTFS	46
3.3.3	Données météorologiques et calendrier des jours fériés	49
3.4	Préparation des données	50
3.4.1	Pré-traitement des données	50
3.4.2	Composition des vecteurs pour l'analyse de la demande	52
3.5	Analyse descriptive	53
3.6	Segmentation	57
3.6.1	Méthodes de transformation des données	58
3.6.2	Méthodes de segmentation	62
3.6.3	Métriques de distance	65
3.6.4	Validation des résultats de la segmentation	67
3.6.5	Choix de la segmentation optimale	69
3.6.6	Analyse détaillée des groupes	73
3.7	Application	75
CHAPITRE 4 PRÉSENTATION DES RÉSULTATS		76
4.1	Analyse descriptive des données	76
4.2	Analyse détaillée des résultats de la segmentation sur l'année 2020	88

4.2.1	La méthode des K-moyennes (« <i>K-means</i> »)	89
4.2.2	La segmentation hiérarchique	101
4.2.3	Comparaison des résultats du K-moyennes et de la segmentation hiérarchique..	109
4.2.4	La méthode DTW combiné au K-moyennes et segmentation hiérarchique.....	110
4.2.5	Agrégation demi-horaire	111
4.3	Segmentation sur le trimestre d'automne 2020	113
4.3.1	L'ensemble du réseau.....	114
4.3.2	Par mode.....	115
4.4	Segmentation par ligne.....	118
4.4.1	Lignes de tram A et B	118
4.5	Segmentation par arrêt	120
4.5.1	Les arrêts de tram	121
4.6	Synthèse des résultats obtenus des différentes approches.....	122
4.6.1	Comparaison des résultats des quatre algorithmes.....	122
CHAPITRE 5 CONCLUSION		124
5.1	Contributions	124
5.2	Limitations	125
5.3	Perspectives	126
RÉFÉRENCES.....		127
ANNEXES		132

LISTE DES TABLEAUX

Tableau 2-1 - Algorithmes de segmentation sur les données brutes	28
Tableau 2-2- Segmentation basé sur les modèles.....	33
Tableau 3-1- Informations recueillies à chaque validation tarifaire du réseau TAO (à l'embarquement)	45
Tableau 3-2 - Validations de l'année 2020	51
Tableau 3-3- Composition du jeu de données	51
Tableau 4-1 - Vecteur "jour-heure"	88
Tableau 4-2 - 10 meilleurs résultats sur la segmentation de l'année 2020 (K-moyennes)	89
Tableau 4-3 - Évaluation des trois meilleurs groupements K-moyennes.....	92
Tableau 4-4 - 10 meilleurs résultats de la segmentation hiérarchique	101
Tableau 4-5 - Trois meilleurs résultats de la segmentation hiérarchique.....	104
Tableau 4-6- Évaluation des trois meilleurs groupements	104
Tableau 4-7 - Vecteur "demi-heure"	111
Tableau 4-8 - Résultats des 4 méthodes	112
Tableau 4-9 - Validations par demi-heure pour le réseau, trimestre d'automne 2020	114
Tableau 4-10 - Résultats des 4 méthodes - réseau, automne.....	114
Tableau 4-11 - Résultats des 4 méthodes - tramways, automne	115
Tableau 4-12 - Résultats des 4 méthodes - bus, automne	117
Tableau 4-13 - Résultats Tram A et B.....	120

LISTE DES FIGURES

Figure 2-1 - Résumé des approches majeures en classification des séries temporelles	27
Figure 2-2 - Approches basées sur l'extraction de caractéristiques	29
Figure 2-3 – Une taxonomie de la classification des séries temporelles basée sur la distance	36
Figure 3-1 - Méthodologie générale du mémoire.....	39
Figure 3-2 - Plan du réseau d'Orléans (France).....	42
Figure 3-3 - Validations totales sur le réseau TAO en 2020	43
Figure 3-4 - Schéma simplifié d'un fichier GTFS	46
Figure 3-5 - Intensité d'utilisation des arrêts à Orléans (Données GTFS)	48
Figure 3-6 - Étape de prétraitement des données	50
Figure 3-7 - Transformations appliquées aux vecteurs	59
Figure 3-8 - Différentes segmentations appliquées aux données	63
Figure 3-9 - Critères d'évaluation des segmentions	67
Figure 3-10- Sélection du groupement optimal.....	70
Figure 3-11 - Évaluation détaillée du groupe final	73
Figure 4-1 - Validations mensuelles totales sur le réseau TAO en 2020	76
Figure 4-2 - Validations mensuelles par type de ligne du réseau TAO en 2020.....	77
Figure 4-3 - Validations mensuelles des dix lignes les plus fréquentées du réseau TAO en 2020...	78
Figure 4-4 - Validations par type de transactions sur le réseau TAO en 2020	79
Figure 4-5 - Les dix (10) arrêts les plus fréquentés sur le réseau TAO en 2020.....	80
Figure 4-6 - Les vingt (20) arrêts les plus fréquentés sur l'ensemble du réseau TAO en 2020	80
Figure 4-7 - Pourcentage de validations pour chaque jour de la semaine par type de ligne du réseau TAO en 2020	82
Figure 4-8 - Validations par jour, segment de la journée et saison sur le réseau TAO en 2020 ...	82

Figure 4-9 - Validations par jour sur le réseau TAO en 2020	83
Figure 4-10 - Histogramme des validations par arrêt pour le réseau TAO en 2020	83
Figure 4-11 - Histogramme des validations par arrêt pour la ligne de tram A du réseau TAO en 2020	84
Figure 4-12 - Histogramme des validations par arrêt pour les lignes de proximité du réseau TAO en 2020	84
Figure 4-13 - Histogramme des validations par arrêt pour les lignes structurantes du réseau TAO en 2020	84
Figure 4-14 - Validations mensuelles sur le réseau TAO en 2020.....	85
Figure 4-15 - Validations pendant les jours de la semaine sur le réseau TAO en 2020	85
Figure 4-16 - Validations par segment sur le réseau TAO en 2020	87
Figure 4-17 - Scores de silhouette pour la transformation standard K-moyennes	90
Figure 4-18 - Scores SSE pour les trois meilleures transformations K-moyennes	91
Figure 4-19 - K-Gap pour les trois meilleurs résultats du K-moyennes	92
Figure 4-20 - Groupes pour la transformation standard K-moyennes	94
Figure 4-21 - Corrélations pour le groupe 0, transformation standard (K-moyennes)	95
Figure 4-22 – Densité pour 5 heures, transformation standard (K-moyennes)	96
Figure 4-23 - Box plot des groupes obtenus K-moyennes - Standard	97
Figure 4-24 - Calendrier des groupes K-moyennes, standard	98
Figure 4-25 - Moyenne des validations par heure par segment de la journée par groupe - K-moyennes, standard	100
Figure 4-26 - Dendrogramme de la transformation standard	102
Figure 4-27 - Dendrogramme de la transformation quantile uniforme	103
Figure 4-28 - Dendrogramme transformation quantile normale	103
Figure 4-29 - Groupes de la transformation finale standard	105

Figure 4-30 - Carte des corrélations du groupe 1 de la segmentation hiérarchique - transformation standard	106
Figure 4-31 - Densité à 5 heures du matin pour la segmentation hiérarchique, transformation standard	107
Figure 4-32 - Calendrier résultat de la segmentation hiérarchique	108
Figure 4-33 - DTW combiné au K-moyennes.....	110
Figure 4-34 - DTW combiné à la segmentation hiérarchique	111
Figure A-1 - Histogramme des validations de la ligne B du Tram	133
Figure A-2 - Histogramme des validations des lignes de Bus standards	133
Figure A-3 - Distribution du pourcentage de validations par segment de la journée des lignes de tram	133
Figure A-4 - Distribution en pourcentage des validations par segment de la journée des lignes structurantes	134
Figure A-5 - Les 20 arrêts les plus fréquentés du réseau TAO en 2020	134

LISTE DES SIGLES ET ABRÉVIATIONS

ABM	<i>Agent-Based Model</i>
ABSM	<i>Activity-Based Simulation Model</i>
AFC	<i>Automatic Fare Collection</i>
APC	<i>Automatic Passenger Counting</i>
AR	Modèle Autorégressif
ARIMA	Modèle Autorégressif Intégré à Moyenne Mobile
ARTM	Autorité Régionale de Transport Métropolitain
AVL	<i>Automatic Vehicle Location</i>
COTE	<i>Collection of Transformation Ensembles</i>
COVID-19	<i>Coronavirus Disease 2019</i>
CSV	<i>Comma-Separated Value</i>
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i>
EM Learning	<i>Expectation-Maximization Learning</i>
EMV	Europay, Mastercard, Visa
GPS	<i>Global Positioning System</i>
HMM	<i>Hidden Markov Model</i>
IRSG	<i>Initial Route Set Generation</i>
K-moyennes	Un algorithme de segmentation en fonction de la similarité
K-NN	<i>K-Nearest Neighbors</i>
LATM	Mémoire à long et à court terme
MAPE	<i>Mean Absolute Percentage Error</i>
OD	Origine-Destination
SARIMA	<i>Seasonal Autoregressive Integrated Moving Moyenne</i>

SARIMAX	Moyenne mobile intégrée autorégressive saisonnière
SCP	<i>Shortest Constrained Path</i>
SDLTM	<i>Stochastic Dynamic Linear Time Model</i>
STAR	<i>System for Travel Analysis and Research</i>
STD	<i>Spatio-Temporal Data Representation</i>
STM	Société de Transport de Montréal
TOPSIS	Technique pour l'Ordre de Préférence par Similarité à la Solution Idéale

LISTE DES ANNEXES

Annexe A – Analyse descriptive	132
Annexe B – Résultats de la segmentation	134

CHAPITRE 1 INTRODUCTION

Dans un contexte de croissance urbaine rapide et de transition vers des villes plus durables, les réseaux de transport en commun occupent une place centrale dans la structuration de la mobilité urbaine. L'efficacité de ces réseaux repose non seulement sur une planification rigoureuse mais aussi sur leur capacité à s'adapter aux évolutions constantes des besoins des usagers.

Au cours des dernières années, des bouleversements majeurs ont redéfini les habitudes de déplacement. La pandémie de COVID-19 a été un déclencheur important, entraînant une réduction significative des déplacements domicile-travail et une baisse marquée de la demande de transport aux heures de pointe. Ce phénomène s'inscrit dans un cadre plus large, où l'adoption croissante du télétravail, les nouvelles dynamiques sociales, et les changements dans les modes de vie jouent un rôle clé. Ces facteurs ont accru la flexibilité des horaires de déplacement, augmentant la variabilité des flux passagers et posant de nouveaux défis pour la planification des services de transport. La variabilité de l'utilisation désignant ici les fluctuations observées dans les flux de passagers selon les jours, les heures ou les événements spécifiques, reflétant des besoins hétérogènes et en constante évolution. Cette évolution interroge sur l'adéquation des infrastructures existantes aux besoins en mutation, tout en révélant des opportunités pour réinventer la gestion des réseaux.

Traditionnellement, la planification des horaires de transport en commun repose sur des approches uniformes. Bien qu'efficaces dans un contexte stable, ces approches peinent à répondre aux nouvelles exigences. Par exemple, de nombreux réseaux appliquent encore des horaires fixes identiques du lundi au vendredi, sans tenir compte des différences dans la demande de transport entre les jours de la semaine. De plus, les fréquences de passage restent souvent rigides et ne s'ajustent pas suffisamment en fonction des fluctuations de la demande au cours de la journée ou entre les saisons. Les ajustements d'horaires (modifications apportées à la fréquence et aux plages horaires des services pour mieux répondre aux variations identifiées dans la demande) apparaissent comme une solution essentielle. Cette uniformité s'étend également aux zones urbaines et périurbaines, où les mêmes schémas sont appliqués malgré des réalités de demande très différentes. Ces limites soulignent la nécessité d'adopter des stratégies plus flexibles et adaptées.

Face à ces transformations, il devient impératif de repenser ces approches uniformes pour les adapter aux nouvelles réalités de déplacement. Ce mémoire se distingue par l'utilisation de données issues de cartes à puce, offrant une analyse précise et détaillée des comportements de déplacement.

En mobilisant des techniques avancées de segmentation, elle identifie les schémas de variabilité dans l'utilisation des réseaux de transport public à Orléans. Ces données permettent non seulement d'éclairer les disparités actuelles mais aussi de proposer des ajustements ciblés. Cette démarche innovante vise à développer une planification adaptative, centrée sur l'utilisateur et alignée sur des résultats mesurables.

L'hypothèse principale est que les comportements des usagers varient significativement selon les jours de la semaine et les plages horaires. Ces variations peuvent être exploitées pour affiner la planification des services et mieux répondre aux besoins identifiés. Les objectifs de cette étude sont:

1. D'identifier les différences de demande en fonction des jours et des heures grâce à une analyse fine des données de validation des cartes à puce et des caractéristiques des flux passagers.
2. De proposer une méthodologie robuste d'ajustement des horaires, basée sur des techniques avancées de segmentation et de transformation des données, permettant de mieux adapter l'offre de transport à la demande observée.
3. De tester et comparer les différentes méthodes de segmentation et de transformation des données, en évaluant systématiquement leurs performances selon des métriques pertinentes. Cette analyse approfondie inclut un travail statistique rigoureux pour valider la qualité des ajustements proposés, leur applicabilité pratique et leur impact sur les principaux objectifs de gestion des réseaux.

Cette étude propose des recommandations visant à mieux adapter les réseaux de transport aux variations de la demande. En ajustant les horaires et les fréquences, elle cherche à répondre aux attentes des usagers tout en améliorant l'efficacité opérationnelle des exploitants. Les résultats pourraient également inspirer des adaptations similaires dans d'autres environnements urbains.

1.1 Mise en contexte

Les réseaux de transport des grandes villes doivent faire face à des contraintes croissantes, ce qui oblige les planificateurs à mieux comprendre les comportements des usagers tout en utilisant efficacement les ressources disponibles. Ces défis, liés à une demande de mobilité diversifiée et à des contraintes budgétaires strictes, requièrent une analyse approfondie des données existantes. En particulier, les informations issues des systèmes d'aide à l'exploitation et des cartes à puce jouent un rôle clé pour ajuster les services de manière précise et adaptée aux besoins réels ainsi qu'aux tendances évolutives de déplacement.

La pandémie de COVID-19 a durablement impacté l'achalandage des transports publics. Par exemple, au niveau régional, seulement 70 % des usagers sont retournés dans les métros, trains et bus du Grand Montréal (ARTM, juillet 2022). Cette diminution persistante, renforcée par le fait que seules 57,7 % des entreprises ayant adopté le télétravail envisagent un retour complet sur site (Statistique Canada, juillet 2021), oblige les opérateurs à repenser leurs approches de planification. Le maintien du télétravail à des niveaux élevés pourrait instaurer une « nouvelle normalité », nécessitant des stratégies plus flexibles. Parmi celles-ci, la modulation des horaires ou l'augmentation ciblée de l'offre pour des segments d'usagers spécifiques pourrait mieux répondre à des schémas de déplacement moins homogènes.

Ce contexte appelle à revisiter les paradigmes traditionnels, comme les horaires constants du lundi au vendredi, en envisageant des ajustements différenciés selon les groupes d'usagers ou les jours spécifiques. Par exemple, les services pourraient être adaptés pour répondre aux fluctuations des heures creuses ou aux besoins particuliers des usagers périurbains. Cette démarche cherche à répondre aux besoins immédiats tout en alignant l'offre de transport sur des objectifs environnementaux ambitieux. Par ailleurs, une telle transition nécessite de prendre en compte la qualité des données disponibles, notamment celles issues des cartes à puce, qui peuvent ne pas refléter de manière exhaustive les comportements des usagers occasionnels ou non réguliers. De plus, l'absence de mises à jour en temps réel pourrait limiter la réactivité des ajustements proposés face aux évolutions rapides des flux de déplacements. Comprendre ces nouvelles dynamiques est crucial pour bâtir une mobilité décarbonée et adaptée aux besoins émergents. Ces ajustements pourraient également réduire l'empreinte environnementale des réseaux en adaptant les ressources aux flux réels d'usagers.

Pour relever ces défis, il est essentiel d'exploiter pleinement les données d'exploitation actuelles, en s'appuyant sur des méthodes avancées issues de la fouille de données et de l'apprentissage machine. Ces outils permettent non seulement d'analyser et d'expliquer les comportements évolutifs des usagers, mais aussi d'identifier des segments spécifiques ou des lignes sous-utilisées, tout en formulant des recommandations pratiques pour améliorer l'offre de transport. En intégrant ces approches, il devient possible d'anticiper les tendances futures et d'améliorer la résilience des réseaux face aux évolutions sociétales et environnementales.

1.2 Problématique

Les systèmes de transport jouent un rôle fondamental dans le maillage social et économique des villes, en facilitant l'accès aux services essentiels tels que l'emploi, l'éducation et les loisirs, tout en renforçant la cohésion sociale et en favorisant l'inclusion des populations vulnérables. En tant que vecteurs essentiels de la mobilité urbaine, ils facilitent non seulement l'accès aux opportunités d'emploi, d'éducation et de loisirs mais jouent aussi un rôle déterminant dans l'intégration sociale et la réduction de l'exclusion, notamment pour les populations vulnérables. La planification efficace des horaires de transport est au cœur de cette fonction, assurant une réponse adaptée aux besoins fluctuants de mobilité au sein des communautés urbaines (Delbosc & Currie, 2011). Cette fonction est particulièrement importante dans le contexte de l'étalement urbain, où les gens vivent loin de leur lieu de travail, de leur école et des commodités.

La fonction sociale du transport s'étend également à la réduction de l'exclusion sociale en améliorant l'accessibilité, en particulier pour les groupes défavorisés tels que les personnes âgées, les personnes handicapées et les personnes à faible revenu (Lucas, 2012). Ces groupes sont souvent confrontés à des obstacles dans l'accès aux services de transport en raison de facteurs tels que la mobilité physique, le coût ou le manque de services appropriés. En fournissant des services de transport accessibles et abordables, les sociétés peuvent s'assurer que ces personnes puissent participer pleinement à la vie sociale, contribuant ainsi à leur estime de soi et à leur sentiment d'appartenance.

La programmation des services de transport public doit prendre en compte la variabilité et les pics de demande pour améliorer la répartition des ressources et répondre efficacement aux attentes des usagers. Cette optimisation passe par une analyse précise des données de mobilité, incluant les volumes de passagers à différents moments de la journée, les jours de la semaine, et lors

d'événements spéciaux, afin de définir des horaires qui répondent au mieux à la demande tout en maximisant l'efficacité du système.

Les réseaux de transport jouent également un rôle essentiel dans le renforcement de la résilience sociale, en particulier dans les situations d'urgence telles que les catastrophes naturelles ou les crises sanitaires. Ils assurent l'approvisionnement en fournitures, services et réponses d'urgence nécessaires, contribuant ainsi à la résilience de la communauté (Kenyon et al., 2002). Par conséquent, la capacité des systèmes de transport à répondre de manière agile aux imprévus et aux évolutions à long terme de la demande est également cruciale pour leur durabilité et leur résilience. Cela inclut la gestion des crises, comme les pandémies. Lors de la pandémie de COVID-19, des réseaux efficaces ont été essentiels pour assurer la fourniture de services de santé et de denrées alimentaires. Ils ont également permis aux travailleurs essentiels de se rendre sur leur lieu de travail ou les catastrophes naturelles, où la flexibilité et l'adaptabilité des horaires peuvent jouer un rôle vital dans la réponse d'urgence et le soutien à la communauté.

La planification des transports est cruciale pour utiliser efficacement les ressources publiques et offrir des services optimaux aux usagers. Elle impacte l'économie en facilitant le commerce, par exemple en favorisant l'efficacité des déplacements professionnels, et en attirant les investissements grâce à une infrastructure de transport fiable et accessible (Shaw, 2019). Sur le plan environnemental, elle atténue les effets négatifs des transports en réduisant les émissions de gaz à effet de serre et en promouvant des modes de transport durables (Shaw, 2019). Socialement, elle garantit l'équité en offrant à tous, y compris les groupes défavorisés, un accès fiable aux services essentiels (Lucas, 2012). En outre, la planification stratégique permet de gérer les tendances et incertitudes futures, comme la croissance démographique, les avancées technologiques et les changements climatiques, assurant ainsi la résilience des systèmes de transport (Hall, 2007). En somme, une planification stratégique et durable des transports est essentielle pour répondre aux besoins actuels et futurs des communautés (Banister, 2008).

Selon de Dios Ortúzar et Willumsen (2011), les données jouent un rôle fondamental dans cette planification. La collecte et l'analyse des informations sur les comportements actuels des voyageurs, les demandes et les performances des systèmes permettent d'identifier les problèmes, de fixer les priorités et d'évaluer l'efficacité des politiques et des interventions. Par exemple, les données peuvent révéler des itinéraires encombrés, une faible utilisation des transports publics ou

des zones mal desservies, ce qui permet de cibler les interventions. Les progrès technologiques, comme le GPS et les applications mobiles, facilitent la collecte de données en temps réel et détaillées, rendant la prise de décision plus précise et rapide (Lee & Kam, 2014).

La programmation des services est un autre aspect essentiel. Une gestion efficace améliore l'efficacité et la fiabilité des transports publics, augmentant ainsi la satisfaction des usagers et leur utilisation (Ceder, A., 2015). Les prévisions de la demande future, basées sur divers facteurs tels que la croissance démographique et les tendances économiques, sont cruciales pour anticiper les besoins et garantir la durabilité et la résilience à long terme (de Dios Ortúzar & Willumsen, 2011).

L'étude de la variabilité dans la recherche sur les transports est essentielle ; en comprenant et en analysant la variabilité (les temps de parcours, les flux de passagers, les modèles de demande et les conditions d'infrastructure), les planificateurs peuvent prendre des décisions éclairées, améliorer les performances des systèmes et renforcer l'efficacité globale des réseaux. Ces informations sont cruciales pour la conception et la gestion des services afin de répondre aux attentes des usagers.

De plus, la variabilité des flux peut entraîner des surcharges, des goulots d'étranglement et des perturbations. L'étude de cette dernière aidera à identifier les causes et les impacts des fluctuations des flux de passagers, tels que les surcharges aux heures de pointe ou les incidents. Ainsi, une meilleure connaissance permettra de développer des stratégies et des outils pour une gestion efficace des flux de passagers, l'atténuation des encombrements et la réponse aux incidents d'une part, et d'identifier les facteurs sous-jacents qui influencent le comportement des voyageurs, tels que les caractéristiques socio-économiques, les variations saisonnières et les événements spéciaux d'autre part. Cette connaissance permet de prévoir avec précision la demande et de concevoir des systèmes capables de répondre à des demandes de déplacement variables.

Dans cette perspective, les réseaux de métro et de bus constituent des piliers fondamentaux des services de transport urbain, rendant essentielle la compréhension des déplacements des usagers, tant sur le plan spatial que temporel. L'analyse des variations dans l'utilisation de ces systèmes permet aux planificateurs d'améliorer la répartition des ressources disponibles en fonction des besoins identifiés. Ceci peut se traduire par un ajustement du nombre de véhicules en circulation, une modification des horaires ou encore une révision des itinéraires, afin de répondre de manière plus précise et efficace aux besoins de mobilité de la population. Une gestion avisée de la demande peut améliorer significativement l'efficacité des opérations et l'expérience des voyageurs, réduisant

les retards, augmentant le confort, et assurant des services plus ponctuels et fiables. Analyser la variabilité de la demande, notamment lors d'événements spéciaux ou de crises comme une pandémie, permet de mieux adapter les services. Adapter le transport public selon la demande favorise son utilisation, contribuant à diminuer les embouteillages et les émissions de CO₂. Investir dans des capacités accrues pour les itinéraires à forte demande se révèle ainsi crucial pour développer un système durable et performant.

Ce mémoire vise à proposer des outils pour analyser la variabilité de l'utilisation des transports publics, en vue d'ultimement permettre une re planification des horaires. En examinant les flux de passagers et leurs évolutions, cette étude cherche à établir des bases pour une gestion adaptée des horaires, contribuant ainsi à la mise en place d'un service réactif et efficace face aux fluctuations de la demande de transport public.

1.3 Objectif

La problématique présentée met en évidence l'importance d'analyser la variabilité de la demande en transport en commun afin de concevoir des réseaux flexibles, performants et alignés sur les besoins réels des usagers. Cette analyse permet d'ajuster les services en fonction des fluctuations observées, améliore ainsi la satisfaction des passagers tout en renforçant la durabilité et l'efficacité globale du système.

L'objectif principal de ce mémoire est de développer une méthodologie avancée pour analyser et mesurer l'usage des réseaux de transports en commun en s'appuyant sur des données issues des systèmes de cartes à puce et des GTFS. Ces données serviront à établir une typologie calendaire, incluant des profils de demande spécifiques pour les jours, les mois ou les saisons. Cela permettra de distinguer les jours usuels des jours exceptionnels, comme les jours fériés ou les vacances, et d'évaluer l'influence des événements planifiés ou imprévus, en orientant directement les ajustements spécifiques des horaires.

L'analyse se déploie à différentes échelles, incluant le réseau global, les modes de transport (bus et tramway), les lignes et les arrêts spécifiques. Elle fournit une compréhension fine des comportements des usagers pour identifier les points de saturation ou de sous-utilisation. En utilisant des méthodes de segmentation (K-moyennes, segmentation hiérarchique) et en intégrant DTW comme mesure de distance, les données sont regroupées pour révéler des modèles sous-jacents. Ces modèles se valident à l'aide de métriques rigoureuses comme WCSS, Davies-Bouldin et

Calinski-Harabasz, qui mesurent la compacité des groupes et identifient les regroupements les plus pertinents pour orienter les ajustements des horaires et des fréquences.

Ces techniques sont appliquées sur le nombre de passagers (mesure directe et représentative de l'utilisation des transports en commun) à différents intervalles de temps et sur divers objets du réseau :

1. **L'ensemble du réseau** : pour fournir une vue globale des modèles d'utilisation quotidienne et horaire.
2. **Le mode (bus et tramway)** : pour mettre en évidence des caractéristiques d'utilisation spécifiques à chaque mode et ajuster les services en conséquence.
3. **La ligne** : pour analyser les demandes propres à chaque ligne et identifier des modèles d'utilisation distincts.
4. **L'arrêt** : pour détecter les zones à fort achalandage et mieux comprendre les spécificités des flux par arrêt.

Les résultats obtenus permettent d'ajuster la planification des horaires en fonction des variations observées, qu'elles soient quotidiennes, hebdomadaires ou saisonnières. Ces ajustements contribuent directement à la réduction des temps d'attente, à une meilleure couverture en heures creuses et à une expérience utilisateur améliorée, tout en tenant compte des attentes en termes de confort, ponctualité et fréquence des services. Cette approche réduit directement l'empreinte environnementale des réseaux en adaptant les ressources aux flux réels d'utilisateurs, tout en assurant leur pérennité et leur performance face aux défis d'une mobilité urbaine en constante évolution.

1.4 Structure du mémoire

Ce travail de recherche est organisé en plusieurs chapitres, chacun apportant une vision approfondie des thèmes abordés. Le chapitre 2 présente une revue de la littérature, structurée en quatre sections. La première section aborde le contexte des données de carte à puce, leur processus de collecte et leurs usages dans les transports. La seconde section traite des séries temporelles appliquées au domaine du transport, ainsi que des méthodes d'analyse nécessaires pour comprendre et exploiter ces données complexes. La troisième section expose un panorama des différentes méthodes de classification, en mettant en lumière celles qui sont particulièrement pertinentes pour cette étude. Enfin, la quatrième et dernière section de ce chapitre se concentre sur la classification du

comportement des passagers, soulignant comment ces approches contribuent à mieux cerner, comprendre et prévoir leurs habitudes de déplacement.

Le troisième chapitre détaille la méthodologie globale de la recherche. Il débute par la présentation des données de carte à puce du réseau TAO, ainsi que des données GTFS, suivie d'une analyse exploratoire approfondie. Ce chapitre décrit ensuite la construction de différents types de vecteurs servant d'intrants pour la classification, tels que les vecteurs « jour-heure » et « jour-demi-heure », ou encore les vecteurs combinant les données relatives aux bus et aux trams pour étudier la demande globale. D'autres types de vecteurs, fondés par exemple sur les comptages horaires par ligne et par arrêt, sont également conçus. Le chapitre détaille également les transformations de données utilisées (log, racine carrée, Yeo-Johnson, etc.), les méthodes de segmentation (distances euclidienne, Manhattan, Cosinus) et les indicateurs servant à évaluer la qualité des groupes formés (WCSS, Davies-Bouldin Index, etc.). Enfin, il précise les approches de visualisation et d'évaluation approfondie des groupes obtenus.

Le quatrième chapitre présente les résultats issus des applications pratiques des modèles et des méthodes développés. Ces résultats illustrent la valeur ajoutée des approches employées pour la gestion et l'optimisation des réseaux de transport.

Pour conclure, le chapitre 5 propose une discussion complète des résultats, en soulignant les contributions majeures de cette recherche, les limites identifiées et les perspectives de travaux futurs. Cette dernière partie offre ainsi une synthèse et des pistes pour élargir ou approfondir la démarche engagée.

CHAPITRE 2 REVUE DE LITTÉRATURE

Cette revue de littérature s'articulera en deux grandes sections. La première portera sur l'utilisation des données de séries temporelles dans les transports, aspect central de notre projet de recherche. La seconde section traitera de l'exploitation des données de carte à puce, en abordant notamment la segmentation des usages et des usagers, les différentes méthodes de segmentation, ainsi que l'emploi de ces données pour la planification opérationnelle des réseaux de transport.

2.1 Conception des horaires

Rainville Jr (1947) considère que l'exploitation des services de transport urbain selon des horaires scientifiquement prédéterminés est d'une importance cruciale pour les usagers : les horaires de transport en commun, conçus de manière à répondre à la demande des passagers tout au long de la journée, garantissent des intervalles appropriés entre les véhicules, évitant ainsi la surpopulation et les longs temps d'attente, ce qui assure confort et commodité aux usagers. De plus, ces horaires déterminent les conditions de travail des employés, offrant des heures de travail stables et prévisibles, cruciales pour le moral et la performance des opérateurs. Sur le plan économique, une gestion efficace des horaires permet de minimiser les coûts opérationnels tout en maintenant la qualité du service, servant de référence pour la supervision des opérations et l'application des mesures correctives nécessaires.

Les données sur la charge des passagers sont essentielles pour la planification des horaires de transport. Elles permettent de mesurer avec précision la demande en différents points du réseau, garantissant ainsi une réponse adéquate en termes de fréquence et de capacité des services. Les méthodes couramment utilisées incluent la méthode du facteur de charge en période de pointe, qui ajuste la fréquence des bus pour éviter la surcharge en fonction des pics de demande. L'approche du facteur de charge égal cherche à uniformiser la répartition des passagers en ajustant les fréquences selon les charges observées à intervalles réguliers. La règle de la racine carrée, quant à elle, optimise la fréquence des bus en fonction de la relation entre la fréquentation et la distance, minimisant ainsi les temps d'attente et les coûts. Enfin, des modèles d'optimisation avancés utilisent des données détaillées pour allouer dynamiquement les ressources, garantissant un équilibre entre la qualité du service et l'efficacité économique (Furth & Wilson, 1981). Ceder, Avishai (1984) a mis en lumière deux approches principales: les méthodes de Point de Contrôle (Charge Maximale)

et les méthodes de Vérification du Trajet (Profil de Charge). Les méthodes de « Point de Contrôle », se basent sur des données recueillies à des points spécifiques de forte demande le long du trajet, ajustant la fréquence des bus pour garantir que la capacité réponde aux charges maximales. Par ailleurs les méthodes de « Vérification du Trajet » utilisent un profil de charge complet sur l'ensemble du trajet, permettant un ajustement plus nuancé des horaires pour mieux répondre à la variabilité de la demande des passagers. Bien que les méthodes de « Point de Contrôle » soient pratiques pour une utilisation régulière, elles peuvent ne pas saisir pleinement les fluctuations de la demande, rendant les méthodes de « Vérification du Trajet » plus adaptées aux trajets présentant une variabilité significative, malgré leur coût de mise en œuvre plus élevé.

Selon Hammerle et al. (2005), l'intégration des données de comptage des passagers avec les systèmes de localisation automatique des véhicules (AVL) permet d'améliorer la fiabilité des services de bus en ajustant les horaires selon la demande réelle. L'analyse des données de montée des passagers et de la régularité des intervalles entre les bus aide à adapter les horaires pour mieux répondre aux périodes de pointe, évitant ainsi la surcharge et les retards. En utilisant des techniques de fouille de données, telles que la segmentation et les arbres de décision, Patnaik et al. (2006) ont montré que l'intégration des données des compteurs automatiques de passagers (APC) permet d'optimiser la planification des horaires de bus en ajustant dynamiquement les intervalles en fonction de la demande réelle des passagers. Enfin, Sun et al. (2014), proposent des modèles de conception d'horaires sensibles à la demande qui utilisent les données de comptage des passagers pour ajuster dynamiquement les horaires des trains en fonction de la demande en temps réel : les données de cartes à puce permettent une analyse plus détaillée des schémas de demande spatiaux et temporels et fournissent des profils de charge des passagers précis et en temps réel qui alimentent la conception d'horaires plus efficaces. En résumé, l'incorporation des données de comptage dans le processus de conception des horaires permet de créer des horaires de bus plus réactifs, efficaces et centrés sur les besoins des passagers, ce qui conduit finalement à une amélioration de la qualité du service et des performances opérationnelles.

2.2 Carte à puce en transport collectif

L'utilisation des données des cartes à puce, dans le contexte des déplacements en transport en commun, a permis aux prestataires de services de transport de disposer d'une ressource de données nouvelle et importante qui peut être utilisée à de multiples fins. La qualité et la facilité d'utilisation

des données générées sont influencées par la conception, les capacités d'interface et le traitement de l'information des cartes à puce. Néanmoins, les avantages d'une participation réduite de l'utilisateur à la collecte des données sont contrebalancés par de nombreux facteurs qui ont une incidence sur la qualité des données. Bien que l'utilisation des données des cartes à puce permette de mener des recherches continues et approfondies, il est important de noter que cette approche peut ne pas englober les passagers qui persistent à payer leurs billets en espèces. Par conséquent, cette limitation se traduit par des informations inadéquates concernant des clients spécifiques. La collecte d'informations spatiales, en particulier de données relatives au débarquement, peut présenter des difficultés en l'absence de relevés de sortie. Même avec des enquêtes, il peut être nécessaire de recueillir des informations supplémentaires afin de déterminer avec précision les points de départ et d'arrivée (Bagchi & White, 2005).

Afin d'exploiter pleinement le potentiel des données issues des cartes à puce, il est impératif que les fournisseurs de services de transport adoptent une stratégie pragmatique qui reconnaisse à la fois les avantages et les limites de ces données. Les cartes à puce permettent de collecter de grandes quantités de données personnelles sur les déplacements, y compris des informations continues sur les trajets, lesquelles peuvent être associées à des cartes ou à des voyageurs spécifiques. Cependant, pour que ces données soient véritablement exploitables, il est crucial de mettre en place des processus méthodologiques rigoureux pour leur traitement et leur agrégation. Cela exige de prendre des décisions éclairées sur la manière de regrouper les données de manière cohérente, afin d'extraire des informations précises sur les tendances de déplacement et de faciliter la prise de décisions stratégiques en matière de gestion des flux de passagers. De plus, il est essentiel d'établir des limites spatiales et des périodes de rotation afin d'analyser le comportement des voyageurs sur une période spécifique (Bagchi & White, 2005).

Morency et al. (2006) en utilisant les données obtenues à partir des systèmes de perception automatique des tarifs par carte à puce (SCAFC) à la Société de transport de l'Outaouais (STO) située à Gatineau, au Québec, ont révélé la présence de quatre groupes distincts de semaines-usagers présentant des schémas de déplacement comparables. Deux de ces groupes présentent des comportements de déplacement discernables et significatifs en utilisant des techniques d'exploration de données.

Morency et al. (2006), en analysant les arrêts de transport en commun utilisés pour l'embarquement, ont mis en évidence des disparités spatiales dans les schémas de déplacement des usagers, influencées par des activités ponctuelles et des variations temporelles. Cette étude démontre l'utilité des données de cartes à puce pour saisir la variabilité des comportements des usagers dans le temps, notamment en observant l'évolution des habitudes de déplacement à travers différents jours. Dans le prolongement de ces analyses, l'utilisation des données provenant des systèmes de collecte automatique de billets par carte à puce (SCAFC) peut fournir des informations statistiques significatives sur la dynamique de l'offre et de la demande des réseaux de transport en commun (Trépanier et al., 2007).

Chu et Chapleau (2008) se concentrent principalement sur l'estimation des temps d'arrivée des bus aux arrêts et sur l'identification des trajets connectés grâce à l'utilisation de concepts spatio-temporels. Ils ont pu reconstruire des itinéraires de déplacement individuels, analyser les schémas de transfert et produire des profils de charge de véhicules. Les données améliorées offrent aux planificateurs une représentation spatio-temporelle complète de chaque voyage, y compris les points de départ et d'arrivée de chaque transaction, ainsi que les activités d'embarquement et de débarquement à chaque arrêt. L'utilisation de cet ensemble de données amélioré s'est avérée utile pour évaluer l'efficacité du système de transport actuel. Park et al. (2008) démontrent que l'utilisation des données des cartes à puce aux fins de planification et d'exploitation ont un impact significatif sur le transport en commun.

L'introduction des cartes à puce a des effets favorables, notamment une augmentation de la fréquentation et une plus grande commodité pour les usagers, même s'il reste des obstacles à surmonter en termes d'amélioration de la technologie, de garantie de l'exactitude des données, de résolution des problèmes de confidentialité et de mise en œuvre de mesures de sécurité robustes (Pelletier et al., 2011). Les systèmes de cartes à puce contribuent à l'amélioration de la planification et de la performance des transports publics grâce à la fourniture de données importantes. Agard et al. (2013) a utilisé la technique d'exploration de données K-moyennes pour analyser les transactions par carte à puce de la Société de transport de l'Outaouais (STO) à Gatineau, au Québec (douze mois, 9,4 millions de transactions effectuées à l'aide de 52 825 cartes à puce sur un réseau de 200 autobus). Une nouvelle approche pour le calcul de la distance a été introduite, incorporant des informations de position des observations dans des vecteurs binaires représentant chaque jour de carte. Elle a permis de créer trois groupes distincts, et la technique peut être appliquée à d'autres

types de vecteurs, tels que les temps de transaction, les lieux d'embarquement, les séquences d'itinéraires et les types d'itinéraires.

Ma et al. (2013) présentent une nouvelle approche qui utilise une chaîne de Markov et un algorithme d'arbre de décision bayésien pour estimer les arrêts d'embarquement des usagers des transports en commun. Leur méthode démontre un niveau de précision notable dans l'attribution des données d'embarquement. L'utilisation de l'algorithme de regroupement K-moyennes++ et la théorie des ensembles approximatifs ont permis de catégoriser les régularités des déplacements.

Munizaga et Palma (2012) soulignent le potentiel des données des cartes à puce en tant que source d'information viable pour la construction de matrices origine-destination (OD) pour les systèmes de transport en commun multimodaux et démontrent la viabilité et le potentiel de l'utilisation des données de la carte à puce pour la recherche sur les transports, suggérant son potentiel pour supplanter les enquêtes origine-destination conventionnelles à grande échelle à l'avenir. Plus loin, Munizaga et al. (2014) confirment l'efficacité et la fiabilité de l'approche suggérée pour estimer les flux origine-destination des transports publics et évaluer le niveau de service grâce à l'utilisation de données de cartes à puce et GPS.

Kusakabe et Asakura (2014) présentent une méthodologie de fusion de données qui intègre les données des cartes à puce aux données des enquêtes sur les déplacements des personnes. L'intégration de cette technique permet l'examen des caractéristiques comportementales persistantes et étendues des buts des déplacements des usagers du transport en commun, une tâche qui n'est pas facilement réalisable simplement au moyen de données d'enquête conventionnelles. La méthodologie employée facilite la surveillance et l'exploration de données des modèles de comportement des voyageurs observés dans les données des cartes à puce.

Robinson et al. (2014) introduisent diverses contributions, dont l'une concerne la proposition de règles d'agrégation de trajets individuels en trajets unifiés. Ce processus est crucial pour la modélisation basée sur l'activité et pour garantir une tarification précise des passagers. Il sert à identifier les véhicules défectueux ou la fourniture de données en utilisant une « approche de comparaison avec les pairs », facilitant ainsi l'amélioration des systèmes de localisation automatique des véhicules pour les agences de transport en commun.

Bhaskar et Chung (2014), utilisent l'algorithme DBSCAN (« *Density-Based Spatial Clustering of Applications with Noise* ») pour répartir les usagers du transport en commun en groupes distincts à

l'aide de données de carte à puce, extrayant des schémas origine-destination (OD) réguliers et des temps de trajet habituels. Les résultats ont permis d'acquérir des connaissances importantes sur les segments de clientèle, leur permettant ainsi de prendre des décisions éclairées concernant l'amélioration des services et la croissance des systèmes de transport en commun. Ali et al. (2016) ont utilisé les données de cartes à puce - comprenant des attributs importants tels que l'identification de la carte, les horodatages d'embarquement et de débarquement et le mode de transport - dans le développement d'un modèle de simulation par agents basé activités pour l'analyse des modèles de transport dans la région métropolitaine de Séoul en utilisant la boîte à outils « *MATSim* » pour la microsimulation afin d'examiner les attributs de comportement de voyage des personnes qui utilisent des cartes à puce.

Wang et al. (2016) se concentrent sur l'exploitation des techniques adaptées aux mégadonnées pour calculer efficacement le temps de trajet des bus et les besoins des passagers : une première couche utilise MapReduce pour analyser un volume important de signaux GPS et d'enregistrements (« *Automated Fare Collection – AFC* ») pour calculer la durée du trajet en bus et identifier les endroits spécifiques où les passagers montent à bord du bus. Ensuite une deuxième couche utilise une analyse de régression pour tirer des informations des données traitées. Cette recherche met l'accent sur la nécessité d'une technologie de surveillance de la position telle que les systèmes GPS, AVL et APC pour améliorer les données existantes pour le transport urbain. Néanmoins, les méthodologies conventionnelles ont tendance à combiner les demandes des passagers sans tenir compte des habitudes de voyage uniques des individus. La technologie proposée vise à extraire des informations détaillées à partir de données non traitées, offrant une alternative rentable et largement applicable aux systèmes AVL-APC.

He et al. (2017) examinent le processus d'extraction du comportement des utilisateurs dans les systèmes de transport en commun, en soulignant sa nature intrinsèquement dynamique en appliquant la méthodologie des séries chronologiques qui est couramment utilisée pour prédire les modèles de transport, avec un accent particulier sur l'application des méthodes de corrélation croisée et de déformation temporelle dynamique. Les distances de corrélation croisée et de déformation temporelle dynamique ont été calculées et ont abouti à la génération d'une matrice de distances englobant plusieurs combinaisons. Puis, une technique de regroupement hiérarchique fut utilisée pour classer les données en groupes distincts aux fins d'analyse de corrélation croisée. De plus, les données sont partitionnées en groupes à l'aide d'une méthodologie de déformation

temporelle dynamique. Les auteurs présentent une nouvelle approche qui intègre la distance de corrélation croisée avec le regroupement hiérarchique afin de répondre aux contraintes associées à l'utilisation de la distance euclidienne pour la segmentation des séries chronologiques : la distance de corrélation croisée présente des performances supérieures et la méthodologie proposée offre un niveau de précision plus élevé dans la compréhension du comportement des clients à l'intérieur des réseaux de transport en commun par rapport à la technique de déformation temporelle dynamique.

Ma et al. (2017) ont présenté un cadre théorique qui utilise des techniques d'exploration de données pour analyser les données des cartes à puce afin de découvrir les modèles de navettage en transport en commun. La méthodologie proposée a la capacité de discerner les navetteurs en transport en commun en examinant les schémas spatio-temporels de déplacement sur de longues périodes d'observation en utilisant un algorithme de regroupement spatial en conjonction avec la méthodologie TOPSIS (Technique pour l'ordre de préférence par similarité à la solution idéale). De plus, il a la capacité d'extraire des détails spécifiques concernant les lieux de résidence et de travail des individus. L'efficacité et la précision du cadre pour discerner les schémas de déplacement en transport en commun ont été démontrées grâce à une enquête à un niveau désagrégé, résultant en un taux d'exactitude notable de 94,1 % dans l'identification des navetteurs de transport en commun.

Ghaemi et al. (2017) ont utilisé des techniques d'apprentissage sur une base de données de cartes à puce afin de catégoriser les utilisateurs en sous-populations distinctes en fonction de leur comportement. La méthodologie proposée consiste en une approche visuelle en deux étapes qui utilise une projection en demi-cercle pour réduire la dimensionnalité des données et applique ensuite un regroupement hiérarchique pour générer des structures de groupes pour une analyse plus approfondie. L'approche décrite dans cette étude implique la transformation de données binaires en variables continues à l'aide de la méthode d'investigation en composantes séquentielles (SCP). Cette méthodologie permet une analyse efficace sur les modèles de comportement des utilisateurs et fournit des contributions significatives à la compréhension du comportement des passagers du transport en commun en utilisant des techniques d'exploration de données et en effectuant des analyses sur les transactions par carte à puce.

Mohamed et al. (2016) mettent l'accent sur l'accessibilité croissante des empreintes numériques à l'intérieur des environnements urbains, notamment dans les réseaux de transport en commun,

donnant ainsi naissance à la discipline de l'informatique urbaine. Ils ont analysé les données des cartes à puce du Service des Transports en commun de l'Agglomération Rennaise (STAR) à Rennes (France) afin d'avoir un aperçu du fonctionnement du système de transport public, avec un accent particulier sur l'utilisation des stations et le comportement des passagers au fil du temps. Les données sont enrichies et prétraitées afin de prédire les lieux de débarquement des passagers et de recréer leurs parcours. L'approche de classification des stations utilise un algorithme de regroupement basé sur un modèle afin de découvrir des groupes de stations qui ont des profils d'activité temporels similaires. Les résultats indiquent la présence de plusieurs groupes caractérisés par des modèles d'utilisation temporels variables, englobant des stations avec une utilisation équitable ainsi que celles qui présentent une activité constante tout au long de la journée. Les passagers ont été regroupés en fonction de leur comportement temporel. De plus, des profils temporels pour les passagers individuels ont été créés, suivie de l'application d'une technique de regroupement basée sur un mélange de modèles d'unigrammes. Cela a permis d'identifier les groupes de passagers qui présentent des habitudes de voyage similaires. Enfin, l'étude se termine en soulignant la nature complémentaire de l'utilisation des méthodologies de regroupement des gares et des passagers.

Zhao et al. (2017) se sont intéressés à la détection d'anomalies et la découverte de régularités dans les habitudes de déplacement des passagers individuels en utilisant des algorithmes basés sur des statistiques et des groupes non supervisés. L'identification des anomalies est effectuée en utilisant l'analyse des paires OD et l'analyse des anomalies du coût temporel. Le regroupement des passagers est effectué en utilisant des caractéristiques temporelles et spatiales pour discerner des cohortes de passagers présentant des comportements comparables. Enfin, ils soulignent l'importance de l'incorporation de variables supplémentaires telles que la démographie des passagers et les préférences d'itinéraire afin de procéder à un examen plus complet des comportements de voyage des passagers individuels.

Deschaintres et al. (2018) soulignent l'importance de comprendre les écarts individuels dans l'utilisation des transports en commun en étudiant la variabilité interpersonnelle et intrapersonnelle chez les personnes au cours d'une année complète. La méthodologie utilisée englobe l'estimation de variables temporelles et spatiales au niveau individuel afin de caractériser les comportements de déplacement de chaque groupe. Des typologies d'utilisateurs sont développées en analysant les statistiques d'utilisation mensuelles et les modèles de comportement hebdomadaire, offrant ainsi

des informations précieuses sur les modèles de régularité individuels et hebdomadaires. Ils ont aussi souligné l'importance de prendre en compte à la fois l'hétérogénéité inter et intrapersonnelle lors de l'examen de l'utilisation des transports en commun. Les auteurs proposent en outre d'étudier la dimension spatiale comme moyen d'approfondir la régularité spatiale dans l'utilisation des transports publics.

Li et al. (2018) se sont centrés sur l'exploitation des données massives générées par les cartes à puce pour l'estimation des destinations des passagers. Ils soulignent que:

- L'acquisition de données spatio-temporelles précises concernant les points d'embarquement et de débarquement est cruciale pour une analyse rigoureuse des données issues des cartes à puce. Cela permet non seulement d'améliorer la prévision de la demande, mais aussi d'optimiser l'allocation des véhicules dans les réseaux de transport.
- Les modèles d'enchaînement de trajets sont fréquemment utilisés dans la déduction des destinations des passagers dans les systèmes de transport public. Le processus de validation des modèles et la réalisation d'analyses de sensibilité sont essentiels pour affiner la précision et accroître l'efficacité de ces méthodes.

Richer et al. (2018) ont utilisé les données des cartes à puce afin d'interroger et d'analyser les pratiques d'intermodalité mises en œuvre au sein du réseau bus-métro de Rennes Métropole. Nonobstant les avantages associés aux données des cartes à puce, elles ne suffisent pas à elles seules pour répondre à toutes les questions concernant la mobilité intermodale. L'étude souligne la possibilité d'intégrer ces données avec d'autres sources, telles que les données socio-économiques, les enquêtes de mobilité, ou encore les informations provenant d'autres modes de transport (comme le train ou le covoiturage), afin d'améliorer la compréhension des dynamiques intermodales et de proposer des solutions plus adaptées pour la gestion des réseaux de transport urbain.

Plus loin, Alsger et al. (2018) montrent qu'en intégrant les données d'enquêtes de mobilité, d'utilisation des sols et de GTFS, la précision de l'inférence des objectifs de déplacement est significativement améliorée. De plus, même en l'absence d'informations explicites sur l'objectif du déplacement, les données des cartes à puce, associées à des caractéristiques temporelles, ont un impact substantiel sur la précision de la prédiction des objectifs de déplacement, en particulier pour les trajets liés à la maison, au travail et aux courses.

Egu et Bonnel (2020) à partir des méthodes de segmentation, parviennent à identifier et à analyser la variabilité quotidienne des comportements de déplacement au sein du réseau de transport en commun de Lyon. Contrairement à l'idée reçue selon laquelle les usagers des transports en commun suivent des schémas de déplacement réguliers et prévisibles, l'étude démontre que les comportements de déplacement peuvent varier de manière significative d'un jour à l'autre, tant au niveau individuel qu'entre différents usagers; d'où la nécessité d'adopter une approche plus nuancée pour analyser les données de transport, permettant ainsi de mieux comprendre les besoins et les préférences des usagers. He et al. (2020) en appliquant une démarche similaire ont démontré que la distance de corrélation croisée (CCD) présente des performances supérieures par rapport à la déformation temporelle dynamique (DTW) pour la classification des séries chronologiques des données de cartes à puce. En effet, l'algorithme CCD a permis d'obtenir des groupes plus cohérents et homogènes que ceux formés avec DTW. La corrélation croisée s'est ainsi révélée plus adaptée pour la classification des profils temporels dans les données de transport en commun. Toutefois, ils ont également reconnu certaines limites, notamment en ce qui concerne la sélection des mesures appropriées aux problématiques spécifiques de transport, et ont souligné la nécessité de poursuivre les recherches pour identifier la métrique la plus adéquate dans différents scénarios.

Torti et al. (2022) examinent le besoin croissant de transport en commun dans les zones urbanisées et mondialisées, en se concentrant spécifiquement sur le service ferroviaire régional en Lombardie en proposant une approche polyvalente pour caractériser les schémas de déplacement, examiner les interactions spatiales et temporelles entre les passagers et l'infrastructure de transport et détecter les situations de congestion excessive afin d'améliorer la qualité du service. Leurs résultats fournissent des preuves de l'efficacité de la méthode de bi-segmentation suggérée pour révéler des modèles complexes qui resteraient non détectés par des analyses de composants individuels.

Cats (2022) propose une revue systématique des approches de segmentation utilisées pour identifier les modèles de déplacement, avec une distinction clé entre la variabilité intra-personnelle (stabilité des habitudes de déplacement d'un individu au fil du temps) et la variabilité interpersonnelle (différences de comportements entre les usagers). Elle suggère des pistes potentielles de recherche pour améliorer la qualité du service notamment à travers la mise en œuvre de stratégies de contrôle en temps réel et la prédiction du flux de passagers à court terme. Cependant, des lacunes subsistent, telles que la nécessité d'une meilleure intégration des dimensions spatiales et une gestion plus efficace des grands ensembles de données. Pour combler

ces lacunes, il est proposé de développer des produits tarifaires plus adaptés à des segments de marché spécifiques, en tenant compte des comportements variés des usagers.

2.3 Séries temporelles en transport

Les données de séries temporelles sont devenues un outil de plus en plus critique dans la recherche et la planification des transports en raison de la richesse des informations qu'elles permettent d'obtenir sur les différents phénomènes de transport. Elles font référence à une séquence de points de données enregistrés à des intervalles de temps uniformes. Dans le contexte des transports, ces points de données peuvent concerner les volumes de trafic, la fréquentation des transports en commun, les taux d'accidents ou l'usure des infrastructures, entre autres (Thacker et al., 2019).

La nature chronologique des données de séries temporelles permet de comprendre les modèles temporels et de prévoir les tendances futures dans les systèmes de transport.

La prédiction (estimation, calcul) et la prévision (plus d'incertitudes et de variables complexes) sont des domaines d'application importants des données de séries temporelles dans les transports. Le travail de Kamarianakis et Prastacos (2005) montre comment l'analyse des séries temporelles, associée à des données spatiales, peut apporter des informations précieuses pour la prévision du trafic. Leur étude a intégré des corrélations temporelles et spatiales dans les données de trafic pour générer des prévisions de flux de trafic à court terme. De même, Williams et Hoel (2003) ont exploité des données de séries temporelles pour prédire les volumes de trafic futurs, en utilisant des modèles de moyennes mobiles intégrées autorégressives (ARIMA) et des méthodes de lissage exponentiel, illustrant le potentiel des données de séries temporelles pour permettre des modèles de prévision précis.

Outre la prédiction, les données de séries temporelles ont permis d'améliorer la sécurité routière. Chen et al. (2009) ont utilisé des données de séries temporelles pour estimer des modèles d'accidents de la circulation, en repérant les lieux et les moments où les risques d'accident sont les plus élevés. Cela a permis de déployer des interventions ciblées, démontrant l'utilité pratique des données de séries temporelles dans l'élaboration de stratégies de sécurité efficaces.

L'avènement des données massives et des technologies de l'Internet des objets (« *IoT* ») a multiplié le volume et la granularité des données de séries temporelles disponibles dans les transports. Dans une étude réalisée par Erhardt et al. (2019), ces développements permettent une modélisation et

une planification plus précises des transports. Les données de séries temporelles à haute résolution permettent d'analyser les schémas de circulation minute par minute, ce qui augmente la précision des prédictions et des efforts d'optimisation.

Malgré ces avancées remarquables, l'exploitation des données de séries temporelles dans le domaine des transports reste un défi de taille. Comme l'ont noté Vlahogianni et al. (2014), des questions telles que la qualité des données et la protection de la vie privée restent des préoccupations majeures. En outre, il est nécessaire de disposer de méthodes robustes pour traiter la dimensionnalité et la granularité élevées des données. L'interprétation des données de séries chronologiques soulève également des questions, car les relations ne sont pas toujours causales, mais plutôt déterminées par des facteurs latents sous-jacents. Ainsi, des techniques avancées d'apprentissage automatique sont explorées pour extraire des informations significatives de ces données, comme le soulignent Kieu et al. (2019).

2.4 Séries temporelles des données de comptage et de cartes à puce dans les transports

L'adoption de technologies de cartes à puce et de systèmes automatisés de collecte de données (ADCS) dans le domaine des transports fournit aux planificateurs la possibilité de rassembler une quantité énorme et en constante augmentation de données chronologiques sur les comportements et les habitudes de déplacement des navetteurs. Pelletier et al. (2011) ont montré que ces données peuvent être très utiles aux planificateurs du transport en commun, de l'exploitation quotidienne du système de transport en commun à la planification stratégique à long terme du réseau. En outre, les données de carte à puce renseignent sur un réseau de transport à l'échelle globale et peuvent suivre l'emplacement d'un utilisateur spécifique tout au long de la journée, de la semaine ou pendant une période plus longue (Decouvelaere et al., 2022). Toutefois, la croissance explosive des bases de données temporelles a largement dépassé la capacité des planificateurs de transport à interpréter ces données à l'aide de techniques statistiques conventionnelles, créant un besoin urgent de nouvelles techniques pour aider l'analyste à transformer les données en informations et connaissances exploitables.

L'un des moyens de comprendre la séquence d'activités des personnes au cours d'une période est l'application des techniques de classification aux données d'utilisateurs des cartes à puce (He et al., 2020). Mariñas-Collado et al. (2022), par exemple, utilisent les données collectées par les

systèmes de comptage automatique des passagers à Stockholm pour étudier la relation entre les volumes de passagers et la fréquence et la durée des arrêts (« *dwell times* ») des trains de banlieue. Deschaintres et al. (2022) ont utilisé plus de 429 millions de validations, réalisées par près de 2 millions de cartes, pour étudier la variabilité d'utilisation du transport en commun, définissant des indicateurs sur la dispersion des déplacements, la variabilité de la fréquence d'utilisation, la variance temporelle du nombre de déplacements par mois et la diversité spatiale des lieux d'embarquement.

À partir de 51 semaines de données de validation des cartes d'abonnés Deschaintres et al. (2019) ont étudié la variabilité interpersonnelle et intrapersonnelle dans l'utilisation hebdomadaire du transport public. Les résultats montrent que les détenteurs d'abonnements annuels sont plus réguliers et voyagent principalement durant les jours de la semaine, tandis que les détenteurs d'abonnements mensuels sont plus versatiles et ne sont pas affectés par les jours fériés et les fins de semaine. Les utilisateurs de carnets de billets sont plutôt des occasionnels. Cette étude souligne l'importance de segmenter les utilisateurs pour mieux répondre à leurs besoins spécifiques, tout en tenant compte des comportements intrapersonnels et interpersonnels. Cependant, elle présente des limites, notamment l'utilisation de la distance euclidienne pour la typologie des semaines, qui ne prend pas en compte l'ordre des jours et peut être sensible à la méthode de standardisation. Des métriques plus avancées et des cycles alternatifs méritent d'être explorés pour affiner la compréhension des comportements de mobilité.

Morency et al. (2007) ont utilisé des données de carte à puce pour mesurer la variabilité de l'utilisation du réseau de transport en commun urbain avec des données recueillies au cours de 277 jours consécutifs. Des mesures de la variabilité spatiale et temporelle de l'utilisation des transports en commun pour différents types de carte ont été définies et des techniques d'exploration de données ont été estimées. Les cycles d'utilisation du transport en commun et les jours et semaines homogènes de déplacement parmi les segments de carte ont pu être identifiés.

Lee et Kam (2014) explorent l'utilisation potentielle de l'exploration de données chronologiques en intégrant des techniques d'analyse de séries chronologiques et d'exploration de données conventionnelles pour découvrir des informations exploitables à partir des données temporelles de transport. Leur étude de cas sur les trains à Singapour applique ces techniques sur plus de 60 millions de voyageurs, générant des modèles de déplacement pour 102 gares.

Ghosh et al. (2009) ont introduit les séries chronologiques structurales (STM). Le modèle structurel de séries chronologiques (STM) (sous sa forme multivariée) a été introduit pour développer un algorithme de prévision des conditions de trafic à court terme multivarié, parcimonieux et simple sur le plan informatique. Les différentes composantes d'un ensemble de données de séries chronologiques telles que les variations tendanciennes, saisonnières, cycliques et calendaires sont modélisées séparément dans la méthodologie STM. Cela indique que l'algorithme de prévision proposé est une approche efficace pour prédire le flux de passagers en temps réel à plusieurs jonctions au sein d'un réseau de transport urbain.

Cyprich et al. (2013), ont mené une étude sur la prévision à court terme de la demande de passagers pour les séries temporelles de transport par bus de banlieue en utilisant plusieurs modèles, dont Box-Jenkins (ARIMA), le lissage exponentiel, et la régression linéaire multiple. L'objectif principal était de concevoir un modèle plus précis et fiable par rapport à ceux couramment utilisés. Le modèle ARIMA s'est avéré être le plus performant, répondant à presque tous les critères statistiques à l'exception de la normalité des résidus du modèle. Malgré la limitation, le modèle ARIMA a démontré des capacités de prévision supérieures par rapport aux autres méthodes étudiées.

Andreoni et Postorino (2006) ont testé des modèles ARIMA univariés et multivariés pour estimer la demande de transport aérien à l'aéroport régional de Reggio Calabria, situé dans le sud de l'Italie. L'étude met en lumière les différences entre les modèles univariés et multivariés, en utilisant la procédure Box-Jenkins pour calibrer les modèles. La comparaison a montré que les modèles univariés ARIMA fournissent des résultats satisfaisants et s'adaptent mieux aux fluctuations brusques (pics) que les modèles ARIMAX multivariés. Cependant, l'étude conclut qu'il n'est pas possible d'affirmer de manière absolue que les modèles univariés sont meilleurs que les modèles multivariés. Le meilleur pouvoir de prévision des modèles univariés est compensé par leurs limites, qui dépendent de la stabilité des conditions aux limites. À l'inverse, les modèles multivariés, bien qu'ils résolvent ce problème grâce à l'inclusion de variables explicatives, rencontrent des difficultés liées à la disponibilité des séries chronologiques nécessaires à leur estimation.

Kumar et Vanajakshi (2015) ont rapporté que l'approche basée sur les données utilisant « *Box-Jenkins Autoregressive Integrated Moving Average (ARIMA)* » exige une base de données solide pour la construction de modèles. Dans les endroits où la disponibilité des données pourrait être un

problème, la validité de ces modèles devient discutable. La solution proposée consiste en un schéma de prédiction utilisant le modèle saisonnier (SARIMA) pour la prédiction à court terme du flux de passager en utilisant uniquement des données d'entrée limitées.

Cyril et al. (2019) ont utilisé les données des transactions tarifaires lors de l'émission des billets dans les bus de la Kerala State Road Transport Corporation à l'aide d'une machine de billetterie électronique. Les performances prévisionnelles de quatre modèles Holt-Winters sont comparées à l'aide de MAPE. L'analyse des données sur la demande de passagers d'autobus révèle que la saisonnalité quotidienne est présente dans les données. La saisonnalité de sept indique que les données décalées d'une semaine montrent une corrélation significative avec les observations de la série chronologique d'origine. Les modèles multiplicatifs de Holt-Winters avec et sans amortissement surpassent les modèles additifs. Cependant, le MAPE de tous les modèles est inférieur à 10%, le modèle multiplicatif amorti ayant la plus petite valeur de 8,29%. Les écarts-types des modèles multiplicatifs sont négligeables, ce qui suggère que l'ajustement du modèle est bon. Par conséquent, la méthode multiplicative amortie de Holt-Winters fournit de meilleures prévisions dans le cas des données saisonnières.

Pour prédire le nombre de demandes de passagers d'un certain bus b à l'arrêt de bus à l'instant t , Zhou et al. (2016) prédisent le nombre de demandes de passagers dans l'espace pour une période de P -minutes. Sur la base de la localisation GPS historique et des données de service de prise en charge et de dépose des passagers, des histogrammes de séries chronologiques sont créés pour chaque arrêt de bus contenant le nombre de demandes de passagers avec une agrégation de P -minutes. Ainsi, la demande de passagers est prédite en fonction des prévisions concernant la répartition de la demande de passagers sur la période $[t, t + P]$. Pour y parvenir, trois modèles de prédiction distincts sont étudiés et un cadre d'ensemble de flux de données est considéré. D'abord un modèle de Poisson variant dans le temps est évalué, puis un modèle de Poisson variant dans le temps est pondéré pour s'adapter aux motifs irréguliers en rafales. Cependant, ces approches présentent plusieurs limites, notamment la sensibilité au bruit à des granularités temporelles fines et l'incapacité de capturer des tendances non stationnaires complexes. Menon et Lee (2017), montrent comment une telle demande à court terme peut être modélisée avec précision avec un processus de Poisson inhomogène, en utilisant un réseau de neurones comme intensité sous-jacente. Ce choix d'intensité subsume les modèles existants comme des cas particuliers et est suffisamment puissant pour saisir certains faits stylisés (composante cyclique forte, non-linéarité

des tendances, influence de la granularité temporelle) de la demande réelle. Badoe (2007) a testé le normal tronqué, le modèle de Poisson, le modèle binomial négatif et un modèle logit ordonné afin de corriger les lacunes sous-jacentes du modèle linéaire avec comme unité de modélisation l'individu. Les modèles sont ensuite évalués en fonction de leur capacité à reproduire les trajets produits par chaque individu dans l'ensemble de données et par chaque zone de circulation. Cette évaluation de la performance en prédiction est effectuée sur un ensemble de données d'estimation recueillies dans la région de Toronto en 1986, et sur un ensemble de données indépendant recueilli dans la même région géographique, 10 ans plus tard, en 1996. Les résultats montrent que, nonobstant la simplicité du simple modèle linéaire et son absence de théorie comportementale de voyage sous-jacente explicite, il prédit les voyages dans les années de référence et de prévision avec moins d'erreur par rapport à l'un de ces autres modèles plus complexes.

En se concentrant sur les membres adultes des ménages allemands, Frondel et Vance (2011) examinent le déterminant de l'achalandage des transports en commun dans le but de quantifier les effets des prix du carburant, des tarifs, des attributs au niveau de la personne et des caractéristiques du système de transport en commun sur le nombre de déplacements sur une semaine de cinq jours. Le recours à des données individuelles a soulevé plusieurs problèmes conceptuels et empiriques, dont le plus fondamental est la grande proportion de valeurs nulles dans les comptages de passagers. Pour tenir compte de cette caractéristique des données, des procédures de modélisation appelées modèles à zéro gonflé (« *ZIMs* »), qui ordonnent les observations en deux régimes latents définis par le fait que l'individu n'utilise jamais les transports en commun, sont utilisées. Le problème a été résolu en employant des modèles de « *Zero-inflated Poisson* » particulièrement appropriés lorsque la question posée nécessite de faire la distinction entre ceux qui n'utilisent jamais le transport en commun et ceux qui ont une probabilité non nulle d'un nombre de déplacements même s'ils n'ont pas utilisé le transport en commun pendant la période d'observation.

En se basant sur un algorithme heuristique qui génère un ensemble de routes réalisables et un algorithme génétique qui trouve le réseau optimal en termes de routes et leurs fréquences associées, Cipriani et al. (2012) ont introduit l'adoption d'une procédure de concentration des flux dans un processus de génération d'itinéraires plus large et l'application de la méthodologie de conception du réseau de transport en commun à une grande zone urbaine réelle, caractérisée par une topologie de réseau routier complexe et un système de transport public multimodal plus une demande de déplacements plusieurs à plusieurs. Les expériences de Bourbonnais et al. (2021), menées auprès

de trois villes de taille moyenne au Québec, démontrent que la réalisation d'optimisations par algorithmes génétiques à l'aide de données précises sur le réseau routier local et de données représentatives de la demande en transport en commun peut générer des scénarios plausibles qui sont entre 10 et 20 % plus efficaces que les réseaux existants, en utilisant les mêmes paramètres et flottes de véhicules de taille similaire.

Enfin, Jha et al. (2019) tout en tenant compte de la demande symétrique pour la facilité de calcul, ont proposé une procédure IRSG combinée à un algorithme génétique pour résoudre le problème de conception du réseau de transport en commun afin de déterminer un ensemble efficace d'itinéraires pour fournir le meilleur pourcentage de trajets directs possible (dans quelques cas, il atteint 100 %) et de réduire le temps de trajet moyen des passagers. Les solutions obtenues se sont considérablement améliorées en termes de pourcentage de trajets directs et de temps de trajet moyen, et surpassent les résultats précédents Arbex et da Cunha (2015), Kechagiopoulos et Beligiannis (2014) & Nikolić et Teodorović (2013).

2.5 Classification des données de séries temporelles

L'objectif de la segmentation ou classification est d'identifier la structure d'un ensemble de données non étiquetées en organisant objectivement les données en groupes homogènes où la similarité entre les objets d'un même groupe est maximisée et la dissimilarité entre les objets de groupes différents est maximisée. La réduction de la dimensionnalité des données est un objectif primordial dans l'utilisation de la représentation des séries temporelles (Fu, 2011). La représentation d'une série temporelle est primordiale en raison de la difficulté inhérente à la manipulation des données de séries temporelles sous leur forme originale. De plus, elle joue un rôle crucial dans le regroupement des séries temporelles, notamment grâce à sa capacité à réduire les besoins en mémoire (Aghabozorgi et al., 2015). Cela est particulièrement utile car le stockage de toutes les données brutes de séries temporelles en mémoire principale peut s'avérer impraticable.

Han et al. (2006) ont classé les méthodes de segmentation ou classification mises au point pour traiter diverses données statiques en cinq grandes catégories : les méthodes de partitionnement, les méthodes hiérarchiques, les méthodes basées sur la densité, les méthodes basées sur la grille et les méthodes basées sur un modèle (Figure 2-1 - Résumé des approches majeures en classification des séries temporelles).

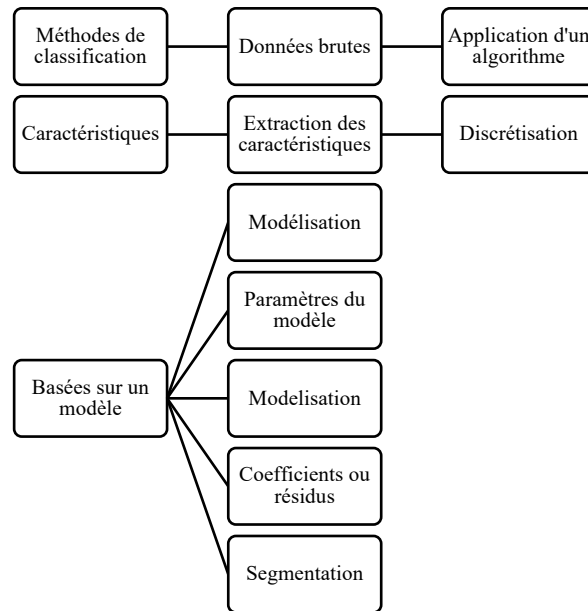


Figure 2-1 - Résumé des approches majeures en classification des séries temporelles

2.5.1 Approches basées sur les données brutes

Les algorithmes de regroupement de séries temporelles basés sur des données brutes utilisent les données telles quelles, sans effectuer de prétraitement ou d'extraction de caractéristiques (ou *features*). Ces caractéristiques désignent des propriétés spécifiques des séries temporelles (comme des tendances, des fréquences ou des statistiques résumées) qui sont extraites pour simplifier leur représentation et en faciliter le traitement. À l'inverse, certains algorithmes de regroupement appliquent un prétraitement pour extraire ces caractéristiques, réduisant ainsi la complexité des données. Le Tableau 2-1 - Algorithmes de segmentation sur les données brutes regroupe les approches couramment utilisées pour le regroupement des séries temporelles à partir de données brutes.

Tableau 2-1 - Algorithmes de segmentation sur les données brutes

Variable	Longueur des séries	Algorithme de segmentation
Univariée	Égale	<i>Fuzzy-means</i>
Multivariée	Égale	Hiérarchique agglomératif
Multivariée	Égale	Procédure modifiée de segmentation par relocalisation
Univariée	Égale	Hiérarchique agglomératif
Multivariée	Égale & Inégale	Moyenne et <i>Fuzzy-means</i>
Univariée	Égale & Inégale	Regroupement génétique basé sur les médioïdes
Univariée	Égale	Moyenne et <i>Fuzzy-means</i>
Univariée	Égale	Fuzzy clustering
Multivariée	Égale	Hiérarchique agglomératif
Univariée	Égale	Hiérarchique agglomératif
Univariée	Égale	Regroupement de réseaux neuronaux

2.5.2 Approches basées sur l'extraction de caractéristiques

Les approches basées sur les caractéristiques pour la classification de séries temporelles consistent à extraire des caractéristiques significatives à partir des données brutes des séries temporelles. Ces caractéristiques sont ensuite utilisées pour représenter chaque série temporelle, avant d'appliquer les algorithmes de regroupement traditionnels (Figure 2-2 - Approches basées sur l'extraction de caractéristiques).

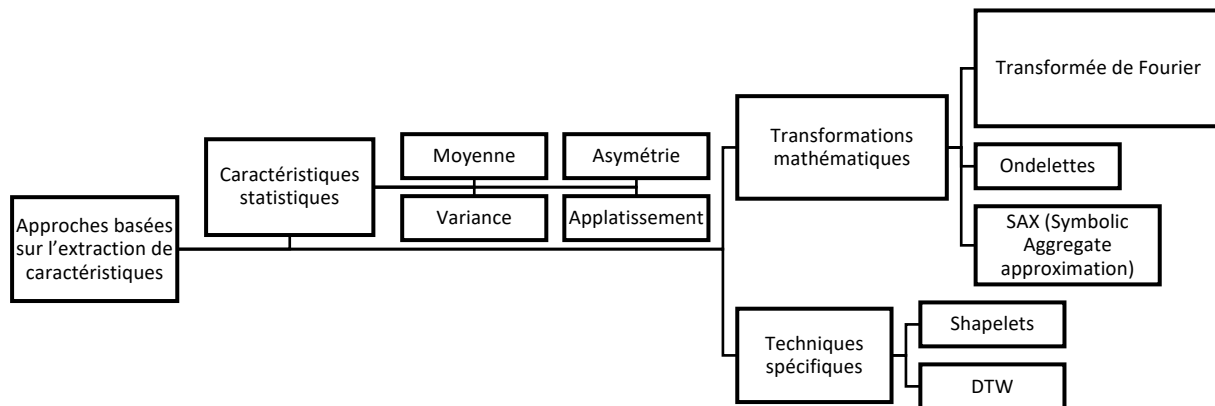


Figure 2-2 - Approches basées sur l'extraction de caractéristiques

L'extraction de caractéristiques est une technique utilisée pour réduire la dimensionnalité des données et capturer des caractéristiques significatives, ce qui permet aux algorithmes de regroupement d'identifier plus efficacement des modèles similaires.

1. **Les approches basées sur l'extraction de caractéristiques** ou (« *feature-based methods* ») consistent à représenter une série temporelle à l'aide de propriétés spécifiques, telles que la moyenne, l'écart-type, les caractéristiques spectrales ou d'autres descripteurs statistiques. Ces méthodes permettent de simplifier la représentation des séries temporelles, réduisant ainsi la complexité informatique tout en facilitant leur analyse et leur regroupement. Elles sont souvent utilisées en combinaison avec des algorithmes de classification ou de clustering pour améliorer la performance du processus.
2. **Les caractéristiques basées sur la forme et les modèles autorégressifs** consistent à calculer des propriétés statistiques telles que la moyenne, la variance, l'asymétrie et l'aplatissement, afin de décrire la forme globale de chaque série temporelle. Ces descripteurs permettent de comparer et de regrouper les séries temporelles en fonction de leurs similitudes structurelles.
 - a. Les modèles autorégressifs (AR), quant à eux, représentent les séries temporelles en les exprimant comme une combinaison linéaire de leurs valeurs passées. Les coefficients de ces modèles, qui capturent les dépendances temporelles au sein des

séries, peuvent être utilisés comme attributs pour regrouper des séries temporelles partageant des dynamiques similaires.

3. **La transformation de Fourier** est une opération mathématique qui convertit les données du domaine temporel au domaine fréquentiel. Cette méthode extrait des propriétés distinctives, telles que les composantes d'amplitude et de phase, qui permettent de mieux comprendre les motifs cycliques ou répétitifs des séries temporelles. Cette méthode est particulièrement utile pour détecter des fréquences dominantes dans les données.
4. **La transformation en ondelettes** est une technique permettant de décomposer une série temporelle en composantes de fréquence distinctes. Cette méthode préserve à la fois les informations temporelles et fréquentielles, ce qui la rend adaptée à l'analyse des variations locales. Les coefficients d'ondelettes obtenus peuvent être utilisés pour regrouper des séries temporelles présentant des modèles de fréquence similaires.
5. **L'approximation par agrégat symbolique (SAX)** transforme les données de séries temporelles en séquences symboliques. Chaque symbole correspond à une plage de valeurs prédéfinie, réduisant ainsi la dimensionnalité des données tout en conservant leurs motifs essentiels. Ces séquences symboliques peuvent ensuite être utilisées pour effectuer une segmentation ou un regroupement.
6. **Le « Dynamic Time Warping » (DTW)** est une méthode utilisée pour aligner deux séries temporelles en tenant compte de leurs déformations temporelles. Contrairement aux approches basées sur l'extraction de caractéristiques, DTW travaille directement avec les données brutes. L'algorithme identifie l'alignement optimal entre les points des deux séries en minimisant une distance cumulée, ce qui le rend particulièrement adapté aux séries présentant des variations de vitesse ou des décalages temporels.
7. **La transformation en shapelets** consiste à identifier des sous-séries discriminantes, appelées **shapelets**, dans un ensemble de données de séries temporelles. Ces shapelets servent de caractéristiques clés pour représenter et classifier des séries temporelles individuelles en fonction de leur similarité avec ces sous-séries.
8. **L'approche par sac de motifs (« Bag-of-Patterns, BoP »)** consiste à extraire des motifs représentatifs, ou sous-séquences, d'une série temporelle. Ces motifs sont regroupés pour construire un histogramme représentant la distribution des modèles dans chaque série. Ces

histogrammes servent ensuite à la segmentation ou au regroupement des séries temporelles en fonction de leurs distributions de motifs.

9. **L'approche connue sous le nom de « *GrammarViz* »** est une méthode basée sur des techniques de compression utilisant des grammaires pour analyser des séries temporelles. Chaque série est convertie en une grammaire sans contexte, permettant une compression efficace tout en préservant les motifs significatifs. Le regroupement est ensuite effectué sur ces représentations symboliques pour identifier et catégoriser les séries temporelles présentant des similitudes.

Les techniques basées sur les caractéristiques s'avèrent particulièrement adaptées lorsque les données brutes présentent une grande dimensionnalité ou lorsque des propriétés spécifiques sont identifiées comme essentielles pour capturer les modèles sous-jacents. Ces approches permettent de réduire la complexité des données, d'extraire des informations pertinentes et de faciliter l'application d'algorithmes de regroupement ou de classification. Le choix de la méthode d'extraction des caractéristiques et de l'algorithme de segmentation repose sur les propriétés intrinsèques des séries temporelles et sur les objectifs spécifiques de l'analyse, qu'il s'agisse d'une segmentation fine, de la détection de motifs ou d'une classification.

2.5.3 Modèles et algorithmes de segmentation

Le choix des modèles et des algorithmes de segmentation dépend des propriétés des séries temporelles, telles que la stationnarité, les dépendances temporelles ou la granularité, ainsi que des objectifs spécifiques de l'analyse, comme la segmentation, le regroupement ou la classification. Les approches décrites ci-dessous offrent des solutions adaptées pour répondre aux défis liés à l'analyse des séries temporelles complexes.

1. **Des modèles ARMA et algorithmes d'optimisation** es modèles ARMA (« *AutoRegressive Moving Average* ») sont souvent utilisés pour modéliser les séries temporelles stationnaires. Leur performance peut être améliorée lorsqu'ils sont associés à des techniques d'optimisation avancées, notamment :
 - a. **Tabu Search**, qui explore efficacement l'espace des solutions tout en évitant de revisiter des configurations déjà analysées;

- b. **Algorithmes génétiques (GA)**, inspirés des processus évolutifs, qui appliquent des principes tels que la mutation et le croisement pour optimiser les solutions ;
 - c. **Recuit simulé**, une méthode probabiliste permettant d'échapper aux optima locaux en simulant un processus de refroidissement progressif.
- 2. **Des modèles de lissage hiérarchique**: Ces modèles s'appuient sur des techniques de regroupement hiérarchique agglomératif, dans lesquelles les segments similaires sont regroupés de manière itérative. Ils sont particulièrement efficaces pour identifier des structures emboîtées ou hiérarchiques dans les séries temporelles.
- 3. **Des modèles de mélange gaussien** (« *GMM*, pour *Gaussian Mixture Models* ») modélisent les séries temporelles comme une combinaison de distributions gaussiennes, chaque composante représentant un sous-ensemble des données. L'algorithme **Expectation-Maximization (EM)** est utilisé pour ajuster les paramètres de ces distributions (moyenne, variance, poids). Il fonctionne en deux étapes répétées jusqu'à convergence :
 - a. **Étape d'espérance (E-step)** : Chaque point de données est attribué de manière probabiliste à une composante gaussienne, en fonction des paramètres actuels du modèle.
 - b. **Étape de maximisation (M-step)** : Les paramètres des distributions gaussiennes sont réajustés pour maximiser la vraisemblance des données assignées.
 - c. Cette combinaison de GMM et d'EM permet de segmenter efficacement les séries temporelles, même en présence de chevauchements significatifs ou de structures complexes.
- 4. **Des modèles autorégressifs et modèles de Markov cachés** : les modèles autorégressifs (AR) expriment une série temporelle comme une combinaison linéaire de ses valeurs passées. Les coefficients AR extraits permettent de capturer les dépendances temporelles et peuvent être utilisés comme caractéristiques dans les processus de regroupement. **Les modèles de Markov cachés** (« *HMM* ») analysent les séries temporelles en identifiant des états latents sous-jacents. Les HMM, qu'ils soient discrets ou continus, permettent une segmentation fine en modélisant les transitions entre les états non observables.

Les modèles statistiques (ARMA, GMM, HMM) et les techniques de transformation (shapelets, Fourier, ondelettes, SAX) offrent une complémentarité pour traiter les séries temporelles complexes. Leur application dépend des propriétés des données (stationnarité, dépendances temporelles, granularité, fréquence) et des objectifs d'analyse, qu'il s'agisse de segmentation, de regroupement ou de classification.

Le Tableau 2-2- Segmentation basé sur les modèles propose une synthèse des principales variables, modèles et algorithmes de segmentation utilisés dans la classification des séries temporelles. Ces outils permettent de maximiser la précision et l'efficacité des analyses, tout en s'intégrant harmonieusement dans des processus analytiques globaux.

Tableau 2-2- Segmentation basé sur les modèles

Variable	Modèle	Algorithme de segmentation
Univariée et multivariée	ARMA	Recherche Tabu, Algorithmes génétiques (GA), Recuit simulé
Univariée	Modèle de lissage hiérarchique	Segmentation hiérarchique agglomérative
Multivariée	Modèle de mélange gaussien	Algorithme EM
Univariée	Autorégressif	Regroupement génétique basé sur les médioïdes
Multivariée	HMM Continu	Quatre niveaux de recherche imbriqués
Multivariée	HMM Continu	Quatre niveaux de recherche imbriqués
Univariée	Coefficients Autorégressif	
Multivariée	HMM Discret (discrétisé par SOM)	DTW suivi d'une opération à point fixe

Tableau 2-2- Segmentation basé sur les modèles (suite et fin)

Variable	Modèle	Algorithme de segmentation
Univariée		Segmentation hiérarchique agglomérative
Univariée (Valeurs discrètes)	Chaîne de Markov	Segmentation agglomérative
Multivariée (Valeurs discrètes)	Chaîne de Markov	Segmentation agglomérative
Univariée	Modèle de mélange gaussien	« <i>Modified fuzzy-means</i> »
Univariée (discretisé par quantification vectorielle à partir des coefficients d'ondelettes)	HMM Discret	Algorithme EM
Univariée	Modèle Autorégressif à moyenne mobile	Algorithme EM

*HMM: « Hidden Markov Model & EM: Expectation-Maximization »

2.5.4 Complément : Approche COTE (Collection of Transformation Ensembles)

L'approche **COTE** « *Collection of Transformation Ensembles* » (COTE), introduite par Bagnall et al. (2017), vise à améliorer la précision et la résilience de la classification des séries temporelles en combinant plusieurs classificateurs. Chaque classificateur est formé sur une représentation distincte des données initiales, obtenue via des transformations variées des séries temporelles.

Les auteurs ont observé que différentes méthodologies de transformation présentent des degrés d'efficacité variables selon les catégories de données analysées. En combinant les résultats de ces classificateurs via une stratégie d'ensemble, COTE exploite efficacement les avantages offerts par ces diverses représentations, offrant ainsi une approche robuste et généralisable pour la classification des séries temporelles complexes.

Mesure de similarité et classification

Le choix d'une mesure de similarité adaptée est un élément fondamental pour la classification des séries temporelles, car il influence directement l'efficacité des algorithmes. Une mesure de similarité idéale doit capturer fidèlement les motifs et structures sous-jacents des données tout en restant résistante au bruit, aux distorsions temporelles et aux variations de longueur. Pour des séries longues ou lorsque la compréhension de leur structure est limitée, les méthodes basées sur la compression et les propriétés structurelles des données peuvent offrir une alternative robuste et sans paramètre (Esling & Agon, 2012).

Taxonomie des approches basées sur la distance

Enfin, Abanda et al. (2019) ont proposé une taxonomie des méthodes basées sur la distance pour la classification des séries temporelles, résumée dans la (Figure 2-3). Cette taxonomie distingue trois grandes méthodologies :

- 1) **Utilisation directe des distances dans les classificateurs K-NN**: la méthode la plus courante consiste à utiliser des mesures de distance préexistantes dans les classificateurs basés sur la proximité, comme le *K-Nearest Neighbors* (K-NN). Bien qu'efficace, cette approche est souvent sensible au bruit présent dans les ensembles de données d'apprentissage, ce qui peut en réduire la robustesse.
- 2) **Transformation des distances en vecteurs de caractéristiques**: cette approche consiste à convertir les séries temporelles en vecteurs de caractéristiques à l'aide de mesures de distance. Cela permet de relier les séries temporelles à des méthodes de classification traditionnelles. Cette stratégie se divise en trois sous-catégories :
 - Utilisation directe des distances comme vecteurs de caractéristiques.
 - Définition des caractéristiques à partir des distances par rapport à des modèles locaux.
 - Intégration des distances après transformation des séries dans des espaces vectoriels.
- 3) **Utilisation de noyaux de distance**: les noyaux de distance sont des structures construites à partir de mesures de distance, utilisées pour des algorithmes comme les machines à vecteurs de support (« *Support Vector Machines*, SVM »). Ces noyaux peuvent être divisés en deux catégories :

- Les **noyaux indéfinis**, qui n'ont pas de garantie de semi-définition positive, mais qui restent utiles pour des données complexes.
- Les **noyaux définis positifs**, prouvés comme étant semi-définis positifs, offrant une meilleure prise en compte des caractéristiques temporelles des séries et améliorant la précision de la classification.

La figure illustre les trois approches principales décrites ci-dessus, mettant en évidence leurs avantages et leurs limites respectifs, ainsi que leurs applications selon le type de données et les objectifs analytiques.

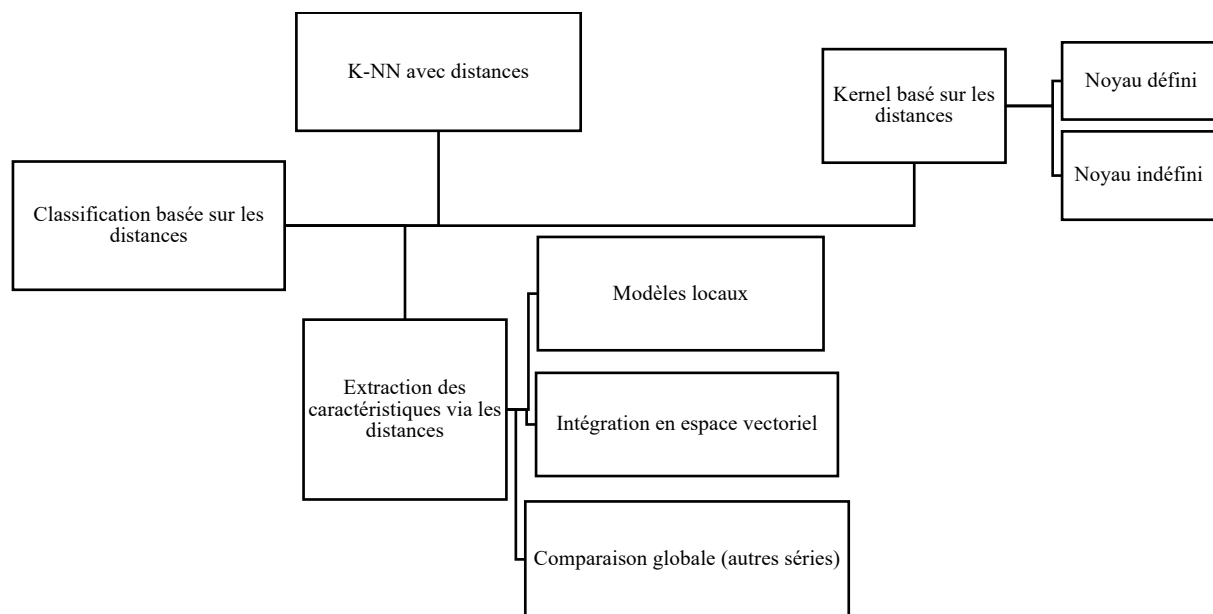


Figure 2-3 – Une taxonomie de la classification des séries temporelles basée sur la distance

2.6 Synthèse de la revue

La majorité des travaux existants (Morency et al. (2007), Lee et Kam (2014)) se sont concentrés sur l'analyse descriptive des données de cartes à puce, sans pleinement exploiter le potentiel des séries temporelles pour améliorer la planification des horaires. Les approches de segmentation (Morency et al. (2006)) permettent de distinguer des groupes d'utilisateurs aux comportements similaires. Toutefois, l'ajustement des horaires et du service en fonction des variations temporelles saisonnières restent largement inexplorés. Cette lacune est particulièrement notable dans un contexte où les schémas de mobilité sont de plus en plus influencés par des facteurs externes tels que les conditions météorologiques, les événements locaux, ou les crises sanitaires, comme celle

du COVID-19. En effet, l'intégration de ces deux types de données reste sous-explorée, surtout lorsqu'il s'agit de capturer la dynamique des comportements des usagers au fil du temps.

Même si des techniques de segmentation comme les modèles autorégressifs (He et al. (2017), Bagnall et al. (2017)) et des approches combinant des mesures de similarité, telles que le *Dynamic Time Warping (DTW)*, ont démontré leur efficacité, leur application au domaine du transport en commun présente des défis spécifiques. Par exemple, le **DTW**, bien qu'il ne soit pas une technique de segmentation, permet d'aligner des séries temporelles présentant des variations temporelles non linéaires. Cependant, les modèles et algorithmes actuellement utilisés ne capturent pas toujours la variabilité intrapersonnelle et interpersonnelle des comportements de déplacement (Deschaintres et al. (2018)). Cette variabilité, qui peut fluctuer d'un jour à l'autre ou même d'une heure à l'autre, rend difficile l'extraction de segments exploitables pour la planification des services.

Face à ces limitations, ce mémoire de recherche se propose d'améliorer la planification des horaires de transport en commun en intégrant des techniques avancées de classification et de segmentation des séries temporelles appliquées aux données de cartes à puce. Plusieurs axes seront explorés pour surmonter les défis identifiés:

1. **L'utilisation des techniques de transformation et de segmentation avancées** : ce travail utilise des transformations de données, telles que la normalisation ou la transformation logarithmique, pour préparer les données de cartes à puce à une analyse approfondie. Ces transformations rendent les schémas de déplacement plus comparables et augmentent la précision des segments obtenus. Par exemple, la transformation logarithmique atténue l'impact des valeurs extrêmes, fréquentes dans les données de validation des cartes à puce.
2. **L'analyse de divers objets du réseau et de fenêtres temporelles flexibles**: contrairement aux approches traditionnelles qui utilisent des segments temporels fixes, ce projet examine des objets variés du réseau (lignes, arrêts, modes de transport) en appliquant des fenêtres temporelles flexibles. Cette méthode capture des schémas de mobilité complexes et permet d'ajuster les horaires de manière plus fine, selon les fluctuations de la demande à différents moments de la journée ou de la semaine.
3. **La combinaison du DTW avec les K-moyennes et la segmentation hiérarchique** : le projet associe le DTW, une méthode d'alignement efficace pour gérer des variations non linéaires entre séries temporelles, avec des techniques de regroupement comme les K-

moyennes et la segmentation hiérarchique. Tandis que le DTW aligne les séries, les K-moyennes et la segmentation hiérarchique regroupent les données en segments exploitables. Cette combinaison vise à pallier les limites des approches traditionnelles et à capturer les nuances des schémas de déplacement.

4. **L'exploration de distances alternatives combinées à ces techniques** : en complément du DTW, le projet explore d'autres mesures de similarité, comme la distance cosinus ou la distance de Manhattan, pour capturer les tendances non stationnaires et les fluctuations fines des données. Cette approche vise à déterminer quelles distances sont les plus efficaces pour segmenter les séries temporelles en fonction de critères spécifiques, afin de produire des segments plus précis et utiles pour la planification des services.

Enfin, l'approche adoptée se distingue par sa profondeur analytique et son potentiel d'application pratique, avec pour ambition d'avoir un impact concret sur l'amélioration des horaires dans les réseaux de transport en commun.

CHAPITRE 3 MÉTHODOLOGIE ET DESCRIPTION DES DONNÉES

3.1 Méthodologie de recherche

La méthodologie développée dans ce projet allie des approches théoriques et pratiques pour traiter efficacement les enjeux de la demande de transport. Structurée en six étapes fondamentales, elle fournit un cadre analytique précis et progressif pour explorer les données. Ces étapes sont résumées dans la Figure 3-1, qui illustre leur articulation et leur enchaînement.

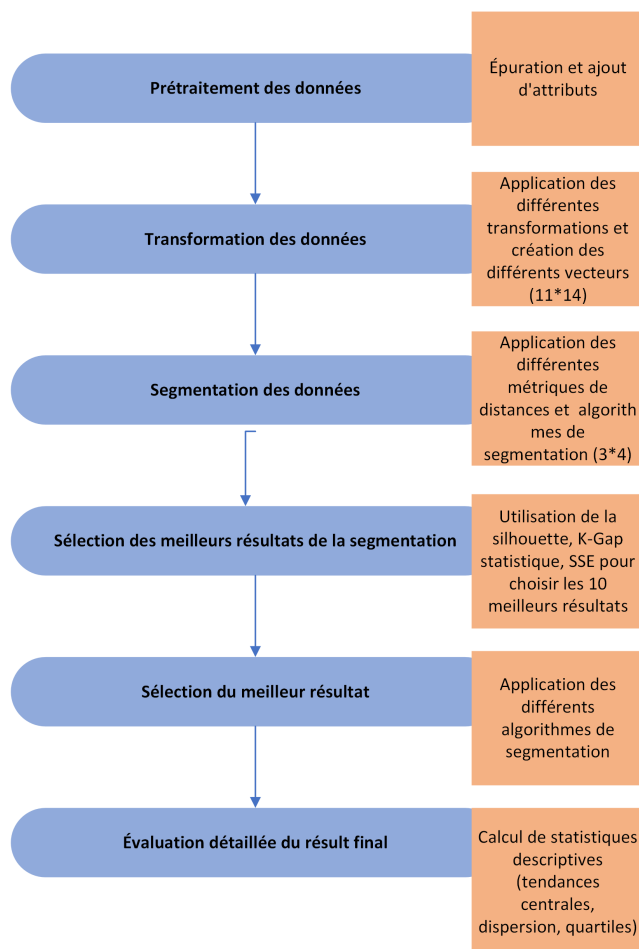


Figure 3-1 - Méthodologie générale du mémoire

Le processus débute par un prétraitement rigoureux des données. Les informations issues des cartes à puce sont enrichies, filtrées, agrégées et validées, tandis que les données GTFS sont nettoyées et

intégrées pour constituer une base consolidée. Cependant, la dépendance aux cartes à puce peut introduire des biais, car ces données ne reflètent pas nécessairement la diversité complète des usagers. Elles capturent principalement les habitudes des utilisateurs réguliers, négligeant potentiellement d'autres catégories telles que les touristes ou les usagers occasionnels. Cette limitation peut affecter la représentativité des résultats. Par ailleurs, bien que les données GTFS apportent une vue essentielle des horaires planifiés, elles n'intègrent pas les perturbations en temps réel, telles que les retards ou annulations de service. Cette absence limite la capacité à évaluer avec précision l'adéquation entre l'offre réelle et la demande observée. Ces données permettent néanmoins d'identifier des objets clés du système de transport, tels que le réseau, les modes, les lignes et les arrêts, qui servent ensuite à créer des vecteurs spécifiques (réseau, mode, ligne, arrêt) pour les analyses futures. De plus, ces vecteurs peuvent constituer une base pour des recommandations pratiques, telles que l'ajustement des fréquences ou l'optimisation des lignes sous-utilisées. Cette étape garantit la qualité des données tout en posant les bases d'une exploration approfondie des aspects essentiels du réseau de transport.

Adapter les données brutes aux objectifs analytiques constitue une étape clé dans cette méthodologie. Les transformations appliquées permettent de répondre à différents besoins. Par exemple, la normalisation (standardisation ou Min-max) homogénéise les échelles, facilitant les comparaisons. En parallèle, des techniques comme la différenciation ou les transformations logarithmiques atténuent l'effet des valeurs extrêmes. Enfin, des approches plus avancées, telles que Box-Cox ou Yeo-Johnson, ajustent les distributions pour une meilleure interprétation. Ces transformations sont centralisées dans un dictionnaire, garantissant leur traçabilité et leur exploitation cohérente.

Les méthodes de segmentation reposent sur des techniques de liaison (simple, complète, moyenne, Ward) associées à des mesures de distance, telles que l'euclydienne, Manhattan, cosinus, ou DTW (« *Dynamic Time Warping* »). Bien que DTW ne soit pas une distance conventionnelle, elle est exploitée comme une mesure de similarité pour analyser les séries temporelles. Les algorithmes classiques, tels que K-moyennes et la segmentation hiérarchique, sont adaptés pour exploiter DTW, qui enrichit l'analyse en capturant les relations temporelles complexes. Cette intégration permet de regrouper des séries similaires en fonction de leur dynamique plutôt que de simples proximités statiques. Ces combinaisons, comme DTW appliqué à K-moyennes ou à la segmentation hiérarchique, produisent des regroupements adaptés aux données temporelles complexes. Les

résultats obtenus sont stockés dans un dictionnaire pour faciliter leur évaluation et leur interprétation.

L'évaluation des résultats de segmentation s'appuie sur des critères variés. Les mesures internes, comme WCSS et le score de silhouette, examinent la cohésion et la compacité des groupes. Les indices tels que K-Gap et le K optimal permettent de déterminer le nombre de clusters le plus pertinent. En complément, des métriques globales, comme Davies-Bouldin et Calinski-Harabasz, analysent la qualité des regroupements en prenant en compte leur séparation et leur compacité. Ces outils d'évaluation permettent de sélectionner les configurations les plus adaptées, garantissant une segmentation robuste et significative.

Les modèles finaux sont visualisés pour faciliter leur interprétation. Les histogrammes et courbes de densité permettent de comprendre la répartition des données, tandis que les cartes de chaleur et les diagrammes de dispersion mettent en évidence les relations entre variables. Les boîtes à moustaches aident à identifier les écarts ou anomalies au sein des groupes. Bien que cette analyse offre des perspectives détaillées, il est important de noter que les résultats obtenus reflètent spécifiquement les données de 2020, une période influencée par des facteurs externes tels que la pandémie de COVID-19. Cela restreint la portée des conclusions, limitant la possibilité d'extrapoler les tendances à long terme ou de saisir les ajustements structurels qui ont pu émerger après cette période. Une perspective pourrait inclure l'analyse de données supplémentaires provenant de périodes postérieures, offrant ainsi une compréhension plus robuste des changements de comportement des usagers. Une analyse intégrant des données postérieures pourrait compléter ces observations et offrir une vision plus équilibrée. Ces visualisations sont accompagnées d'une analyse statistique détaillée, incluant des paramètres tels que la moyenne, l'écart-type et l'asymétrie. Ces outils fournissent des informations clés pour formuler des recommandations pratiques et adaptées aux segments obtenus.

La méthodologie proposée permet une approche cohérente et complète des données de transport en commun. Elle offre également une perspective d'amélioration continue, en intégrant des données en temps réel ou en combinant différentes sources pour enrichir les analyses. En combinant des étapes rigoureuses de prétraitement, transformation, segmentation et évaluation, elle garantit une analyse approfondie des comportements de déplacement, tout en fournissant une base solide pour des recommandations opérationnelles pertinentes.

3.2 Cas d'étude

Le projet de recherche, mené en collaboration entre Kéolis France et Polytechnique Montréal, vise à définir de nouvelles typologies calendaires pour optimiser l'offre et les horaires du réseau TAO à Orléans, France. Située dans la région Centre-Val de Loire, Orléans est une ville où les enjeux liés à l'ajustement des horaires de transport en commun sont particulièrement importants.

Initialement concentré sur l'agglomération orléanaise et le réseau TAO, le projet pourrait s'étendre à d'autres régions et centres d'exploitation en fonction des résultats obtenus, avec les ajustements nécessaires pour s'adapter aux contextes locaux.

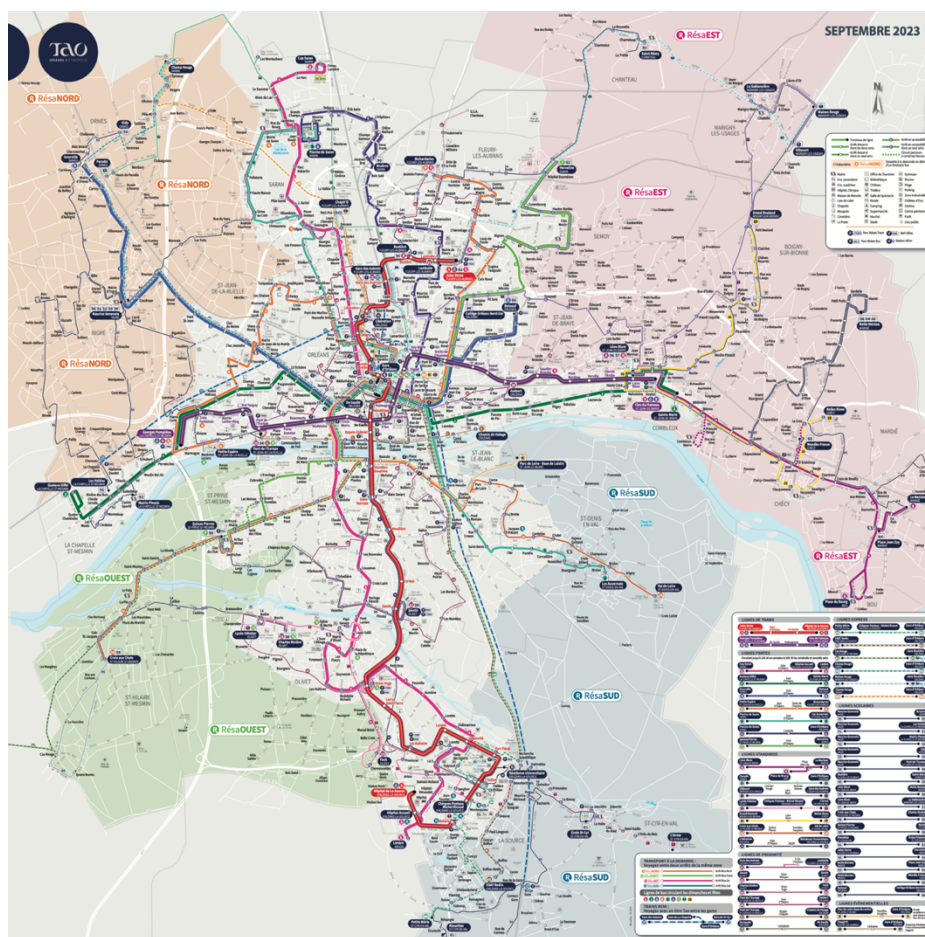


Figure 3-2 - Plan du réseau d'Orléans (France)

Source : TAO mobilité, 2023

Le réseau TAO (Transports de l'Agglomération Orléanaise) est responsable de la gestion des transports en commun pour la ville d'Orléans et ses environs (Figure 3-2). Il se compose

principalement de lignes de bus et de tramways qui desservent les vingt-deux communes de la métropole orléanaise. Ce réseau est complété par un service de location de vélos en libre-service, baptisé Vélo'+, conçu pour des trajets de courte durée.

En complément, le titre P+R (Parking + Relais) facilite la multimodalité en permettant aux automobilistes de stationner et d'effectuer un aller-retour dans la journée. Ce titre couvre l'ensemble des passagers de la voiture voyageant ensemble, dans la limite de sept personnes, conducteur compris.

Le tramway moderne d'Orléans dessert plusieurs quartiers de la ville et constitue un moyen de transport à la fois populaire et efficace pour les déplacements urbains. Le réseau de bus, quant à lui, couvre un vaste territoire, reliant les quartiers résidentiels aux centres commerciaux, aux établissements publics et aux principaux points d'intérêt de la métropole.

Grâce à une organisation en lignes complémentaires et interconnectées, le réseau de transport en commun TAO assure une couverture efficace de l'ensemble du territoire, facilitant les déplacements des habitants et des visiteurs.

Le tramway et les lignes de bus sont autant fréquentés (Figure 3-3).

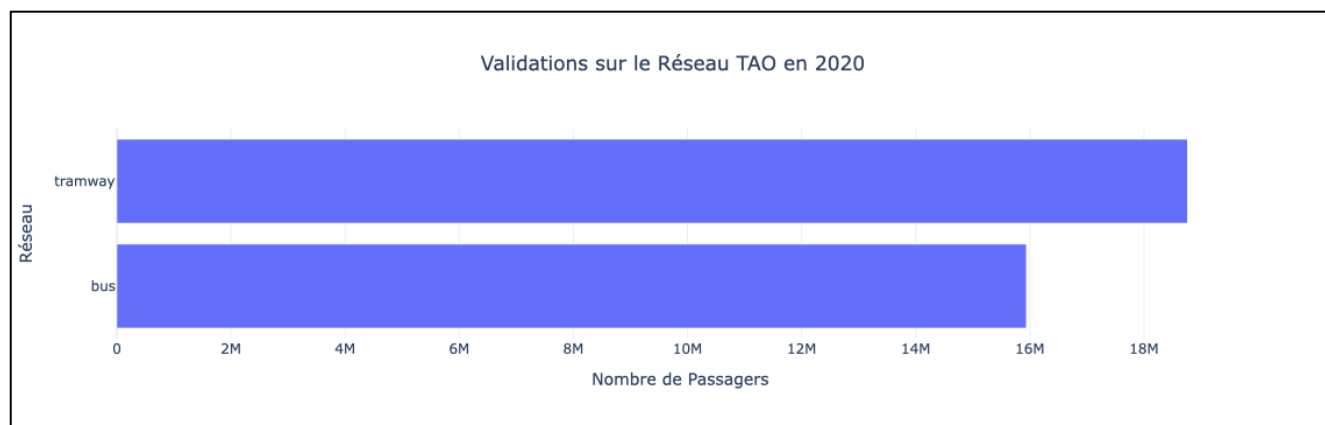


Figure 3-3 - Validations totales sur le réseau TAO en 2020

Composition du réseau

Le réseau TAO d'Orléans, géré par Keolis, dessert les 22 communes de la métropole orléanaise. Il comprend 29,3 km de lignes de tramway (Lignes A et B reliant les axes nord-sud et est-ouest) et des lignes structurantes, circulant de 5h à 23h (jusqu'à 1h les vendredis et samedis) avec des fréquences de 10 à 20 minutes, complétant les tramways. Les lignes standards, incluant la ligne nocturne N, desservent la plupart des communes. Les navettes communales sont regroupées sous les lignes de proximité. Les lignes périurbaines sont numérotées de 33 à 37, tandis que les lignes express (40 à 45) incluent des lignes Rémi interurbaines. Les lignes scolaires (50 à 69) desservent les établissements en périphérie. Le réseau inclut aussi une ligne événementielle (CO), une ligne d'été (IC), des services vers les gares (SNCF Ter Bus) et des bus relais (LA et LB) pour les trams A et B en cas d'arrêt. Le tramway et les lignes structurantes représentent 95 % des trajets.

Utilisateurs du réseau

- Les utilisateurs du réseau TAO d'Orléans¹ sont divers et incluent plusieurs groupes principaux. Les étudiants, attirés par les nombreuses universités et écoles supérieures, utilisent fréquemment les transports en commun pour se rendre sur leur campus ou en ville. Les professionnels et actifs, utilisant quotidiennement bus et tramways, se déplacent pour le travail et les réunions. Les résidents des banlieues proches accèdent régulièrement au centre-ville pour des raisons professionnelles, d'achats ou de loisirs. Les personnes âgées et retraitées se déplacent pour accéder aux services, commodités et activités sociales. Les touristes, attirés par le riche patrimoine historique et culturel d'Orléans, explorent la ville et ses attractions en utilisant les transports en commun. Enfin, les familles avec enfants se servent des bus et tramways pour les trajets scolaires, les activités parascolaires et les rendez-vous médicaux.

¹ Tao. (s. d.). *Tao*. <https://www.tao-mobilites.fr/fr/>

3.3 Données

Dans ce mémoire, les données utilisées proviennent de diverses sources essentielles pour l'analyse des services de transport en commun du réseau TAO d'Orléans. Ces données comprennent les informations collectées par les cartes à puce, les données GTFS ainsi que les données météorologiques, les périodes de vacances scolaires, les périodes de confinement et les données d'exploitation de TAO. Les sections suivantes détaillent chacune de ces sources de données, leur pertinence et leur utilisation spécifique dans le cadre de cette recherche.

3.3.1 Données de transaction de paiement

Les données utilisées pour ce travail proviennent des valideurs de titres de transport. Ces valideurs peuvent scanner efficacement et rapidement les titres de transport dotés de codes-barres 2D affichés sur des smartphones, ainsi que les cartes à puce et les cartes EMV (Europay, Mastercard, et Visa) sans contact. Lorsqu'une carte EMV est utilisée, un code de transaction unique est créé, ajoutant une couche de sécurité supplémentaire. Les données couvrent la période du 1er janvier 2020 au 31 décembre 2020, totalisant 32 972 470 observations (Tableau 3-1).

Tableau 3-1- Informations recueillies à chaque validation tarifaire du réseau TAO (à l'embarquement)

	Métro	Bus
Identifiant carte (anonymisée)	√	√
Véhicule	√	√
Sens	√	√
Courses	√	√
Code Produit	√	√
Type Transaction	√	√
Ligne	√	√
Code arrêt	√	√
Libellé de l'arrêt	√	√
Date et heure	√	√

3.3.2 Données GTFS

Le format GTFS (« *General Transit Feed Specification* ») représente les données sur les horaires, les itinéraires et les arrêts de bus à itinéraire fixe dans une série de fichiers de données délimités par des virgules. La spécification GTFS illustré par la Figure 3-4 définit un format commun pour les horaires des transports publics et les informations géographiques associées telles que l'emplacement des gares. Cela permet aux données de transport en commun d'être mises à disposition dans un format standardisé, facilitant ainsi leur analyse par d'autres intéressés, ou leur utilisation par les développeurs pour créer des applications utiles.

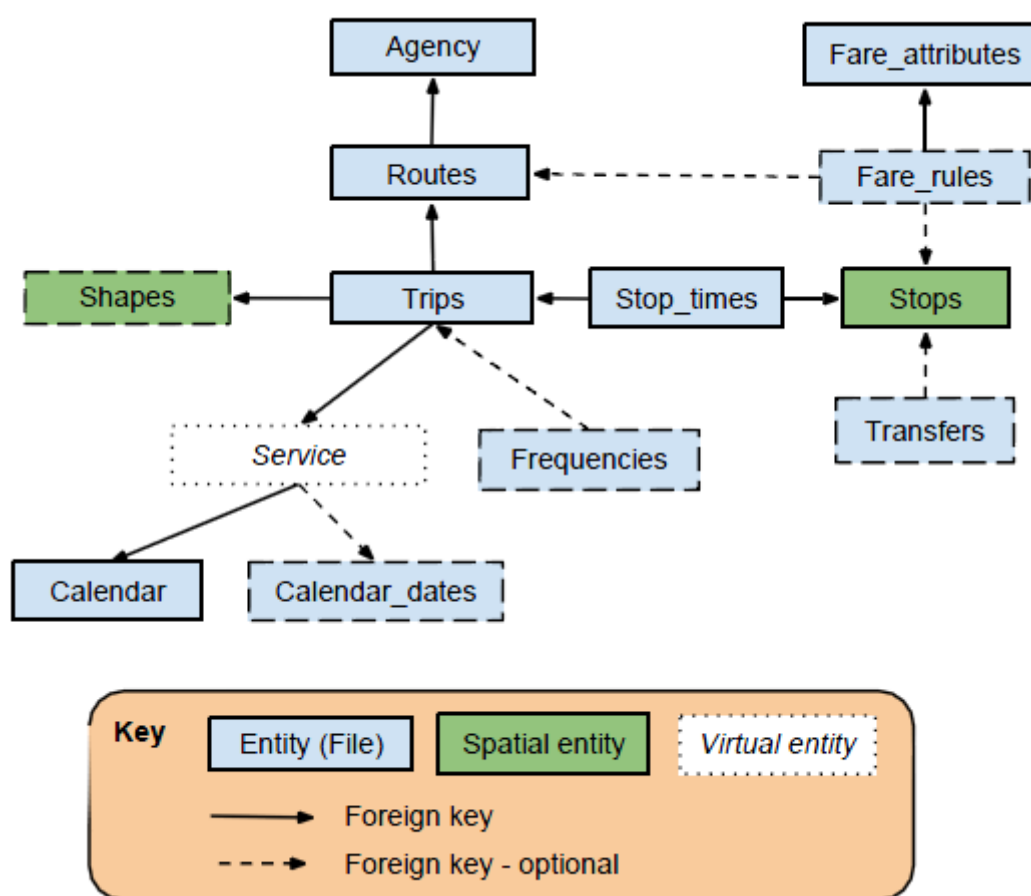


Figure 3-4 - Schéma simplifié d'un fichier GTFS

Source: Martin, David. Data model diagrams for GTFS

Les données GTFS (General Transit Feed Specification) des réseaux urbains et interurbains de la région Centre-Val-de-Loire, brutes pour l'année 2020, représentent un total de 2,04 Go. Ces

données, obtenues à partir du portail ouvert des données de transport, sont composées de neuf fichiers :

1. **agency.txt** : englobe des informations sur l'agence et le fuseau horaire de la ville où le transport en commun fonctionne.
2. **stops.txt** : comporte les noms des stations/arrêts de transport en commun, l'identifiant de l'arrêt et les géolocalisations (latitude, longitude).
3. **trips.txt** : comprend les directions de chaque mouvement du véhicule. Chaque voyage possède un identifiant de service qui précise les jours pendant lesquels il fonctionne.
4. **routes.txt** : renferme les informations sur la façon dont les voyages sont regroupés en services uniques.
5. **stop_times.txt** : contient les informations sur l'heure d'arrivée et de départ à chaque station/arrêt de transport en commun, ainsi que la séquence des stations et les identifiants de trajet correspondants.
6. **calendar.txt** : inclut les informations sur les jours de service hebdomadaires référencés par les déplacements.
7. **calendar_dates.txt** : comporte les informations sur les services hebdomadaires qui étendent calendar.txt et comprend les exceptions sur le service défini dans calendar.txt.
8. **shapes.txt** : comprend des « polygones » géographiques représentant les itinéraires empruntés par un véhicule de transport en commun.
9. **transfers.txt** : spécifie des règles et des remplacements supplémentaires pour les transferts sélectionnés.

Ces fichiers permettent de fournir une représentation complète et détaillée des services de transport en commun, incluant les horaires, les itinéraires et les arrêts, ainsi que les données spatiales et les exceptions de calendrier.

Ces données seront utilisées pour comprendre la fréquence de passage des bus et le nombre de passages aux arrêts, identifier les arrêts très fréquentés et ajuster les horaires de transport en commun ainsi que l'allocation des ressources. La carte ci-dessous l'intensité de l'utilisation des arrêts (nombre de validations de titres de transport effectuées à chaque arrêt) au sein de la Ville

d'Orléans pour l'année 2020, obtenue à partir de kepler.gl (un outil d'analyse géospatiale open source pour les ensembles de données à grande échelle).

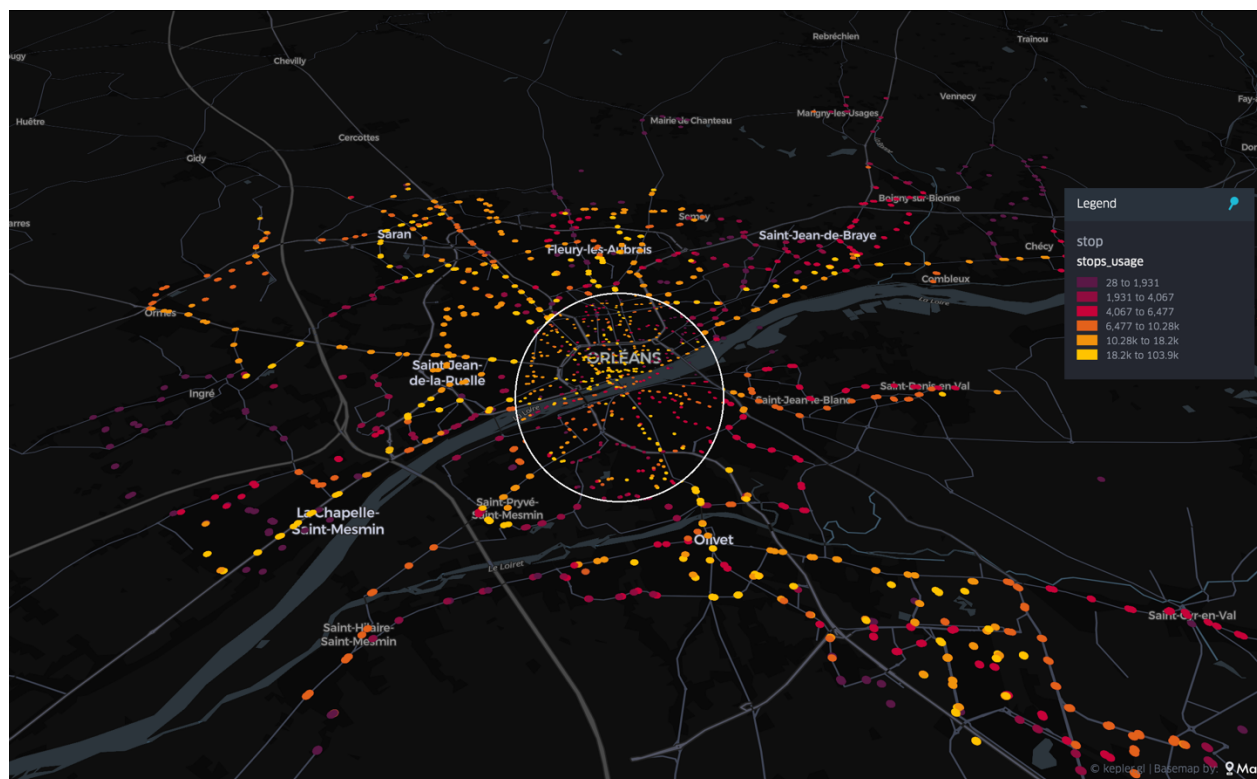


Figure 3-5 - Intensité d'utilisation des arrêts à Orléans (Données GTFS)

Ce travail examine en détail les données GTFS afin de déterminer les jours de service et de non-service pour chaque itinéraire de transport en commun. Les tables GTFS pertinentes — « *trips* », « *calendar* » et « *calendar_dates* » — sont initialement chargées depuis la base de données SQLite de l'environnement de travail. Les colonnes « *start_date* » et « *end_date* » sont converties du format caractères au format « *datetime* » pour générer des plages de dates précises pour chaque période de service.

En itérant sur ces plages de dates, une liste complète des jours de service potentiels est créée pour chaque ID de service, en incluant le jour de la semaine correspondant et les indicateurs de jour de service (du lundi au dimanche). Les indicateurs de jours de service, initialement des caractères, sont convertis en nombres entiers pour garantir un filtrage précis. Les jours de service générés sont ensuite filtrés pour ne conserver que les dates correspondant aux jours de service actifs de la semaine.

Pour tenir compte des exceptions spécifiées dans la table « *calendar_dates* », des dates de service supplémentaires sont incorporées et les dates spécifiques de non-service sont supprimées comme définies. Après l'application de ces exceptions, les jours de service filtrés sont associés à leurs itinéraires respectifs en fusionnant ces données avec le tableau « *trips* ». Cette étape permet de relier chaque jour de service à son itinéraire correspondant.

Enfin, les données par itinéraire et ID de service sont regroupées pour dresser une liste de tous les jours de service actif pour chaque itinéraire. Les jours sans service sont identifiés en déterminant les dates non incluses dans les jours de service, en tenant compte de la plage de dates globale.

Le résultat de cette analyse est un ensemble de données détaillé comprenant les identifiants uniques pour chaque itinéraire et service, des listes de jours de service actifs et des listes de jours sans service. Cet ensemble de données fournit des informations précieuses sur les horaires des services de transport en commun, permettant une meilleure planification et gestion des opérations et garantissant une compréhension claire des modèles de service (représentation structurelle et fonctionnelle de la manière dont un réseau de transport est organisé, planifié, et opéré pour répondre aux besoins de déplacement d'une population dans une zone géographique donnée) sur l'ensemble du réseau d'Orléans.

3.3.3 Données météorologiques et calendrier des jours fériés

Des données météorologiques et le calendrier des jours fériés viennent compléter la base de données.

Pour collecter les données météorologiques utilisées dans cette analyse, plusieurs étapes et API sont mises en œuvre afin d'obtenir un ensemble de données complet. Tout d'abord, l'API Météostat est utilisée pour collecter des données météorologiques quotidiennes pour Orléans, en France. La période considérée s'étend du 1er janvier 2020 au 31 décembre 2020, avec les coordonnées géographiques d'Orléans définies comme point de référence. Les données météorologiques quotidiennes, notamment la température moyenne, sont récupérées à l'aide de la fonction « *Daily* » de la bibliothèque Météostat.

Ensuite, le jeu de données est enrichi en intégrant les jours fériés en France grâce à la bibliothèque « *Workalendar* ». Cette étape implique de déterminer la plage de dates dans l'ensemble de données

et d'identifier les jours fériés au sein de cette plage. Ces jours fériés sont marqués dans l'ensemble de données pour faciliter l'analyse.

De plus, les périodes de confinement dues au COVID-19 sont ajoutées à l'ensemble de données. Cela implique de définir un statut initial à « 0 » pour les jours hors confinement et de mettre à jour le statut à « 1 » pour les dates correspondant aux périodes de confinement. Cette approche globale garantit que l'ensemble de données capture non seulement les conditions météorologiques, mais également les informations contextuelles pertinentes pour la période d'étude.

3.4 Préparation des données

Les données obtenues pour la réalisation de ce travail proviennent de plusieurs fichiers .csv représentant chaque mois de validation sur le réseau TAO, qui ont été agrégés en un seul jeu de données couvrant l'ensemble de l'année 2020. De plus, les données GTFS, initialement constituées de plusieurs fichiers « .zip » pour l'année 2020, ont également été consolidées pour créer une base de données unique pour cette période. Il convient également de souligner que les données de l'année 2019 ont été recueillies à des fins de comparaison, permettant ainsi une analyse plus approfondie des variations et tendances sur deux années consécutives.

3.4.1 Pré-traitement des données

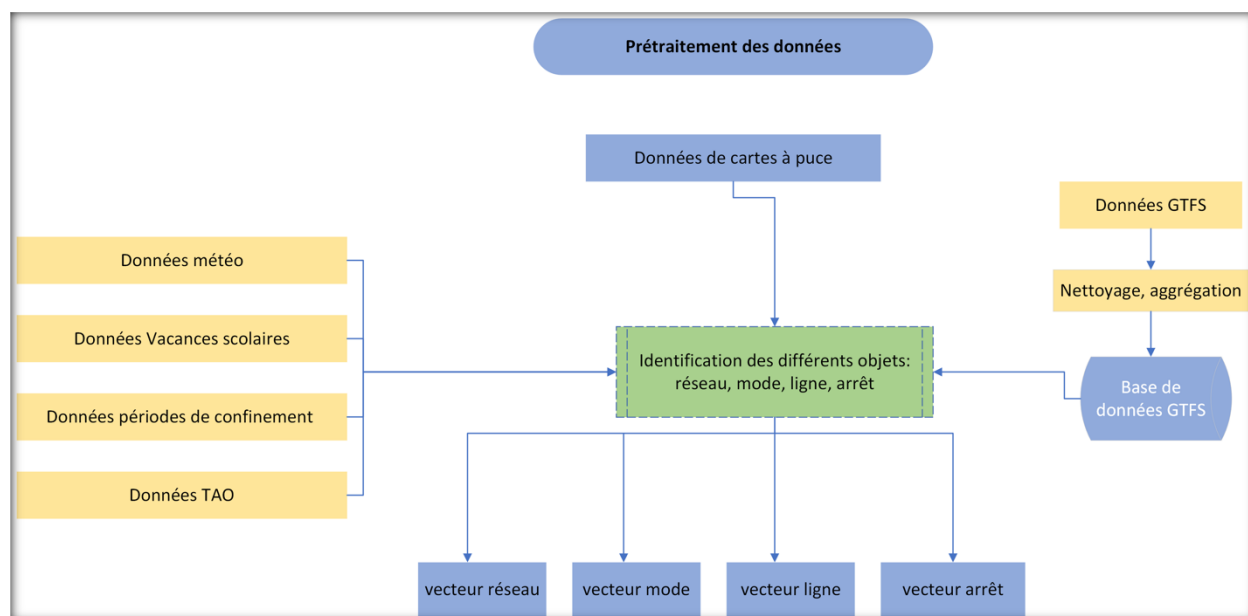


Figure 3-6 - Étape de prétraitement des données

Les données obtenues à partir des fichiers ont été tabulées en utilisant la bibliothèque Pandas du logiciel Python (Figure 3-6). À partir des informations recueillies sur le site tao-mobilites.fr de l'agence, les validations en dehors des heures de service, c'est-à-dire entre 5 heures du matin et 23 heures, ont été supprimées. Cette suppression permet d'éliminer les plages horaires où il n'y a pas d'événements à observer, et donc des points de données non-informatifs qui n'apporteraient pas d'informations significatives à l'analyse. De plus, cela rend les tâches d'analyse, souvent gourmandes en calcul, potentiellement plus réalisables. Les données obtenues renseignent sur le nombre de validations à chaque quart d'heure, avec un pas horaire de 15 minutes (Tableau 3-2).

Tableau 3-2 - Validations de l'année 2020

Datetime	Carte	Vehicule	sens	courses	Code_Produit	Type_Transaction	Ligne	Code_arret	Lib_arret	Nb_Passager
2020-05-22 09:15:00	0	41	0	1466	1	Compostage Magnétique	A	TCOLIG1	Coligny	1
2020-05-28 13:15:00	0	41	0	1466	1	Compostage Magnétique	A	TCOLIG1	Coligny	1
2020-05-22 09:00:00	0	41	0	1466	1	Compostage Magnétique	A	TCSMAR1	Saint-Marceau	1
2020-05-19 15:45:00	0	41	0	1466	1	Compostage Magnétique	A	TGAULA1	De Gaulle	1
2020-05-28 12:30:00	0	41	0	1466	1	Compostage Magnétique	A	THOPAC1	Hopital-Accueil	1

Une fois ce jeu de données obtenu, la première démarche consiste à ajouter des informations additionnelles sur les types de lignes en fonction de la classification faite par l'agence mentionnée plus haut (structurantes, proximité, etc.). Ensuite, une distinction est faite pour séparer les lignes de tramway des bus, ainsi que les bus de remplacement des trams. La composition finale du jeu de données obtenu est la suivante :

- **Type de ligne** : Classification des lignes (structurantes, proximité, etc.)
- **Mode de transport** : Distinction entre les lignes de tramway et de bus
- **Remplacement** : Identification des bus de remplacement pour les trams

Ce jeu de données enrichi permet une analyse plus détaillée et spécifique des différentes composantes du réseau de transport en commun.

Tableau 3-3- Composition du jeu de données

Catégories	Description	Total (unique)
Cartes	Carte de l'utilisateur	97 653

Tableau 3 3- Composition du jeu de données (suite et fin)

Véhicule	Numéro du véhicule	258
Sens	Sens du trajet	2
Courses	Numéro de la course	368
Code Produit	Code du produit	199
Type Transaction	Type de transaction (compostage magnétique, correspondance etc.)	14
Ligne	Numéro de la ligne	46
Code arrêt	Code de l'arrêt	1 457
Lib arrêt	Nom complet de l'arrêt	754

3.4.2 Composition des vecteurs pour l'analyse de la demande

Afin d'étudier la demande des passagers pour l'année 2020 plusieurs agrégations avec des pas horaires différents furent considérées :

1) Analyse spécifique au réseau

- **Agrégation horaire quotidienne** : Un vecteur « jour-heure » pour révéler des informations telles que des modèles de comportement similaires en semaine par rapport à la fin de semaine, et identifier des jours inhabituels s'écartant des modèles typiques.
- **Une deuxième agrégation demi-horaire** : un vecteur « jour-demi-heure » pour identifier les moments de la journée où la demande est la plus élevée, notamment les heures de pointe du matin et du soir. Cela permet de se focaliser sur les tendances et d'identifier les jours de forte ou faible fréquentation, en tenant compte des horaires de travail, des événements hebdomadaires et des habitudes des usagers.

2) Analyse spécifique au mode

- **Vecteurs bus et trams combinés** : deux vecteurs « comptage horaire des passagers des bus » et « comptage horaire des passagers des trams » pour analyser la demande globale de chaque mode. Cette analyse permet de comprendre les cycles de demande uniques et d'améliorer la répartition des ressources entre les bus et les tramways.

3) Analyse temporelle spécifique aux lignes

- **Vecteurs bus et trams** : Chaque groupe de lignes de bus et chaque tram séparé (trams A et B) est représenté par un vecteur distinct de « comptage horaire des passagers », indiquant combien de passagers utilisent chaque ligne de bus et chaque tram à différents moments de la journée.

4) Analyse temporelle spécifique aux itinéraires

- **Un vecteur** « comptage horaire des passagers par ligne et sens » (jour – ligne – sens – heure) afin d’approfondir la dynamique opérationnelle détaillée de chaque itinéraire permettant une compréhension ciblée du fonctionnement des différents itinéraires tout au long de la journée. Il contribuera à identifier les itinéraires les plus fréquentés et les périodes de pic d'utilisation, ce qui est crucial pour la gestion de la congestion et l'amélioration de l'expérience des passagers.

5) Analyse spécifique aux arrêts

- **Vecteur par arrêt** : Un vecteur « comptage horaire des passagers par arrêts » (jour – ligne – sens – code arrêt) pour évaluer le flux de passagers à des endroits spécifiques, identifier les zones à fort trafic, les goulots d'étranglement potentiels et planifier des ajustements de service et des améliorations des infrastructures.

3.5 Analyse descriptive

L'analyse exploratoire des données (« *Exploratory Data Analysis* ») est une étape critique dans la compréhension et l'interprétation des données avant d'appliquer des modèles statistiques ou d'apprentissage automatique. Ce chapitre détaille le processus EDA pour l'ensemble de données contenant les validations des passagers et les données GTFS (« *General Transit Feed Specification* ») de l'année 2020. L'analyse est conçue pour découvrir des modèles, des tendances et des anomalies dans les données, ce qui est crucial pour une compréhension adéquate de la demande et prise de décision efficace dans les services de transport d'Orléans.

La première étape du processus EDA dans le cadre de ce travail fut le nettoyage et le prétraitement des données. Cela garantit que l'ensemble de données est précis, cohérent et prêt pour l'analyse. Les valeurs manquantes sont traitées de manière appropriée pour éviter des résultats faussés. De plus, les types de données sont convertis dans des formats appropriés. Par exemple, les informations de date et d'heure sont converties au format « *datetime* », permettant des analyses

temporelles plus efficaces. De plus, pour faciliter une analyse plus approfondie, de nouvelles variables (« *features* ») sont créées. Des colonnes pour le jour de la semaine (« *DayOfWeek* ») et la date (« *Date* ») sont ajoutées. Ces variables permettent d'explorer les modèles quotidiens et hebdomadaires dans les validations des passagers.

Par ailleurs, les données sont regroupées par semaine qui fournit une vue d'ensemble des tendances des passagers au fil du temps. En résumant le nombre de validations de passagers (« *Nb Passenger* ») pour chaque semaine, les tendances et les modèles qui se produisent sur une base hebdomadaire ont pu être identifiés. Cette agrégation aide à comprendre l'utilisation globale et peut mettre en évidence les semaines avec un nombre de passagers inhabituellement élevé ou faible, ce qui pourrait être en corrélation avec des événements ou des variations saisonnières.

De même, le regroupement des données par jour permet d'analyser les modèles d'utilisation quotidienne. En calculant les validations quotidiennes, des variations au cours d'une semaine, identifiant les jours les plus chargés, ont pu être observées. Ces informations éclairent les décisions concernant l'allocation des ressources, comme l'augmentation du nombre de bus les jours où le nombre de passagers est plus élevé.

Puis, la visualisation des données est utilisée pour découvrir des tendances et des informations qui pourraient ne pas être immédiatement apparentes à partir de chiffres bruts. Plusieurs tracés sont créés pour illustrer les modèles temporels dans l'ensemble de données :

- 1) **Validations par semaine** : une image claire de la façon dont le nombre de passagers évolue au fil du temps sur une base hebdomadaire ; permettant aussi de révéler des tendances telles que des augmentations pendant les périodes de vacances ou des baisses pendant les périodes de mauvais temps.
- 2) **Validations quotidiennes** : permet de visualiser le total des validations quotidiennes, offrant une vue granulaire des fluctuations quotidiennes du nombre de passagers. Cela aide à identifier les journées les plus chargées et les moins chargées, ce qui est crucial pour la planification opérationnelle quotidienne.
- 3) **Agrégation horaire** : Pour comprendre comment le nombre de passagers varie au cours de la journée, les données sont regroupées par heure. Cela implique de définir la colonne « *Datetime* » comme index et d'agréger les données par heure. En résumant le nombre de passagers pour chaque heure, un profil horaire détaillé de l'activité des passagers est obtenu.

Cette agrégation horaire est particulièrement utile pour identifier les heures de pointe. Par exemple, les heures de pointe du matin et du soir sont clairement visibles, guidant les décisions sur la fréquence du service pendant ces périodes.

- 1) **Segmentation par lignes de bus** : différentes lignes de bus peuvent présenter différents modes d'utilisation. Pour analyser cela, l'ensemble de données est segmenté par lignes de bus (« Ligne »). Une fonction classe les lignes de bus en différents groupes en fonction de leurs valeurs numériques ou de codes spécifiques. Cette segmentation aide à comprendre les différences de nombre de passagers sur les différents itinéraires.
- 2) **Agrégation par lignes de bus** : les données sont regroupées par lignes de bus et par intervalle de 60 minutes pour résumer le nombre de passagers. Cette analyse détaillée montre comment les différentes lignes de bus sont utilisées tout au long de la journée. Il met en évidence quels itinéraires sont les plus populaires et lesquels sont moins utilisés, éclairant ainsi les décisions sur la planification des itinéraires.

Pour se concentrer sur les périodes les plus pertinentes, les données sont filtrées pour exclure les heures hors service. Ainsi, l'analyse prend en compte uniquement les données de 5:00 heures à 23:00 heures, ce qui correspond aux heures de service typiques.

- **Analyse annuelle et mensuelle** : en visualisant les validations pour toute l'année 2020, segmentées par différentes lignes de bus, un aperçu de la tendance annuelle peut être observé. Cela aide à comprendre comment le nombre de passagers évolue tout au long de l'année, qui peut être influencé par des facteurs tels que les trimestres scolaires, les vacances et les changements climatiques saisonniers.

Enfin une concentration sur des mois spécifiques (automne 2020) permet une analyse plus détaillée de périodes plus courtes. Cela a mis en évidence les variations saisonnières et fournit un aperçu de la façon dont le nombre de passagers fluctue au cours des différents mois.

En intégrant les données GTFS aux données de validation, les modèles d'utilisation des différents arrêts sont analysés, identifiant les arrêts les plus et les moins utilisés et éclairant les décisions sur les changements potentiels concernant l'emplacement des arrêts ou la fréquence du service. Un autre aspect critique des données GTFS concerne les horaires prévus pour les itinéraires et les déplacements. En comparant les délais de validation réels avec les horaires prévus, le respect de la

planification peut être évalué. Cela aide à identifier les retards et à comprendre leur impact sur la satisfaction des passagers et l'efficacité opérationnelle.

Plus loin, ce travail a procédé à un examen approfondi des données GTFS (« *General Transit Feed Specification* ») afin de déterminer les jours de service et de non-service pour chaque itinéraire de transport en commun. Initialement, nous avons chargé les tables GTFS pertinentes — « *trips* », « *calendar* » et « *calendar_dates* » — depuis la base de données « *SQLite* » dans notre environnement de travail. En itérant sur ces plages de dates, une liste complète des jours de service potentiels pour chaque *ID* de service est créée, en veillant à inclure le jour de la semaine correspondant et les indicateurs de jour de service (du lundi au dimanche).

Sachant que les indicateurs de jours de service étaient initialement des caractères, ces colonnes sont converties en nombres entiers pour garantir un filtrage précis. Par la suite, un filtrage des jours de service générés est réalisé pour conserver uniquement les dates correspondant aux jours de service actifs de la semaine. Pour tenir compte des exceptions spécifiées dans la table « *calendar_dates* », des dates de service supplémentaires sont incorporées et les dates spécifiques de non-service sont supprimées telles que définies dans la table « *calendar_dates* »

Après avoir appliqué ces exceptions, nous avons associé les jours de service filtrés à leurs itinéraires respectifs en fusionnant ces données avec le tableau « trajets ». Cette étape nous a permis de relier chaque journée de service à son itinéraire correspondant. Enfin, les données par itinéraire et ID de service sont regroupées pour dresser une liste de tous les jours de service actif pour chaque itinéraire. De plus, les jours sans service sont identifiés en déterminant les dates non incluses dans les jours de service, en tenant compte de la plage de dates globale.

Le résultat de cette analyse complète est un ensemble de données détaillé qui comprend les identifiants uniques pour chaque itinéraire et service, des listes de jours de service actif et des listes de jours sans service. Cet ensemble de données fournit des informations précieuses sur les horaires des services de transport en commun, permettant une meilleure planification et une meilleure optimisation des opérations. Le résultat final est enregistré et utilisé pour une analyse plus approfondie, garantissant ainsi une compréhension claire du service sur l'ensemble du réseau d'Orléans.

3.6 Segmentation

L'analyse du nombre total de validations par heure sur l'ensemble du réseau offre un aperçu de haut niveau de la demande de transport en commun à Orléans, constituant une étape fondamentale dans le cadre de ce travail : partir d'une agrégation du total des validations par heure sur tous les modes de transport (bus et tramways) permet de :

1. Comprendre les modèles quotidiens, facilitant ainsi la planification à long terme et les ajustements saisonniers du niveau de service et de l'offre de transport.
2. Obtenir une base de référence pour comparaison : ces données agrégées fournissent une base de référence pour comparer les performances de différents segments du réseau (tels que différentes lignes de bus ou lignes de tramway) par rapport à l'ensemble du système.

Cette analyse initiale qui regroupe toutes les données opérationnelles sur le réseau permet d'étudier et de surveiller l'utilisation totale par jour et par heure. Cette perspective globale est essentielle pour identifier les tendances globales en matière d'achalandage et de demande de services, fournissant ainsi une base pour des analyses plus nuancées; elle permet non seulement des ajustements opérationnels immédiats, mais ouvre également la voie à une analyse plus détaillée, spécifique au mode, qui peut conduire à des améliorations ciblées des services de bus et de tramway. La segmentation ultérieure sera réalisée en utilisant les vecteurs identifiés plus haut, à savoir :

1. **Réseau** : agrégation horaire quotidienne et demi-horaire pour comprendre les modèles quotidiens et obtenir une base de référence.
2. **Mode** : analyse spécifique pour les bus et trams, en combinant les données pour chaque mode afin de comprendre les cycles de demande uniques.
3. **Ligne** : analyse horaire par ligne pour comprendre la variation de la demande pour chaque ligne.
4. **Itinéraire** : analyse par ligne et sens pour une compréhension détaillée des itinéraires.
5. **Arrêt** : analyse par arrêt pour évaluer le flux de passagers à des endroits spécifiques.

En disséquant davantage les données, des analyses distinctes pour les bus et les tramways sont effectuées, en ventilant les données par ligne : identifier les caractéristiques et les besoins spécifiques des différents modes et des lignes individuelles, ce qui est crucial pour adapter efficacement les améliorations des services et l'allocation des ressources.

Tout au long de ces analyses, des techniques avancées de transformation des données et des méthodes de segmentation sont utilisées. La segmentation aide à identifier des regroupements naturels au sein des données, utiles pour segmenter les horaires, les stations ou des itinéraires entiers en fonction de modèles d'utilisation similaires. Ces groupes peuvent guider des interventions ciblées, depuis l'ajustement de la fréquence des services jusqu'aux expansions stratégiques. La transformation des données, quant à elle, prépare l'ensemble de données à l'analyse, en normalisant les validations pour garantir que les informations obtenues soient basées sur des mesures comparables.

3.6.1 Méthodes de transformation des données

Dans ce même ordre d'idées, réaliser cette segmentation à partir des données horaires de validation des passagers, il est crucial de sélectionner la bonne technique de transformation des données. Ces données contiennent généralement des variations d'échelle et de répartition dues aux fluctuations du volume de passagers tout au long de la journée et selon les différents jours de la semaine. Il est donc essentiel de conserver une perspective quotidienne afin de capturer les variations intra-journalières et de comprendre les éléments qui peuvent influencer les décisions opérationnelles quotidiennes. Chaque méthode de transformation (Figure 3-7) appliquée aux données ajuste la façon dont elles sont interprétées et peut révéler des informations distinctes qui sont cruciales pour une segmentation efficace. Ce travail explore diverses méthodes de transformation de données, chacune adaptée à différents aspects de la structure et des propriétés statistiques des données, affectant le résultat de la segmentation.

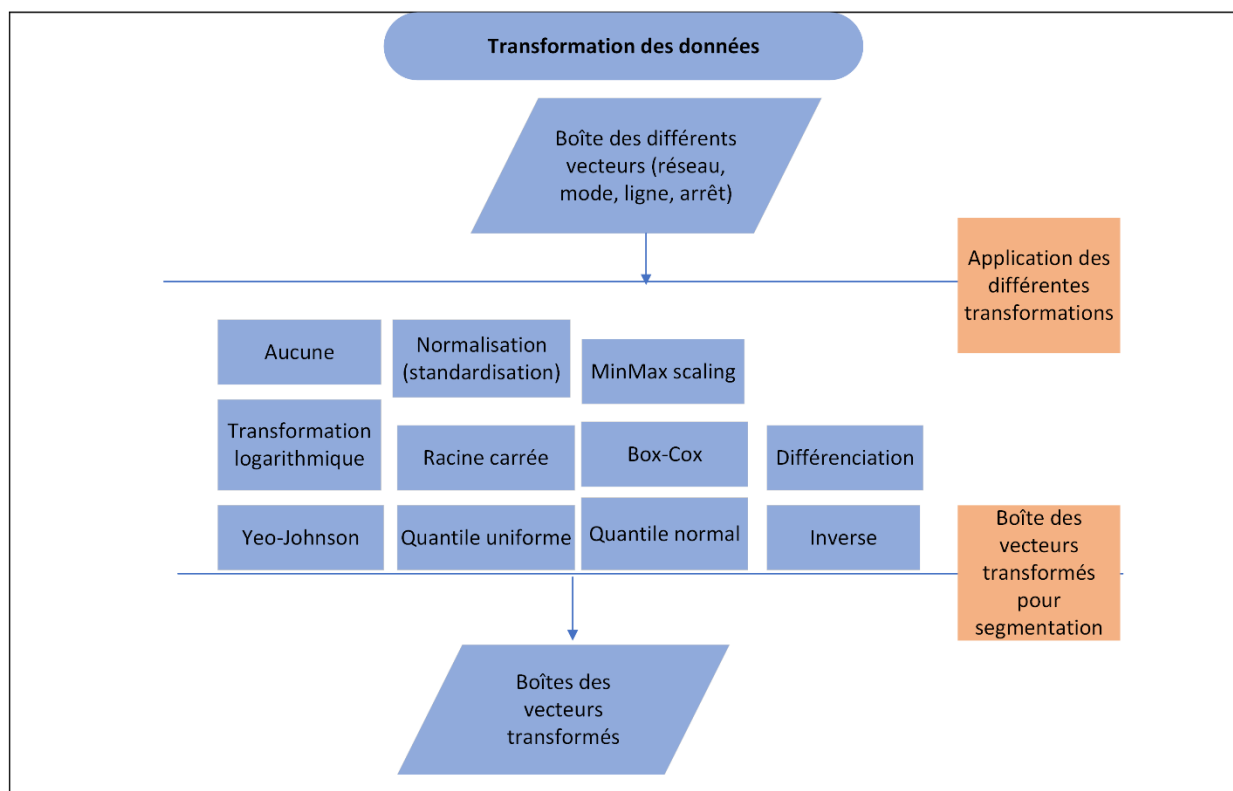


Figure 3-7 - Transformations appliquées aux vecteurs

Soit x et x' les valeurs originales et transformées; les différentes transformations expérimentées dans les analyses subséquentes sont :

1. **La normalisation** : elle ajuste le nombre de validations de chaque heure en supprimant la moyenne et en mettant à l'échelle la variance unitaire sur l'ensemble de l'ensemble de données. Cette méthode est particulièrement utile lorsque les plages horaires de données varient considérablement, garantissant que des volumes d'achalandage plus élevés pendant les heures de pointe ne dominent pas le processus de segmentation. En soustrayant la moyenne quotidienne et en divisant par l'écart-type pour les données de chaque jour, la normalisation garantit que les variations au cours de la journée, telles que les heures de pointe et les heures creuses, sont accentuées sans l'influence du nombre absolu de validations.

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

Où μ est la moyenne des données, et σ est l'écart-type des données. Cette transformation est bénéfique pour les méthodes de segmentation comme K-moyennes, qui supposent que les données sont centrées et mises à l'échelle.

2. Le « *min-max scaling* » : il met à l'échelle les données de chaque jour de 0 à 1. Il est particulièrement utile pour regrouper les jours en fonction de l'intensité relative du flux de passagers au fil des heures, facilitant ainsi les décisions sur l'allocation des ressources et les ajustements d'horaires en fonction des fluctuations quotidiennes de la demande.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

Où $\min(x)$ et $\max(x)$ sont respectivement les valeurs minimale et maximale des données. Il peut aussi améliorer la performance des algorithmes de segmentation comme K-moyennes et les méthodes hiérarchiques.

3. La transformation logarithmique : elle corrige les données asymétriques à droite, en compressant de manière plus significative les heures à volume plus élevé. Elle fait ressortir des variations plus subtiles pendant les heures les plus calmes, contribuant ainsi à différencier les jours, en particulier ceux impactés par des événements qui modifient les habitudes de voyage typiques.

$$x' = \log(1 + x) \quad (3)$$

Où $\log$ est le logarithme népérien. Elle est utile pour les méthodes comme K-moyennes et hiérarchique.

4. La racine carrée : convient aux données contenant des zéros. Elle réduit l'influence des valeurs extrêmes tout en préservant davantage la structure des données d'origine, ce qui la rend utile pour analyser les jours où des volumes d'achalandage plus faibles nécessitent un examen plus approfondi, comme lors d'événements spéciaux ou de vacances.

$$x' = \sqrt{x} \quad (4)$$

Cette transformation est utile pour les méthodes de segmentation comme K-moyennes et les méthodes hiérarchiques.

5. Box-Cox : elle optimise la conformité des données à la normalité, améliorant ainsi les performances de segmentation là où les hypothèses de distribution normale s'appliquent. Cela permet de standardiser les variations sur plusieurs jours, garantissant que la segmentation capture les véritables différences plutôt que les artefacts de distribution des données.

$$x'(\lambda) = \frac{(x^\lambda - 1)}{\lambda}, \quad \text{si } \lambda \neq 0 \quad (5)$$

$$\log(x), \quad \text{si } \lambda = 0$$

Où λ est le paramètre de transformation. Cette transformation permet de normaliser les variations horaires et d'améliorer la comparabilité des jours.

6. L'inverse : met en évidence les jours avec des nombres de validations généralement inférieurs, identifiant potentiellement des anomalies telles que celles qui se produisent lors d'interruptions de service ou de jours inhabituellement peu achalandés.

$$x' = \frac{1}{x} \quad (6)$$

Elle peut influencer les méthodes de segmentation comme K-moyennes et les méthodes hiérarchiques en changeant l'échelle des données.

7. La transformation quantile en uniforme : cette transformation garantit que les données pour chaque heure sont réparties uniformément, lissant la distribution et minimisant les effets des valeurs aberrantes. Elle est particulièrement utile pour rendre plus apparentes les tendances intra-journalières subtiles, prenant ainsi en charge une analyse affinée pour la planification quotidienne.

$$x' = F_{uniform}^{-1}(F(x)) \quad (7)$$

Où $F(x)$ est la fonction de répartition empirique des données et $F_{uniform}^{-1}$ est la fonction quantile de la distribution uniforme.

8. La transformation quantile en normal : transformation des données pour suivre une distribution normale correspond aux attentes de nombreux algorithmes de segmentation et est utile pour établir des comparaisons entre les jours dans l'hypothèse d'une normalité statistique, par exemple en comparant les jours de la semaine aux fins de semaine.

$$x' = F_{normal}^{-1}(F(x)) \quad (8)$$

Où F_{normal}^{-1} est la fonction quantile de la distribution normale.

9. Yeo-Johnson : est idéale pour les ensembles de données qui incluent des valeurs nulles ou négatives, stabilisant ainsi la variance et normalisant efficacement les distributions. Il prend en charge une analyse détaillée des jours à volume élevé et faible, s'adaptant à un large éventail de scénarios de validation quotidiens.

$$x'(\lambda) = \frac{[(x+1)^\lambda] - 1}{\lambda}, \quad \text{si } \lambda \neq 0, x \geq 0 \quad (9)$$

$$\log(x+1) \text{ si } \lambda = 0, \quad \text{si } x \geq 0$$

$$- \frac{[(-x+1)^{\{(2-\lambda)\}} - 1]}{(2-\lambda)}, \quad \text{si } x < 0$$

10. La différence : la différenciation des données d'une heure à l'autre au cours d'une même journée met l'accent sur les changements immédiats dans le nombre de validations, mettant en évidence les changements rapides dans le flux de passagers qui pourraient nécessiter des ajustements rapides des services ou du personnel.

$$x'_t = x_t - x_{t-1} \quad (10)$$

11. Aucun : l'absence de transformation permet d'évaluer les données brutes telles quelles, servant de référence pour évaluer l'impact d'autres transformations.

3.6.2 Méthodes de segmentation

La segmentation dans l'analyse des données de comptage de passagers est une méthode importante et largement utilisée. Elle permet d'analyser des flux distincts de passagers à différents moments de la journée et sur plusieurs jours. Par exemple, les heures de pointe en semaine, les périodes de loisirs la fin de semaine et les jours d'événements spéciaux peuvent chacun former des groupes différents, facilitant ainsi des ajustements de planification ciblés.

Diverses méthodes de segmentation, telles que K-moyennes, la segmentation hiérarchique et DBSCAN, permettent une segmentation précise. De plus, en décomposant l'analyse en segments plus faciles à gérer, tels que l'heure de la journée, le type de ligne de transport en commun ou des arrêts spécifiques, les planificateurs peuvent obtenir des informations plus approfondies sur les

modèles d'utilisation et la variabilité de la demande (Figure 3-8). Cette démarche permet l'application de méthodes et de mesures spécialisées adaptées aux caractéristiques uniques de chaque objet, améliorant ainsi la précision et l'efficacité des stratégies de planification d'horaire.

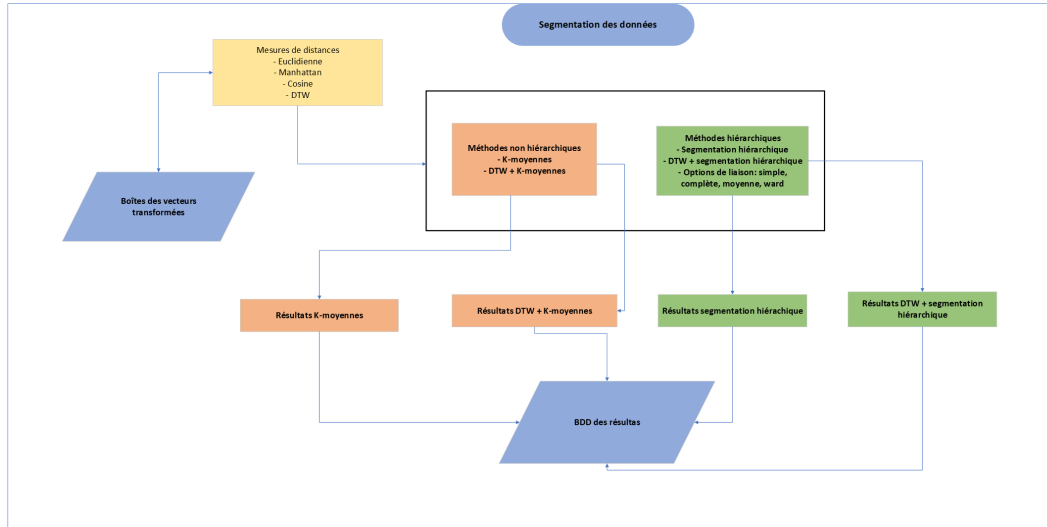


Figure 3-8 - Différentes segmentations appliquées aux données

1) L'algorithme des K-moyennes ou « K-means »

L'algorithme des K-moyennes vise à partitionner un ensemble de données en un nombre spécifié de groupes (**K**) en minimisant la somme des distances euclidiennes entre chaque point et son centroïde. Cela permet de former des groupes cohérents où les points sont proches de leur centre respectif.

Mathématiquement, il peut être explicité ainsi :

$$\operatorname{argmin}_{\{\mu_i\}_{i=1}^K} \sum_{i=1}^K \sum_{j=1}^{n_i} \|x_j^{(i)} - \mu_i\|^2, \quad \text{avec } \mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j^{(i)} \quad (11)$$

Où :

- K est le nombre de groupes
 - n_i est le nombre de points dans le groupe i
 - $x_j^{(i)}$ le j nième point du groupe i
 - μ_i le centroïde du groupe i
- 2) La segmentation hiérarchique construit un arbre (dendrogramme) de groupes en fusionnant ou en divisant les groupes de manière itérative en fonction du critère de liaison choisi. Il ne

nécessite pas de prédéfinir le nombre de groupes. Ce dernier commence par considérer chaque point de données comme un groupe individuel, puis fusionne les groupes les plus proches à chaque étape jusqu'à obtenir un nombre prédéfini de groupes. La distance entre les groupes est calculée à l'aide de diverses méthodes de liaison. Les distances peuvent être mesurées par différentes métriques, comme la distance euclidienne, de Manhattan, la similarité cosinus ou le DTW. L'algorithme fournit une vue complète de la structure des données, adaptée à la visualisation des hiérarchies de groupes. On peut la formuler de la manière suivante :

- a) Liaison simple : la distance entre deux groupes est la distance minimale entre une paire quelconque de points dans les deux groupes

$$d(A, B) = \min_{\{a \in A, b \in B\}} d(a, b) \quad (12)$$

- b) Liaison complète : la distance entre deux groupes est la distance maximale entre une paire quelconque de points dans les deux groupes

$$d(A, B) = \max_{\{a \in A, b \in B\}} d(a, b) \quad (13)$$

- c) Liaison moyenne : la distance entre deux groupes est la moyenne des distances entre toutes les paires possibles de points dans les deux groupes

$$d(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (14)$$

- d) Méthode de Ward : minimise la variance totale au sein des groupes.

$$d(A, B) = \frac{2|A||B|}{|A| + |B|} |\mu_A - \mu_B|^2 \quad (15)$$

μ_A et μ_B Centroides des groupes A et B

- 3) Le DTW combiné au K-moyennes : la combinaison de DTW avec la segmentation K-moyennes permet l'identification de groupes dans les données de séries chronologiques, en tenant compte des désalignements temporels. Chaque série temporelle est attribuée au centroïde le plus proche en fonction de la distance DTW et les centroïdes sont mis à jour en minimisant la somme des distances au sein d'un groupe.

- 4) Le DTW combiné à la segmentation hiérarchique : la matrice de distance DTW est utilisée comme entrée pour la segmentation hiérarchique. Les distances sont calculées selon les mêmes critères que précédemment, en remplaçant la distance classique par DTW.

3.6.3 Métriques de distance

Compte tenu de la nature des données (où chaque ligne représente un jour et les colonnes représentent des intervalles horaires), certaines mesures de distance peuvent fournir des résultats de regroupement plus robustes que d'autres. Plusieurs distances ont été testées :

- 1) Distance euclidienne : Pour les validations horaires de cartes à puce, la distance euclidienne permet d'identifier les jours présentant des modèles globaux similaires en termes de validations totales par heure, fournissant ainsi une méthode simple pour regrouper les jours avec des volumes de passagers similaires. La distance euclidienne est inestimable pour sa simplicité et sa capacité de mesure directe dans un espace multidimensionnel, où chaque dimension peut représenter une heure de la journée. Il calcule la distance en ligne droite entre les points, ce qui en fait un choix intuitif pour les données quantitatives où les différences absolues dans les nombres de validation sont significatives. Pour deux points $\mathbf{p}$ et $\mathbf{q}$ dans un espace euclidien :

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (16)$$

- Avec d distance euclidienne

- 2) La distance de Minkowski est une métrique généralisée qui inclut la distance euclidienne et de Manhattan comme étant des cas particuliers; la distance généralisée :

$$d(\mathbf{p}, \mathbf{q}) = \sqrt[p]{\sum_{i=1}^n (p_i - q_i)^p} \quad (17)$$

- Euclidienne quand $p=1$ et Manhattan quand $p=2$

- 3) Distance de Manhattan (bloc urbain) : étant donné les contextes de transport urbain où le flux de passagers peut ne pas augmenter de façon linéaire au fil des heures mais plutôt par sauts (car les gens ont tendance à se déplacer en fonction des heures de travail, des heures d'école, etc.). La métrique Manhattan peut mieux capturer ces changements discrets. Il est utile pour identifier les jours où le schéma de flux de passagers (par exemple, pointe du

matin, accalmie de midi, pointe du soir) est similaire, quel que soit le nombre exact de validations. La distance de Manhattan calcule la somme des différences absolues entre les coordonnées dans un espace vectoriel. Cela est particulièrement efficace lorsque les dimensions des données représentent des caractéristiques indépendantes, telles que différentes heures de la journée.

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n |\mathbf{p}_i - \mathbf{q}_i| \quad (18)$$

- 4) Cosinus : la similarité cosinusoidale est excellente pour regrouper les jours en fonction de la demande tout au long de la journée, quelles que soient les validations quotidiennes totales. Elle peut efficacement regrouper les jours où la demande est similaire, comme des pics similaires aux heures de pointe, même si le nombre total de passagers varie considérablement. Bien qu'elle soit généralement utilisée dans l'analyse de texte, la similarité cosinus mesure le cosinus de l'angle entre deux vecteurs, ce qui la rend pertinente pour comparer l'orientation plutôt que l'ampleur des vecteurs de données. Cette propriété est cruciale lorsque l'intérêt réside dans la structure ou la forme des données selon les dimensions (heures), plutôt que dans le volume absolu.

$$S(\mathbf{p}, \mathbf{q}) = \frac{\sum_{i=1}^n p_i q_i}{\sqrt{\sum_{i=1}^n p_i^2} \sqrt{\sum_{i=1}^n q_i^2}} \quad (19)$$

- S similarité cosinus d $(\mathbf{p}, \mathbf{q}) = 1 - S(\mathbf{p}, \mathbf{q})$

- 5) Le « *Dynamic Time Warping* » (DTW) est une métrique spécialisée pour l'analyse des séries chronologiques, permettant de mesurer la similarité entre deux séquences temporelles, même lorsque leur vitesse ou leur rythme diffèrent. Cette méthode aligne les séquences de manière à minimiser la distance entre des points similaires, même si ces points se produisent à des moments différents. Le DTW est particulièrement utile pour analyser des ensembles de données où des décalages temporels dans les modèles sont attendus, par exemple lorsque les heures de pointe varient selon les jours.

Cette technique permet de regrouper les jours en fonction de leurs modèles de validation tout en s'adaptant aux changements dans ces modèles, ce qui la rend particulièrement recommandée pour les données présentant une variabilité temporelle.

Pour deux séquences temporelles pp et qq , la distance DTW (d) est calculée à l'aide d'une matrice de coût D de dimensions $n*m$ où chaque élément $D(i, j)$ représente le coût pour aligner les points p_i et q_j :

- a. Calcul de la matrice de coût DTW

$$D(i, j) = |p_i - q_j| + \min \{ D(i-1, j), D(i, j-1), D(i-1, j-1) \} \quad (20)$$

- b. Calcul de la distance DTW

$$DTW(p, q) = D(n, m) \quad (21)$$

La distance DTW est la valeur à $D(n, m)$, qui représente le coût total accumulé pour le chemin d'alignement optimal entre les deux séquences temporelles.

3.6.4 Validation des résultats de la segmentation

La validation des résultats de la segmentation est une étape cruciale pour évaluer la qualité des partitions obtenues. Cette section présente différentes métriques de validation interne, telles que la Somme des Carrés des Erreurs (SSE), le Coefficient de Silhouette et la Statistique « K-Gap ». Ces mesures présentées sur la Figure 3-9 permettent de quantifier la compacité des groupes, la séparation entre les groupes et de déterminer le nombre optimal de groupes. Leur utilisation conjointe assure une évaluation complète et robuste de la performance des algorithmes de segmentation appliqués aux données.

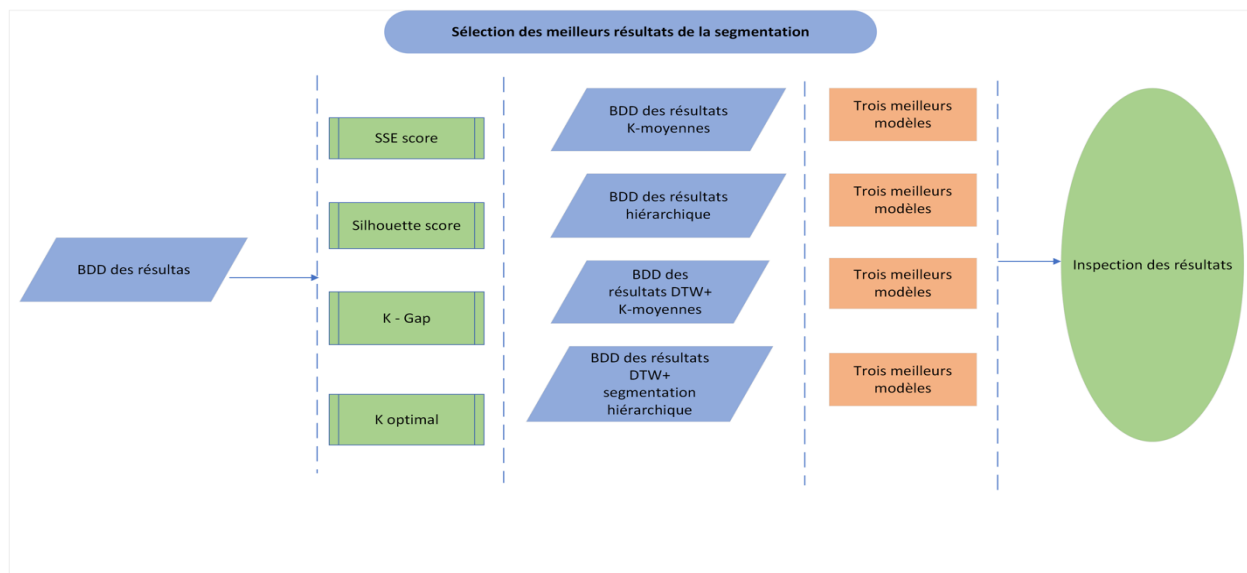


Figure 3-9 - Critères d'évaluation des segmentations

- 1) SSE : la Somme des Carrés des Erreurs (SSE) est une métrique utilisée pour évaluer la performance d'un algorithme de segmentation en mesurant la compacité des groupes formés. Elle est définie comme la somme des distances au carré entre chaque point de données et le centroïde du groupe auquel il appartient.

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} |x - \mu_i|^2 \quad (22)$$

Où :

- k est le nombre de groupes
 - C_i est l'ensemble des points de données dans le groupe i
 - x est un point de données dans le groupe i
 - μ_i est le centroïde du groupe i
 - $\|x - \mu_i\|^2$ est la distance au carré entre le point de données x et le centroïde μ_i .
- 2) Coefficient de Silhouette : le Coefficient de Silhouette est une métrique qui mesure à quel point un point de données est similaire à son propre groupe (cohésion) par rapport aux autres groupes (séparation). Il varie de -1 à 1, où une valeur élevée indique que le point de données est bien associé à son propre groupe et mal associé aux groupes voisins.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (23)$$

Où :

- $s(i)$ est le coefficient de silhouette pour le point de données i
- $a(i)$ est la distance moyenne entre le point de données i et tous les autres points du même groupe
- $b(i)$ est la distance moyenne minimale du point de données i à tous les points du groupe le plus proche qui n'est pas celui auquel i appartient.

Le coefficient de silhouette global est la moyenne des coefficients de silhouette pour tous les points de données.

- 3) Statistique « *K-Gap* » ou « *Gap Statistic* » : compare la variation totale intra-groupe pour différents nombres de groupes avec leurs valeurs attendues sous une distribution de référence nulle des données. Elle aide à déterminer le nombre optimal de groupes en trouvant la valeur de k pour laquelle la structure de segmentation observée est la plus différente de la distribution de référence nulle :

$$Gap(k) = \frac{1}{B} \sum_{b=1}^B \log(W_k^{*b}) - \log(W_k) \text{ et} \\ W_k = \sum_{i=1}^k \frac{1}{2|C_i|} \sum_{x,y \in C_i} \|x - y\|^2 \quad (24)$$

Où :

- W_k est la dispersion intra-groupe pour les données observées pour k groupes
- W_k^{*b} est la dispersion intra-groupe pour les données de référence (générées sous l'hypothèse nulle) pour k groupes
- B est le nombre de jeux de données de référence
- C_i est le nombre de points dans le groupe i .

Le nombre optimal de groupes est la valeur de k qui maximise la statistique de l'écart.

3.6.5 Choix de la segmentation optimale

Le choix du groupement optimal est essentiel pour garantir une segmentation des données efficace et significative. Cette section explore plusieurs indices de validation (Figure 3-10), notamment la Somme des Carrés Intra-Groupe (WCSS), l'Indice de Davies-Bouldin, l'Indice de Calinski-Harabasz, l'Indice de Dunn, la Cohésion et Séparation (intra-groupe et inter-groupe ou BCSS), et le Coefficient de Cophénétique. Ces indices permettent d'identifier la meilleure configuration de segmentation en évaluant la compacité interne des groupes et la séparation entre eux.

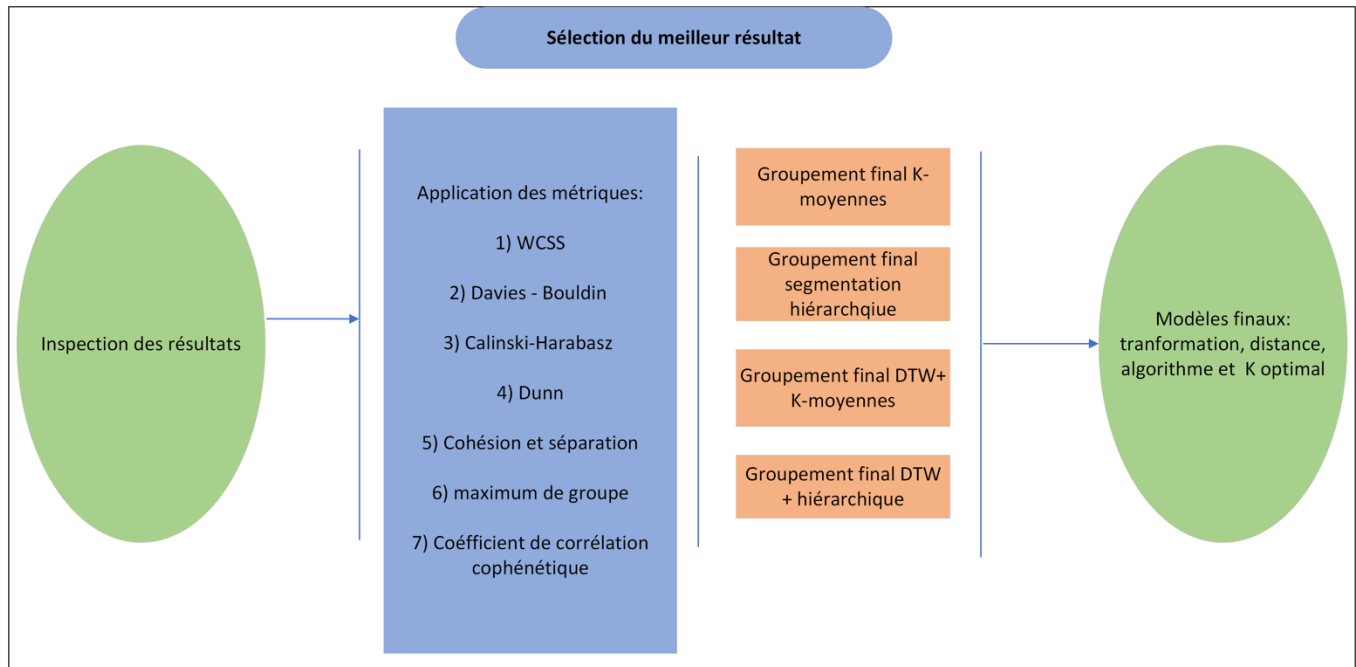


Figure 3-10- Sélection du groupement optimal

1. WCSS : La Somme des Carrés Intra-Groupe (WCSS) est une métrique utilisée pour mesurer la compacité des groupes formés par un algorithme de segmentation. Elle évalue la somme des carrés des distances entre chaque point de données et le centroïde de son groupe.

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (25)$$

Où :

- k est le nombre de groupes
 - C_i est l'ensemble des points données dans le groupe i
 - μ_i est le centroïde du groupe i
 - $\|x - \mu_i\|$ est la distance euclidienne entre le point de données x et le centroïde μ_i .
2. Indice de Davies-Bouldin : L'Indice de Davies-Bouldin évalue la qualité d'une segmentation en mesurant la moyenne des rapports de similarité entre chaque groupe et le groupe le plus similaire. Une valeur plus faible indique une meilleure segmentation.

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{\sigma_i + \sigma_j}{\|\mu_i - \mu_j\|}$$

Où :

- k est le nombre de groupes.
 - σ_i est la moyenne des distances entre chaque point du groupe i et son centroïde.
 - $\|\mu_i - \mu_j\|$ est la distance entre les centroïdes des groupes i et j .
3. Indice de Calinski-Harabasz : L'Indice de Calinski-Harabasz, ou indice de rapport de variance, évalue la qualité d'une segmentation en comparant la somme des carrés inter-groupe (BCSS) à la somme des carrés intra-groupe (WCSS).

$$CH = \frac{BCSS / (k - 1)}{WCSS / (n - k)} \quad (37)$$

Où :

- k est le nombre de groupes.
 - n est le nombre total de points de données.
 - BCSS est la somme des carrés inter-groupe.
 - WCSS est la somme des carrés intra-groupe.
4. Dunn Index : L'Indice de Dunn vise à identifier des groupes compacts et bien séparés en minimisant la distance maximale intra-groupe et en maximisant la distance minimale inter-groupe.

$$Dunn = \frac{\min_{1 \leq i \leq j \leq k} \delta(C_i, C_j)}{\max_{1 \leq k \leq k} \delta(C_k)} \quad (28)$$

Où :

- $\delta(C_i, C_j)$ est la distance minimale entre les points des groupes C_i, C_j
- $\delta(C_k)$ est la distance maximale intra-groupe du groupe k

Un indice de Dunn plus élevé indique des groupes bien séparés et compacts.

5. Cohésion et séparation (intra-groupe et inter-groupe ou BCSS) : la cohésion mesure la compacité d'un groupe en évaluant la somme des distances entre les points de données et le

centroïde de leur groupe (intra-groupe). La séparation mesure la distance entre les centroïdes des groupes (inter-groupe).

$$Cohésion = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

$$BCSS = \sum_{i=1}^k |C_i| \|\mu_i - \mu\|^2 \quad (29)$$

Où :

- μ est le centroïde de toutes les données
 - $|C_i|$ est le nombre de points dans le groupe i .
6. Coefficient de cophénétique : le coefficient de cophénétique mesure à quel point un dendrogramme reflète fidèlement les distances entre les points de données d'origine. Il compare les distances cophénétiques (distances sur le dendrogramme) aux distances originales.

$$Coefficient\ de\ cophénétique = \frac{\sum_{i < j} (d_{ij} - \bar{d})(t_{ij} - \bar{t})}{\sqrt{\sum_{i < j} (d_{ij} - \bar{d})^2 \sum_{i < j} (t_{ij} - \bar{t})^2}} \quad (30)$$

Où:

- d_{ij} est la distance originale entre les points i et j .
- t_{ij} est la distance cophénétique entre les points i et j .
- $\bar{d}$ et $\bar{t}$ sont les moyennes respectives des distances originales et cophénétiques.

Un coefficient de cophénétique plus élevé indique une meilleure fidélité du dendrogramme aux distances originales.

3.6.6 Analyse détaillée des groupes

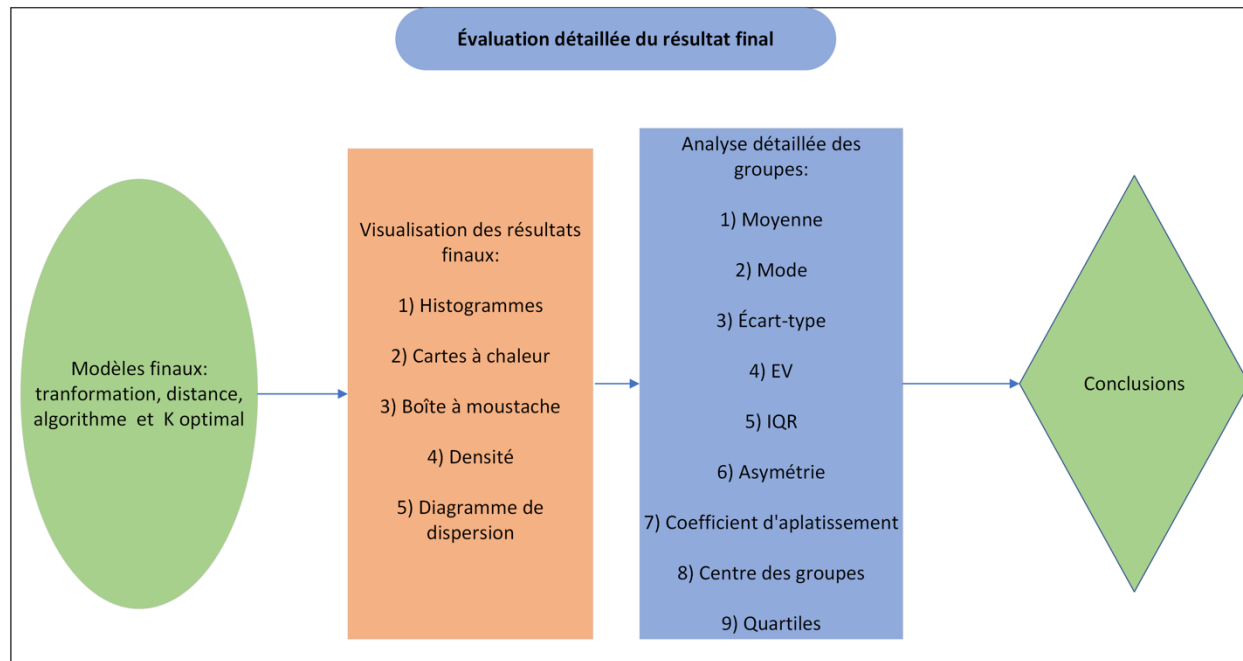


Figure 3-11 - Évaluation détaillée du groupe final

Plusieurs critères statistiques sont essentiels pour analyser et interpréter la distribution des données, la compacité et la séparation des groupes, et pour prendre des décisions éclairées en matière de segmentation. Ils fournissent des mesures quantitatives pour évaluer la qualité des groupes et les caractéristiques des données (Figure 3-11). Mises à part les mesures de tendance centrale et la dispersion, d'autres sont introduites dans le cadre de ce travail :

1. Étendue de variation : est la différence entre la valeur maximale et la valeur minimale d'un ensemble de données.
2. IQR (Intervalle Interquartile) : est la différence entre le troisième quartile (75e percentile) et le premier quartile (25e percentile) d'un ensemble de données. Il mesure la dispersion des valeurs centrales.
3. Asymétrie : mesure la symétrie d'une distribution de données. Une asymétrie positive indique une queue longue à droite, tandis qu'une asymétrie négative indique une queue longue à gauche.

$$\text{Asymétrie} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^3 \quad (31)$$

Où :

- μ est la moyenne de l'ensemble de données.
 - σ est l'écart-type de l'ensemble de données.
 - n est le nombre total de points de données
4. L'aplatissement : l'aplatissement (ou kurtosis) mesure la "pointedness" ou la platitude de la distribution d'un ensemble de données par rapport à une distribution normale.

$$Kurtosis = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^4 - 3 \quad (32)$$

Où :

- μ est la moyenne de l'ensemble de données.
 - σ est l'écart-type de l'ensemble de données.
 - n est le nombre total de points de données
5. 25e Percentile : ou premier quartile (Q1), est la valeur sous laquelle se trouvent 25% des données.
6. 50e Percentile : ou médiane (Q2), est la valeur sous laquelle se trouvent 50% des données.
7. 75e Percentile : ou troisième quartile (Q3), est la valeur sous laquelle se trouvent 75% des données.
8. Nombre d'Observations : est le nombre total de points de données dans l'ensemble de données.
9. Centre du groupe : est le point moyen de tous les points de données dans un groupe. Pour des données multi-dimensionnelles, il est le vecteur moyen des coordonnées de tous les points de données du groupe. Il est donné par :

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (33)$$

Où :

- μ_i est le centroïde du groupe i
- $|C_i|$ est le nombre de points dans le groupe i
- x est un point de données dans le groupe i .

3.7 Application

Diverses techniques de segmentation et d'analyse sont appliquées à différents objets du réseau TAO, y compris le réseau, les lignes, les arrêts et les modes. Au niveau du réseau, une analyse globale de la demande de transport est effectuée en agréant les données de validation sur une base horaire et demi-horaire pour capturer les modèles quotidiens et identifier les tendances globales. Pour l'analyse des lignes, des vecteurs de comptage horaire des passagers sont créés pour chaque ligne afin d'identifier les heures de pointe et les modèles de demande quotidiens, tandis que la segmentation des lignes permet d'acquérir une meilleure connaissance des périodes de forte affluence et de formuler des recommandations sur les horaires. L'analyse des arrêts vise à évaluer la demande aux arrêts spécifiques en analysant les comptages horaires des passagers pour déterminer les zones à fort achalandage et les points de congestion potentiels, guidant ainsi les décisions sur l'ajustement des fréquences et l'amélioration des infrastructures. Au niveau des modes, une analyse distincte des bus et des tramways est réalisée pour comprendre les cycles de demande uniques de chaque mode et équilibrer les ressources en conséquence. Plusieurs techniques avancées de transformation des données, telles que la normalisation et la transformation logarithmique, ainsi que des techniques de segmentation comme K-moyennes et la segmentation hiérarchique, sont utilisées pour segmenter efficacement les données. Les résultats de la segmentation sont validés à l'aide de métriques telles que le SSE, le coefficient de silhouette et la statistique de l'écart pour déterminer le nombre optimal de groupes. En appliquant ces techniques à différents niveaux, des informations exploitables sont dérivées pour informer les décisions stratégiques dans la planification des horaires et la gestion des ressources, garantissant ainsi un système plus efficace et convivial, comme détaillé au chapitre 4.

CHAPITRE 4 PRÉSENTATION DES RÉSULTATS

4.1 Analyse descriptive des données

La Figure 4-1 montre les validations mensuelles totales sur le réseau TAO en 2020, différenciant les passagers des bus et des tramways. De janvier à mars, le nombre de passagers diminue régulièrement, atteignant un creux en avril en raison des confinements liés à la COVID-19. À partir de mai, une reprise progressive est observée, culminant en août, surtout pour les tramways. De septembre à décembre, une légère diminution est notée, plus marquée pour les tramways que pour les bus. Cette tendance illustre l'impact de la pandémie sur les déplacements et les différences d'utilisation entre les bus et les tramways sur l'année.

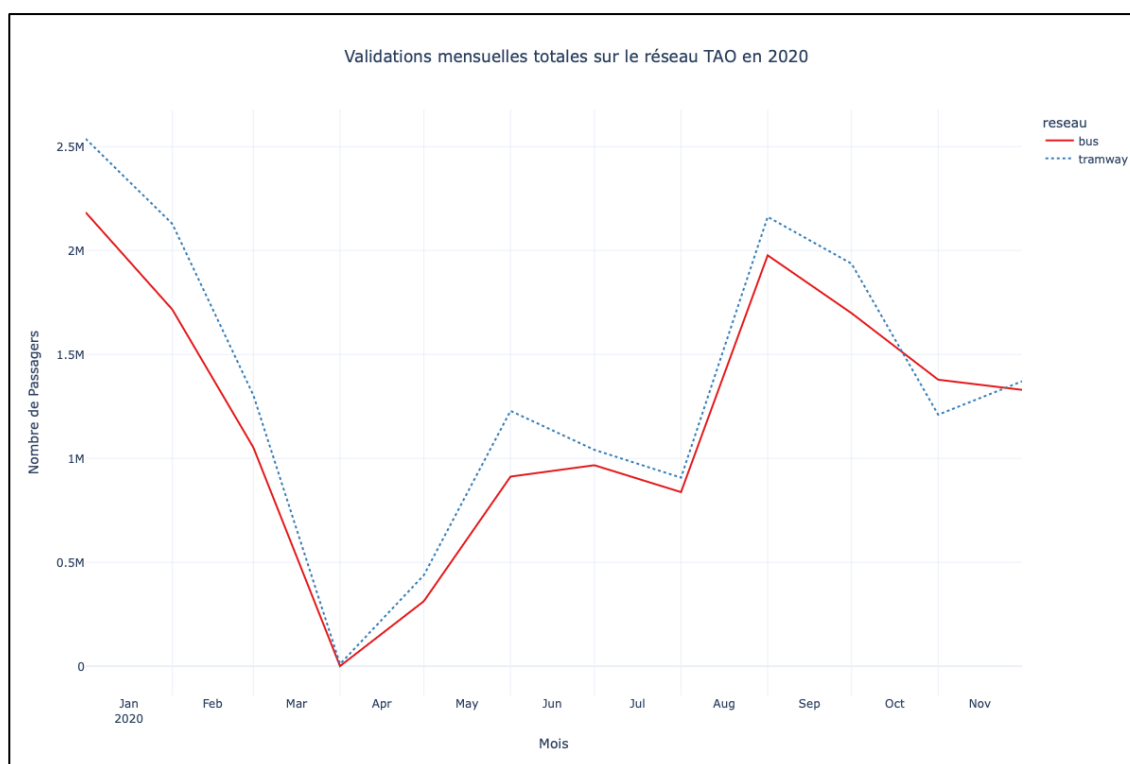


Figure 4-1 - Validations mensuelles totales sur le réseau TAO en 2020

Quant à la Figure 4-2, elle montre les validations mensuelles par type de ligne en utilisant une échelle logarithmique pour le nombre de passagers. On peut observer une baisse significative des validations pour toutes les lignes en avril, due aux restrictions de déplacement liées à la pandémie. Avant et après cette période, les lignes de tram A et B, ainsi que les lignes structurantes,

maintiennent un nombre élevé de validations, indiquant une utilisation régulière et importante. Les lignes périurbaines, de proximité et express montrent également une certaine constance dans leur utilisation, malgré une baisse notable en avril. Les lignes scolaires et événementielles, en revanche, affichent des fluctuations plus marquées, avec des validations quasi nulles pendant certaines périodes, probablement en raison de la fermeture des écoles et l'annulation d'événements. Ce graphique permet de comprendre les variations de la demande en fonction du type de ligne et de mieux planifier les services en fonction des besoins réels.

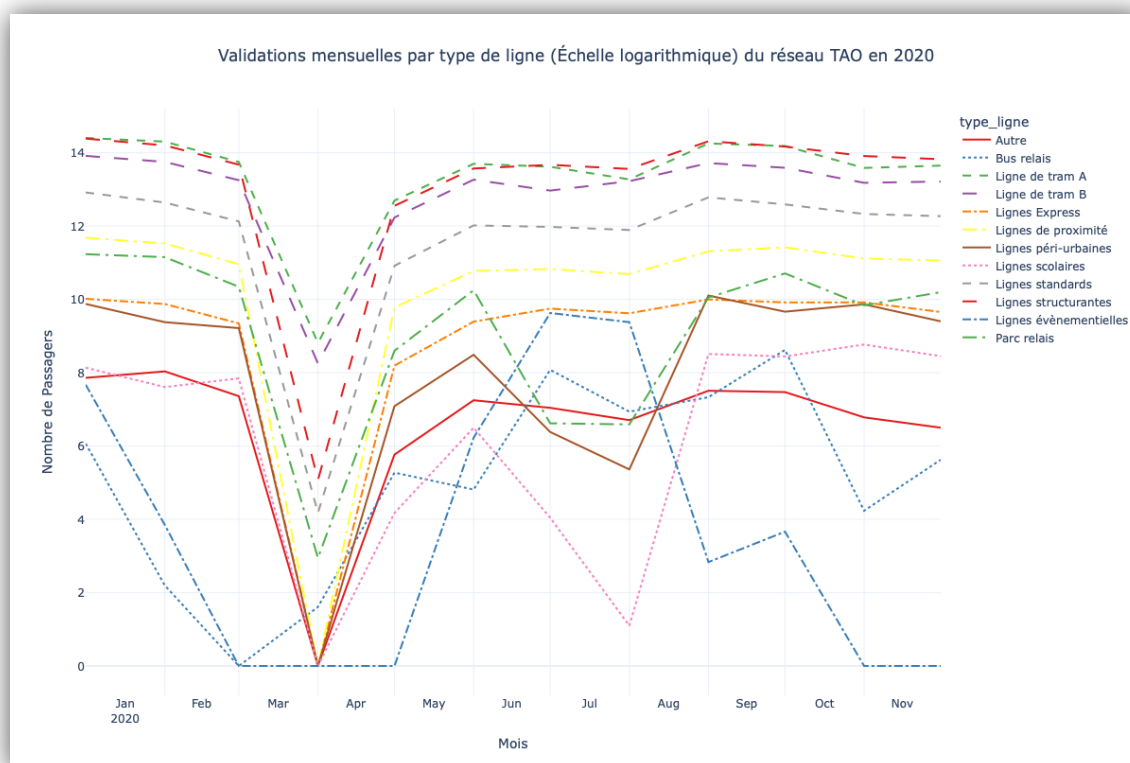


Figure 4-2 - Validations mensuelles par type de ligne du réseau TAO en 2020

La Figure 4-3 illustre les validations mensuelles pour les dix lignes les plus fréquentées du réseau, en utilisant une échelle logarithmique pour le nombre de passagers. Les autres lignes de bus, telles que les lignes 1, 2, 3, 4, 5, 6, 7 et 18, suivent une tendance similaire au tram avec une reprise progressive après le mois d'avril. Cette analyse démontre l'impact significatif des restrictions sanitaires sur l'utilisation des transports en commun et la reprise progressive de la fréquentation post-confinement.

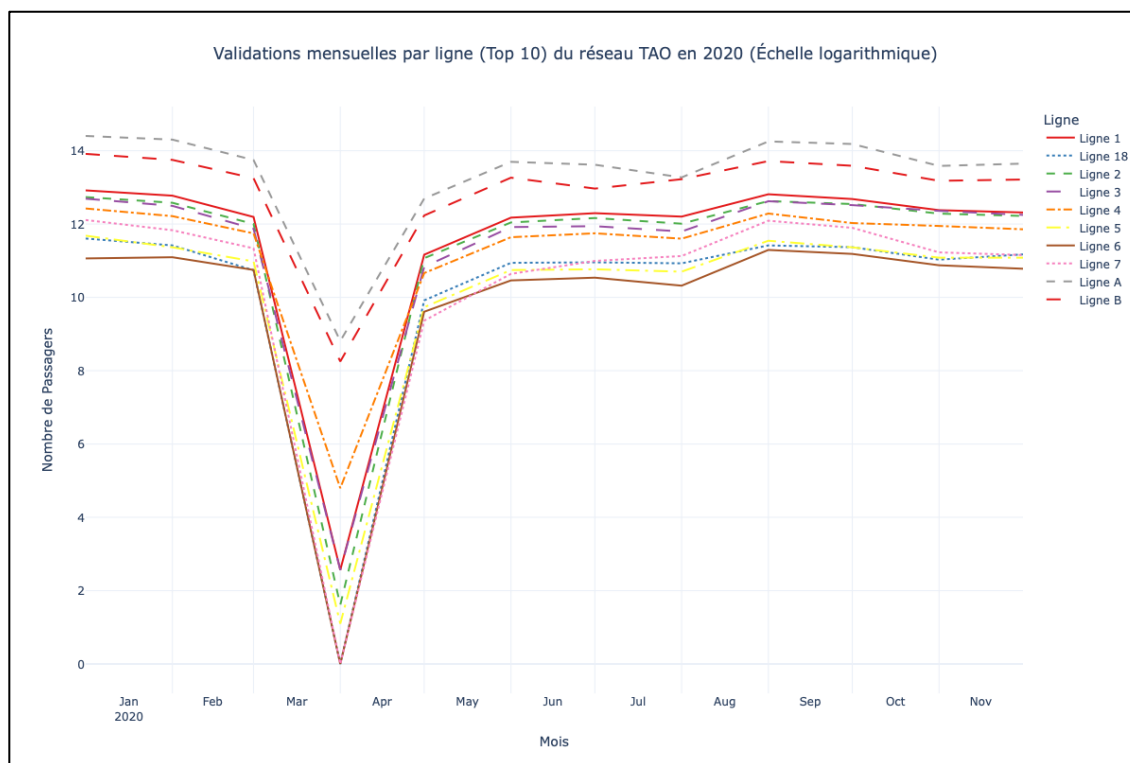


Figure 4-3 - Validations mensuelles des dix lignes les plus fréquentées du réseau TAO en 2020

La Figure 4-4 présente une visualisation des différentes transactions de validation séparant les types de lignes et les modes de transport (tramway et bus) ainsi que les différents types de transactions et correspondances.

En général, le compostage d'un billet signifie qu'un usager doit insérer son billet dans une machine pour le valider avant de pouvoir l'utiliser pour son trajet. Cela peut se faire avec des tickets en papier qui sont physiquement compostés (marqués) par une machine, ou avec des cartes sans contact (CSC), où la validation se fait électroniquement en passant la carte devant un lecteur. Dans le réseau TAO à Orléans, les usagers doivent souvent valider leur titre de transport (composter) avant d'entrer dans le bus ou le tramway. Ils représentent plus de 50% des validations. D'un autre côté, la correspondance magnétique se réfère à l'utilisation de cartes à bande magnétique pour valider et transférer entre différents modes de transport ou lignes sans acheter un nouveau billet. Les cartes magnétiques enregistrent les informations du trajet et permettent de transférer (faire une correspondance) d'un mode de transport à un autre.

Les validations de tramway sont divisées principalement entre les lignes A et B, avec un grand nombre de validations correspondant à des compostages CSC et magnétiques. Les lignes

structurantes, qui comprennent des validations à haut volume, montrent des transactions diverses telles que les compostages CSC, les correspondances CSC, et les validations et correspondances magnétiques. Les autres types de lignes, comme les lignes scolaires, périurbaines, de proximité et les bus relais, ont également des validations variées mais en volumes moindres comparés aux lignes structurantes et de tramway.

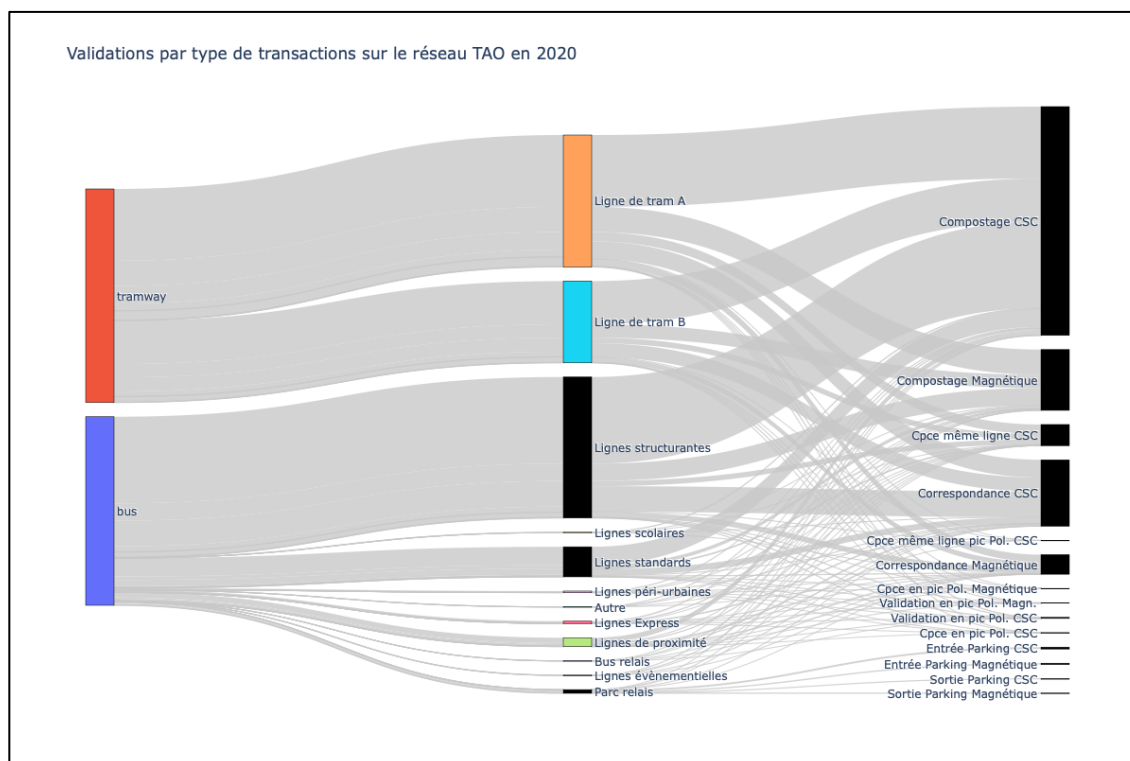


Figure 4-4 - Validations par type de transactions sur le réseau TAO en 2020

Les diagrammes de « Sankey » (Figure 4-5 et Figure 4-6) montrent respectivement les dix et vingt arrêts les plus fréquentés en 2020. Les deux lignes de tram A et B, sept (7) des lignes structurantes et standards constituent l'ensemble des validations sur le réseau avec une moyenne de 16 kms et plus de 35 arrêts. De plus, ils dominent en termes de fréquentation des arrêts. L'arrêt "De Gaulle" apparaît comme le plus fréquenté, suivi par d'autres arrêts majeurs tels que "Gare d'Orléans", "République" et "Royale-Chatelet". Ces arrêts servent de nœuds importants pour les correspondances et les flux de passagers, jouant un rôle crucial dans la fluidité et l'efficacité du réseau de transport en commun.

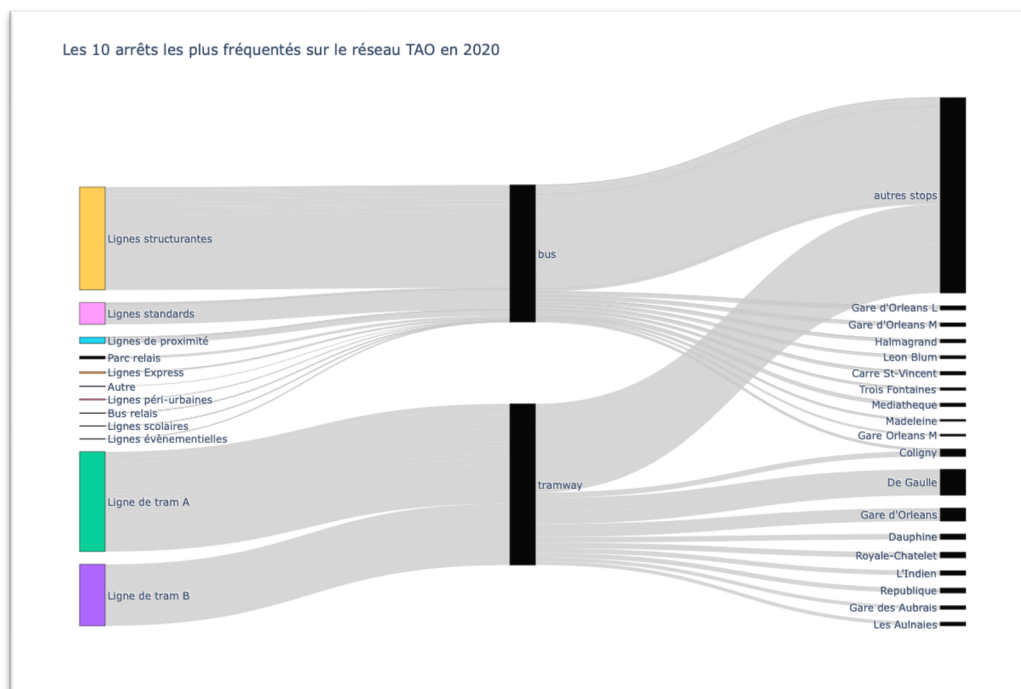


Figure 4-5 - Les dix (10) arrêts les plus fréquentés sur le réseau TAO en 2020

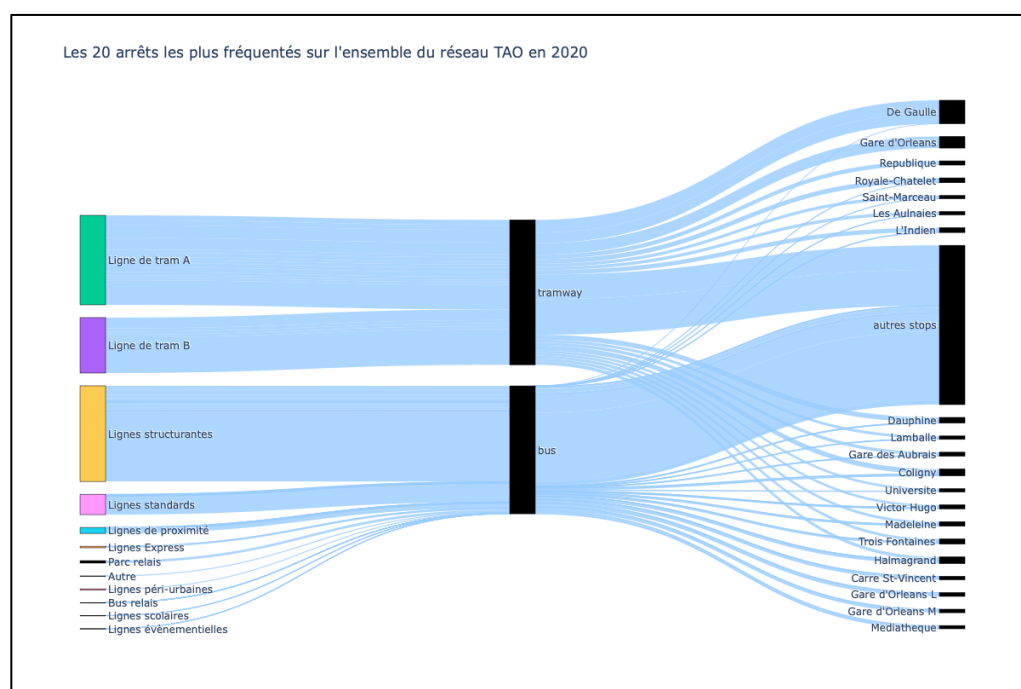


Figure 4-6 - Les vingt (20) arrêts les plus fréquentés sur l'ensemble du réseau TAO en 2020

La Figure 4-7 présente la répartition du pourcentage de validations pour chaque jour de la semaine par type de ligne. Il met en évidence les variations d'utilisation des différentes lignes tout au long de la semaine.

- Les bus relais, par exemple, montrent une utilisation plus élevée en semaine, avec des pics significatifs le lundi (23,7%) et le vendredi (27,4%). Cela peut indiquer une forte dépendance de ces lignes pour les trajets domicile-travail en semaine. En revanche, les lignes événementielles ont une utilisation marquée le vendredi (30,1%), ce qui est attendu car elles sont probablement utilisées pour des événements spéciaux ou des activités de fin de semaine.
- Les lignes structurantes et standards, ainsi que les lignes de tramway A et B, montrent une utilisation relativement constante tout au long de la semaine, avec de légères variations. Ces lignes principales maintiennent un usage stable, reflétant leur rôle crucial dans le transport quotidien des résidents.
- Les lignes scolaires présentent une utilisation plus élevée en semaine, avec des pourcentages plus élevés le lundi (20,1%) et le mardi (19,5%), et une nette diminution pendant la fin de semaine, ce qui est cohérent avec leur objectif de transporter les étudiants.
- Les lignes de proximité et les lignes périurbaines montrent également des variations similaires, avec des pics en semaine et une légère baisse la fin de semaine, suggérant une utilisation principalement pour les trajets domicile-travail et autres activités quotidiennes.

La Figure 4-8 des validations par jour, segment de la journée et saison met en évidence des variations significatives dans l'utilisation des transports en commun, en utilisant la segmentation temporelle suivante : Matinée (5h-10h), Fin de matinée (10h-12h), Midi (12h-14h), Après-midi (14h-18h), Soirée (18h-22h) et Nuit (22h-5h). Au printemps, les validations sont faibles, en raison des restrictions liées à la pandémie, avec une légère augmentation en début de soirée. En été, une légère hausse des validations est observée, particulièrement les lundis et vendredis, suggérant des déplacements liés aux activités estivales. L'automne montre une augmentation notable, surtout en début de soirée et en milieu de journée, reflétant la reprise des activités scolaires et professionnelles. L'hiver atteint le pic de validations, notamment en milieu de journée et en début de soirée, avec une forte activité les vendredis, indiquant des déplacements pour les achats et les préparatifs de la fin de semaine. Ces variations saisonnières et journalières fournissent des informations essentielles pour ajuster les horaires et les ressources du réseau TAO.

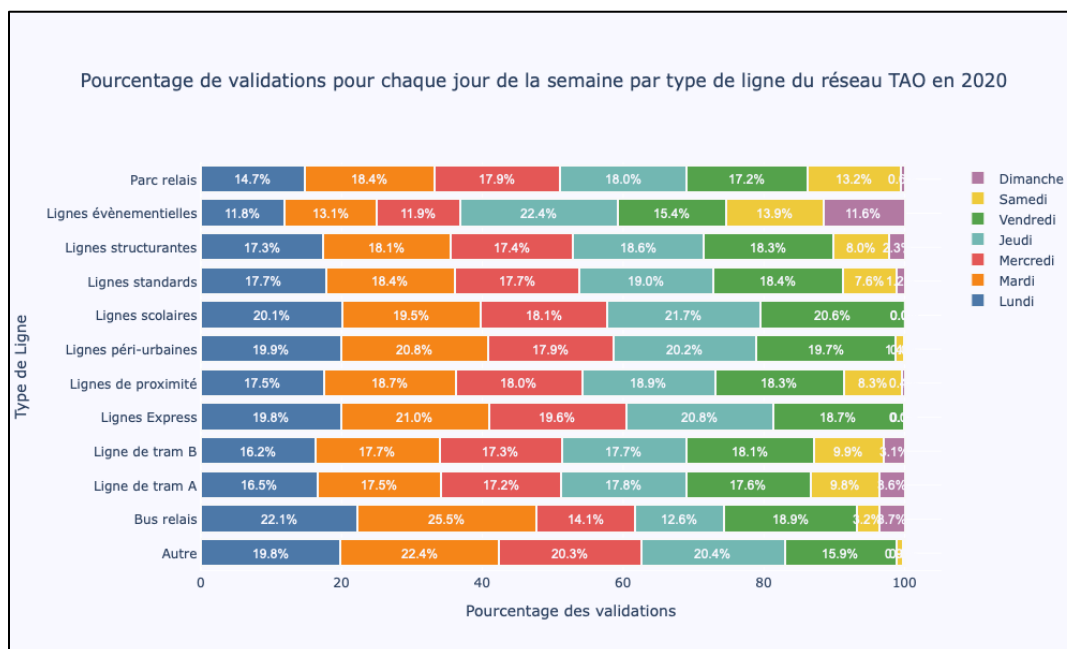


Figure 4-7 - Pourcentage de validations pour chaque jour de la semaine par type de ligne du réseau TAO en 2020

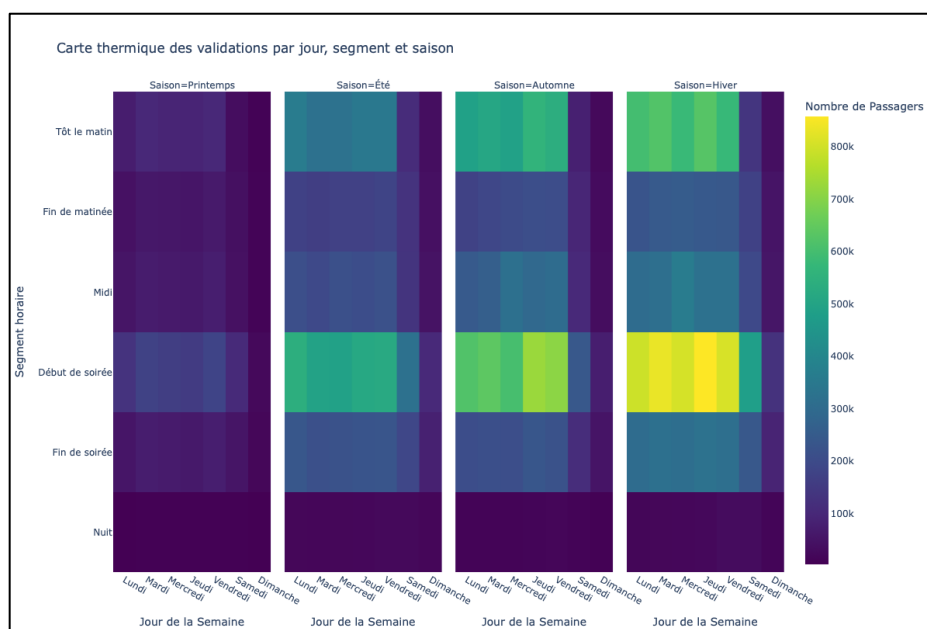


Figure 4-8 - Validations par jour, segment de la journée et saison sur le réseau TAO en 2020

Les deux histogrammes (Figure 4-9 et Figure 4-10) fournissent des perspectives complémentaires sur les validations des passagers. Le premier histogramme montre la distribution des validations journalières sur l'ensemble du réseau. On observe que la majorité des jours enregistrent entre

50 000 et 100 000 validations, avec une moyenne de 94 826 validations par jour. Cette distribution est assez large, indiquant une variabilité significative dans l'utilisation quotidienne des transports en commun, probablement influencée par des facteurs saisonniers, les jours de semaine versus les fins de semaine, et les périodes de confinement.

Le deuxième histogramme représente la distribution des validations par arrêt pour l'ensemble du réseau, en utilisant une échelle logarithmique. La majorité des arrêts enregistrent un nombre modéré de validations, avec une moyenne de 46 029 par arrêt. Ce graphique révèle également une distribution relativement équilibrée, mais avec une dispersion notable vers les arrêts ayant des volumes de passagers très élevés. Quelques arrêts affichent des valeurs exceptionnellement élevées, indiquant leur importance en tant que points de correspondance clés, potentiellement des centres de transfert ou des arrêts situés dans des zones à forte densité urbaine.

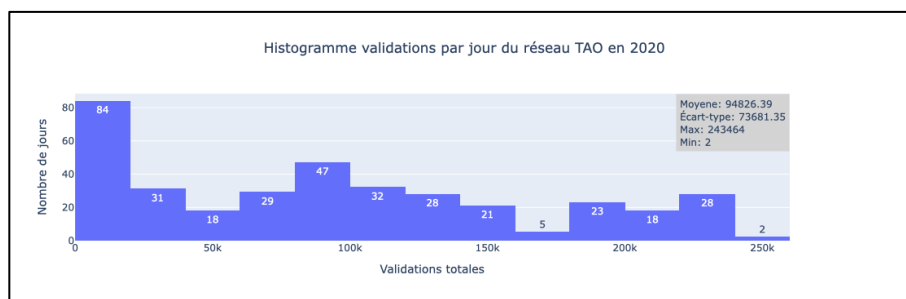


Figure 4-9 - Validations par jour sur le réseau TAO en 2020

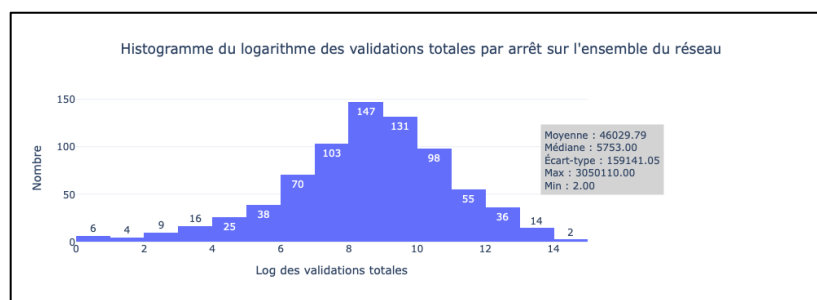


Figure 4-10 - Histogramme des validations par arrêt pour le réseau TAO en 2020

Les validations sur la ligne de tram A (Figure 4-11) sont concentrées sur quelques arrêts principaux, avec une moyenne de 15 401 validations. Cela montre que certains arrêts sont beaucoup plus utilisés que d'autres; ces chiffres sont similaires à ceux du tram B.

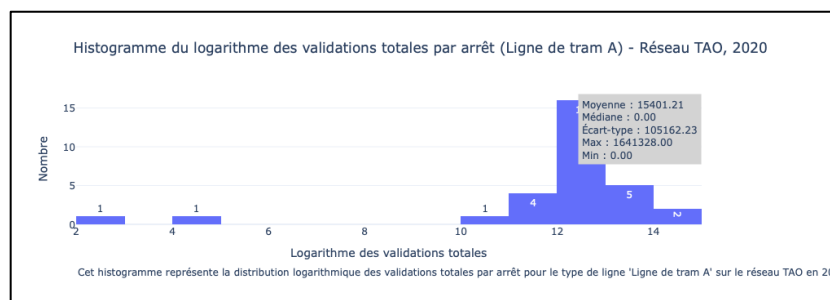


Figure 4-11 - Histogramme des validations par arrêt pour la ligne de tram A du réseau TAO en 2020

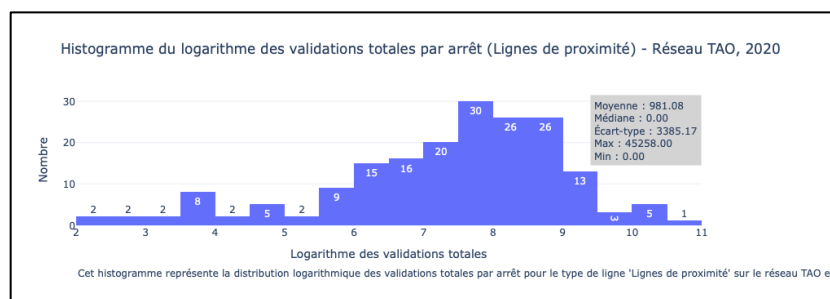


Figure 4-12 - Histogramme des validations par arrêt pour les lignes de proximité du réseau TAO en 2020

Les lignes de proximité (Figure 4-12) montrent une utilisation plus modérée avec une moyenne de 981 validations. Ces lignes desservent probablement des zones résidentielles ou moins denses.

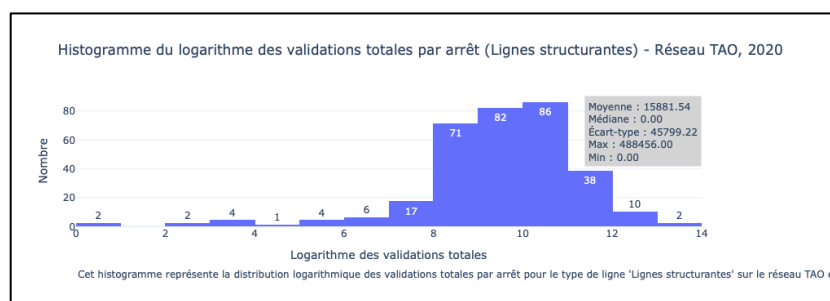


Figure 4-13 - Histogramme des validations par arrêt pour les lignes structurantes du réseau TAO en 2020

Les lignes structurantes (Figure 4-13) montrent une utilisation relativement élevée et uniforme avec une moyenne de 15 881 validations. Cela indique que ces lignes sont essentielles et largement utilisées par les usagers.

Les données montrent une distribution stable des validations au fil des mois (Figure 4-14). Les lignes de proximité et les lignes de tramway représentent la majorité des validations chaque mois, avec des variations mineures. Notons une légère baisse des validations pendant les mois de vacances d'été (juillet et août), ce qui est attendu.

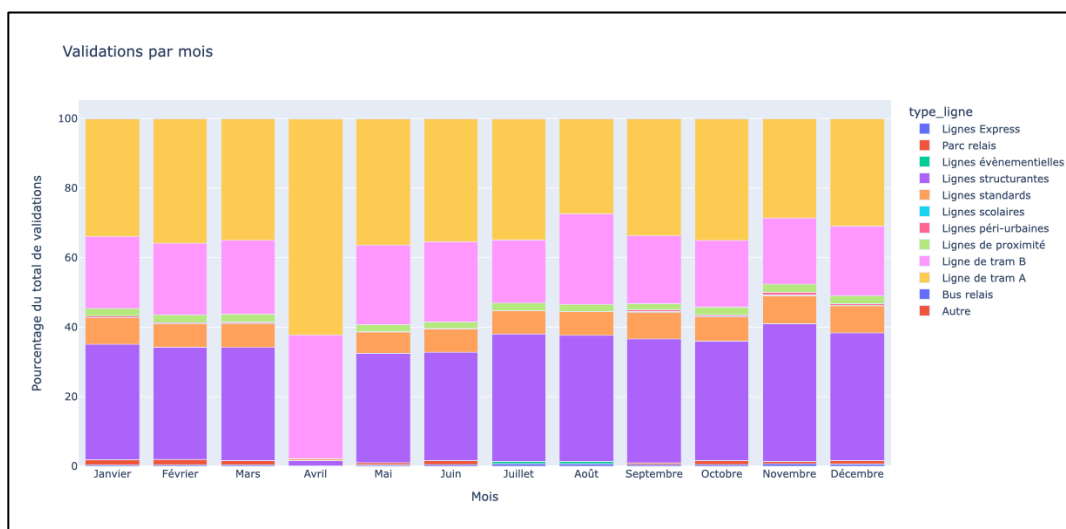


Figure 4-14 - Validations mensuelles sur le réseau TAO en 2020

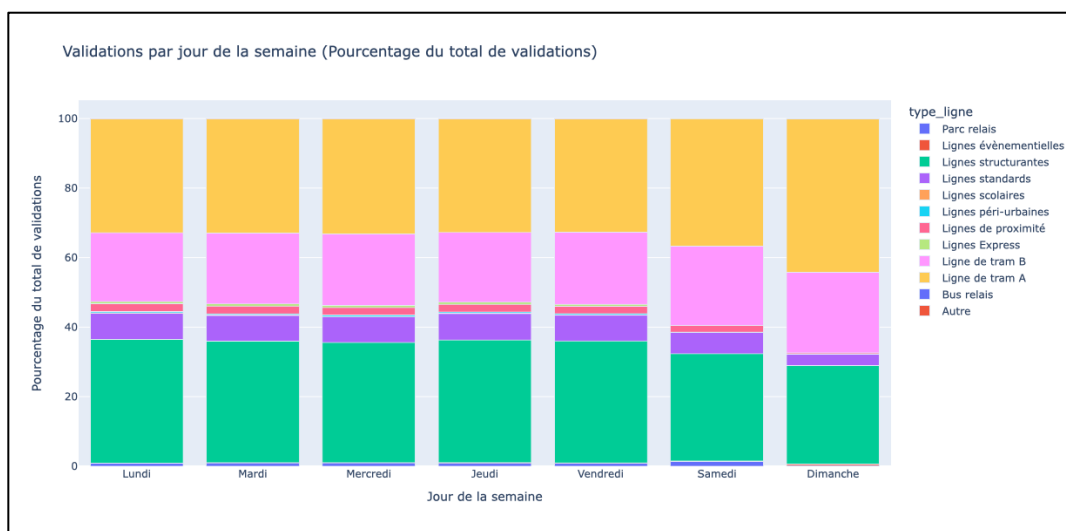


Figure 4-15 - Validations pendant les jours de la semaine sur le réseau TAO en 2020

L'analyse des validations par jour de la semaine révèle que les jours de semaine (du lundi au vendredi) montrent une utilisation relativement constante, avec une légère augmentation les mardis et mercredis. La fin de semaine, en particulier le dimanche, voit une diminution des validations, probablement due à une réduction des activités et des déplacements (Figure 4-15).

Le graphique en anneau (Figure 4-16) et les histogrammes empilés montrent une répartition des validations en fonction des segments de la journée. Voici les définitions des segments utilisés pour cette analyse :

- **Début de matinée (5h-10h)** : Correspond aux trajets matinaux principalement liés aux déplacements domicile-travail ou domicile-école. Il est caractérisé par un fort afflux de passagers, correspondant aux heures de pointe matinales. Ce choix repose sur l'observation des pics d'utilisation des lignes principales et des trams à cette période.
- **Fin de matinée (10h-12h)** : Moins fréquenté, ce segment regroupe principalement les déplacements hors-pointe. Elle inclut principalement des trajets souvent liés à des activités non professionnelles (courses, rendez-vous médicaux, etc.). Ce découpage permet de différencier les heures creuses des périodes plus actives.
- **Midi (12h-14h)** : Lié aux pauses déjeuner ou aux trajets ponctuels. Bien qu'il ne soit pas une période de pointe, il montre une augmentation modérée des validations, en particulier sur les lignes de proximité et les trams.
- **Début de soirée (14h-18h)** : Reflète une reprise progressive des déplacements, liée à des activités sociales, des loisirs, ou à un retour anticipé au domicile. Il permet d'isoler une phase de transition entre les heures creuses et les pointes de soirée.
- **Fin de soirée (18h-22h)** : Correspond au pic d'utilisation pour les trajets domicile après le travail ou les études. Elle inclut également les déplacements liés aux sorties en soirée.
- **Nuit (22h-5h)** : Très peu utilisé, ce segment inclut les déplacements nocturnes. Il couvre des déplacements beaucoup plus rares, souvent associés aux travailleurs de nuit, aux événements spécifiques (spectacles, concerts), ou aux trajets exceptionnels.

Parmi ces segments, le "Début de soirée" est le plus fréquenté avec 35,4% des validations, suivi par le segment "Début de matinée" avec 23,7%. Cela indique que les heures de pointe se situent principalement le matin et en début de soirée, correspondant aux déplacements domicile-travail. Ces tendances confirment l'importance d'ajuster les fréquences des lignes principales pour ces périodes.

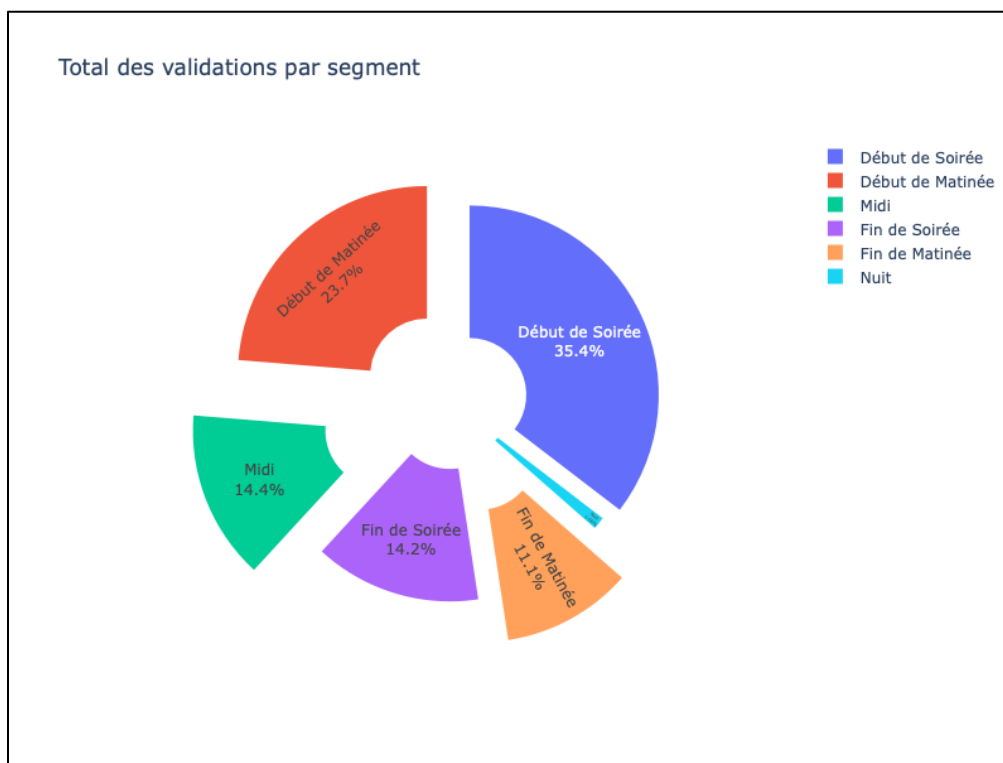


Figure 4-16 - Validations par segment sur le réseau TAO en 2020

Les données montrent une utilisation prédominante des trams et des lignes principales, avec des pics d'utilisation en début de soirée et le matin, reflétant les heures de pointe. Les validations varient légèrement entre les mois et les jours de la semaine, avec une utilisation plus marquée des lignes périurbaines et scolaires en semaine et des lignes de proximité et de tram la fin de semaine. Les bus relais, lignes express et lignes structurantes jouent un rôle essentiel dans les trajets domicile-travail. Les lignes événementielles et les parc relais sont principalement utilisés en début de soirée et le matin, tandis que les lignes scolaires sont surtout utilisées en début et fin de semaine.

Cette analyse descriptive des données de validation des utilisateurs du réseau TAO pour l'année 2020 met en lumière les tendances d'utilisation en fonction des mois, des jours de la semaine, et des segments de la journée. Les résultats montrent une dépendance élevée aux trams et aux lignes principales, avec des pics d'utilisation marqués en début de soirée et le matin. En utilisant ces données comme base, nous pouvons maintenant passer à une analyse plus approfondie par segmentation pour identifier des groupes d'utilisateurs distincts et leurs comportements spécifiques, afin d'affiner encore davantage les stratégies d'amélioration des horaires afin de mieux répondre aux besoins des usagers.

4.2 Analyse détaillée des résultats de la segmentation sur l'année 2020

Pour cette analyse, dix transformations de données sont appliquées sur l'ensemble des données du réseau pour l'année 2020. Les dix transformations sont utilisées sur les validations totales en considérant un pas horaire (60 minutes) sur l'ensemble de l'année. Comme indiqué dans la méthodologie, les quatre méthodes de segmentation sont ensuite testées sur ces données transformées, en utilisant diverses métriques de performance. Les méthodes de segmentation évaluées sont : K-moyennes, segmentation hiérarchique, DTW combiné au K-moyennes et DTW combiné à la segmentation hiérarchique.

Pour chaque méthode, les 10 meilleures transformations sont sélectionnées en fonction de leurs performances, puis les 3 transformations les plus performantes sont identifiées. Enfin, une configuration de segmentation optimale est déterminée pour chaque méthode.

Les étapes de cette analyse sont détaillées dans la méthodologie. Les résultats finaux obtenus pour chaque méthode de segmentation sont présentés de façon détaillée dans les sections suivantes.

Tableau 4-1 - Vecteur "jour-heure"

Hour	0	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Date																				
2020-01-01	110.0	96.0	84.0	218.0	318.0	438.0	620.0	898.0	1100.0	1096.0	1568.0	1786.0	1792.0	2014.0	1920.0	1268.0	746.0	466.0	266.0	160.0
2020-01-02	36.0	746.0	1928.0	4572.0	6362.0	5378.0	5876.0	6992.0	7306.0	7978.0	10952.0	11392.0	12086.0	13488.0	8996.0	5238.0	2622.0	1322.0	540.0	378.0
2020-01-03	52.0	852.0	2040.0	4886.0	6948.0	5984.0	7014.0	8032.0	8550.0	9272.0	11668.0	12280.0	12792.0	13748.0	9118.0	5714.0	2864.0	1774.0	948.0	746.0
2020-01-04	144.0	502.0	694.0	1388.0	2670.0	4654.0	6008.0	7012.0	6966.0	7612.0	10520.0	10022.0	10014.0	9332.0	8138.0	5748.0	2830.0	1464.0	798.0	826.0
2020-01-05	114.0	68.0	88.0	402.0	886.0	1452.0	1650.0	2290.0	2008.0	2504.0	2646.0	2362.0	2806.0	2802.0	2806.0	2222.0	1392.0	796.0	446.0	196.0

Avant de commencer l'analyse, rappelons que le vecteur "jour-heure" (Tableau 4-1) est une structure de données qui agrège les données de validation des passagers par heure pour chaque jour de la semaine. L'agrégation des données de validation des passagers par heure pour chaque jour permet de dégager des tendances claires. Par exemple, en semaine, les heures de pointe matinales et en fin de journée peuvent être facilement identifiées, souvent correspondant aux périodes de trajets domicile-travail. En revanche, les fins de semaine peuvent montrer des pics d'activité à des heures différentes, souvent liées aux activités de loisirs. Enfin, l'utilisation du vecteur "jour-heure" pour l'agrégation horaire quotidienne fournit une vue détaillée et dynamique de l'utilisation du réseau.

4.2.1 La méthode des K-moyennes (« *K-means* »)

Sélection des 10 meilleurs résultats

Le tableau ci-dessous (Tableau 4-2) présente les dix meilleures transformations de données appliquées aux données de l'année 2020, classées en fonction de leurs performances en termes de regroupement. Pour chaque transformation, nous avons indiqué le nombre optimal de groupes (Meilleur K), la métrique de distance utilisée (Meilleure métrique) et le meilleur score de silhouette obtenu (Meilleur score de silhouette). De plus, la colonne "K Optimal (Écart)" montre le nombre de groupes optimal déterminé par la statistique de l'écart.

Tableau 4-2 - 10 meilleurs résultats sur la segmentation de l'année 2020 (K-moyennes)

Transformation	Meilleur K	Meilleure métrique	Meilleur score de silhouette	K Optimal (Écart)
Standardisation	3	Cosinus	0.696277	18
Min-max	3	Cosinus	0.684362	20
Logarithmique	2	Manhattan	0.637234	20
Racine carrée	3	Manhattan	0.536541	20
Box-Cox	3	Manhattan	0.562654	18
Inverse	4	Euclidienne	0.916219	18
Quantile uniforme	2	Cosinus	0.672051	19
Quantile normal	8	Cosinus	0.771808	18
Yeo Johnson	2	Cosinus	0.742942	19
Différenciation	2	Euclidienne	0.700496	17

Tableau 4-2- 10 meilleurs résultats sur la segmentation de l'année 2020 (K-moyennes, suite et fin)

Transformation	Meilleur K	Meilleure métrique	Meilleur score de silhouette	K Optimal (Écart)
Aucune transformation	2	Cosinus	0.590911	19

Parmi les 10 meilleurs résultats présentés plus haut, les trois transformations les plus performantes pour K-moyennes sont identifiées : standardisation, min-max et inverse. Ces choix sont justifiés par des scores de silhouette élevés et des valeurs de SSE faibles, comme illustré dans les graphiques ci-dessous. La transformation standard obtient un score de silhouette de 0,696 (distance cosinus, Meilleur K=3), montrant une bonne définition des groupes. La transformation min-max affiche une performance similaire avec un score de 0,684 (distance cosinus, Meilleur K=3). Enfin, la transformation inverse se distingue par un score exceptionnel de 0,916 (distance euclidienne, Meilleur K=4), démontrant une excellente qualité de regroupement avec des groupes bien définis.

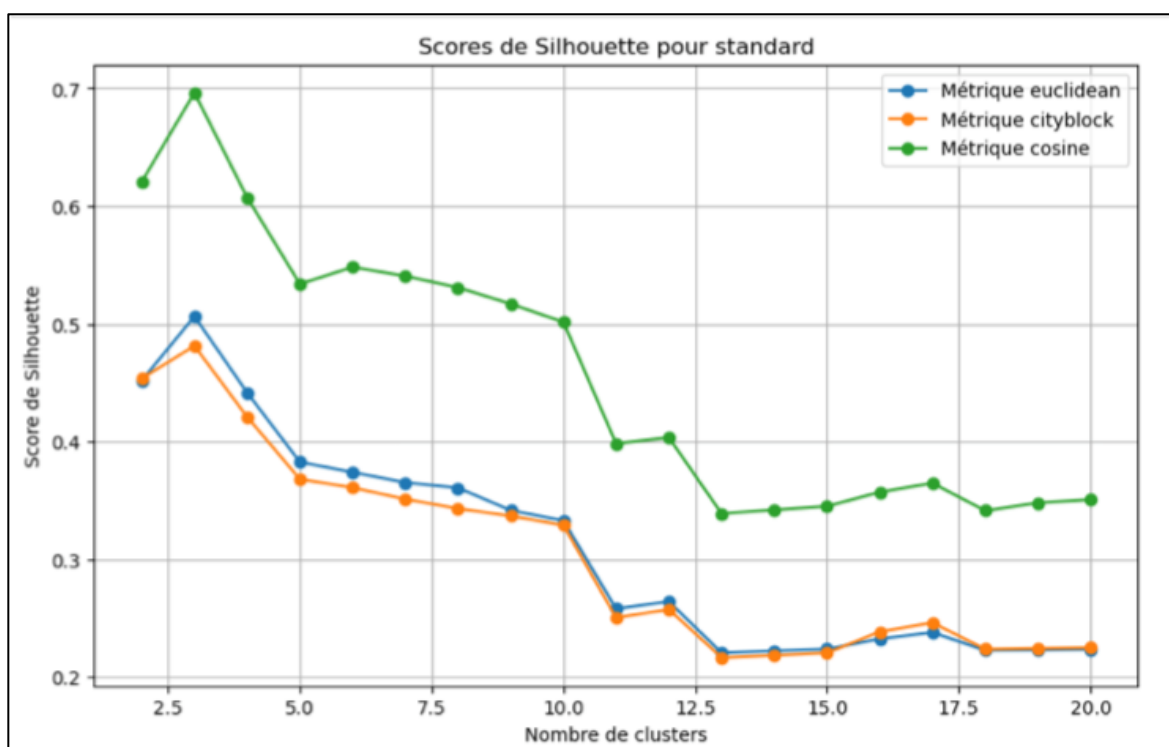


Figure 4-17 - Scores de silhouette pour la transformation standard K-moyennes

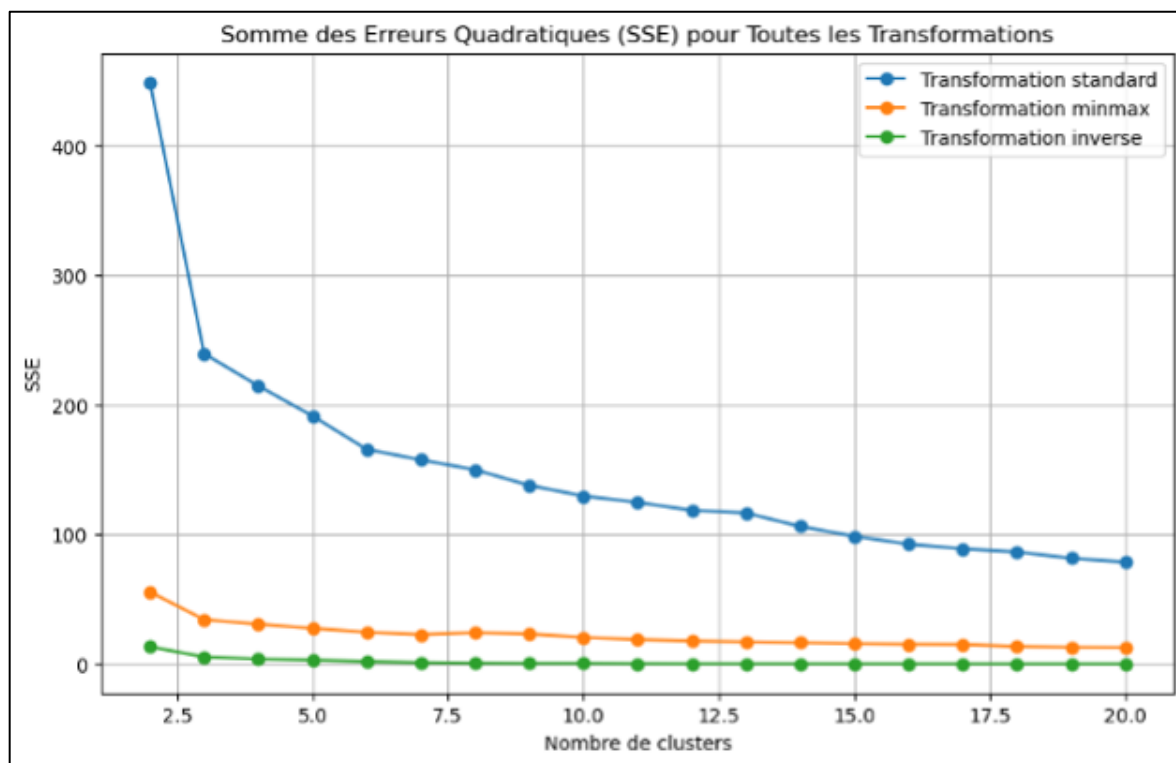


Figure 4-18 - Scores SSE pour les trois meilleures transformations K-moyennes

La figure 33 montre le nombre optimal de groupes selon la statistique de l'écart (« *Gap Statistic* ») et le nombre de groupes sélectionnés (Meilleur K) pour les transformations standard, min-max et inverse avec l'algorithme K-moyennes. Les barres bleues indiquent que la statistique de l'écart suggère un nombre élevé de groupes : 18 pour standard et inverse, et 20 pour min-max. En revanche, les points rouges montrent que les scores de silhouette maximaux indiquent trois groupes pour standard et min-max, et quatre groupes pour inverse. Cette divergence met en évidence que, bien que la statistique de l'écart favorise plus de groupes pour maximiser la variation intra-groupe, les scores de silhouette, qui évaluent la cohésion et la séparation des groupes, préfèrent un nombre plus restreint de groupes. Cela souligne l'importance d'utiliser plusieurs métriques pour déterminer le nombre optimal de groupes, équilibrant la cohésion interne et la séparation externe des groupes.

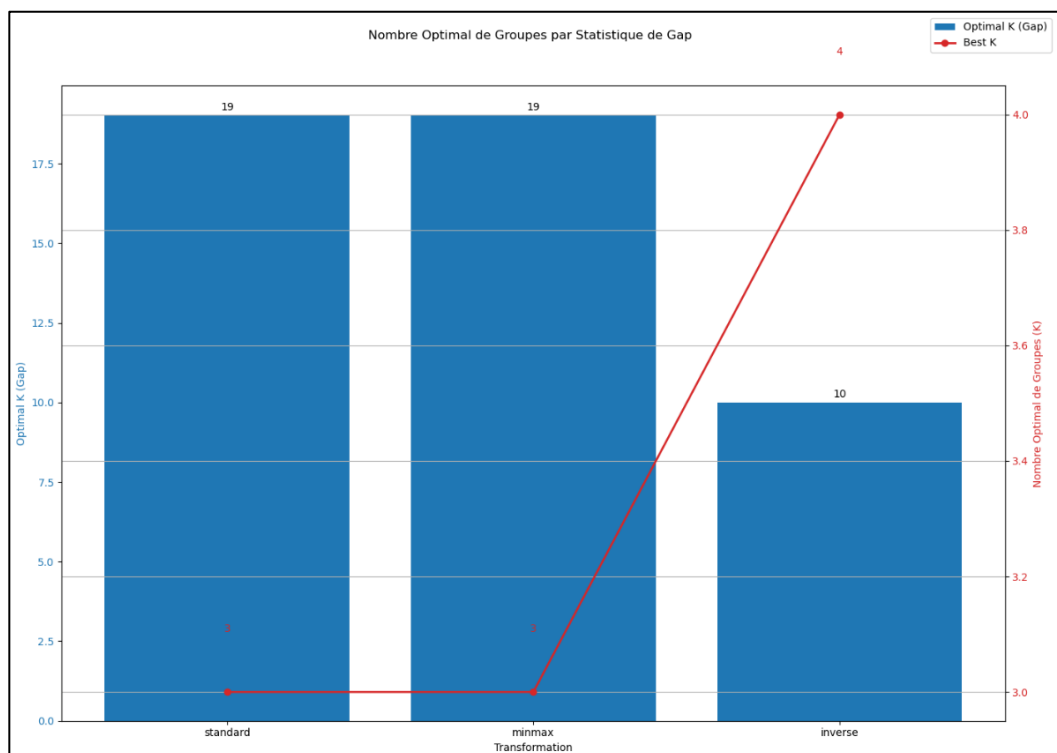


Figure 4-19 - K-Gap pour les trois meilleurs résultats du K-moyennes

Pour sélectionner le groupement final pour l'algorithme K-moyennes, une analyse approfondie des trois meilleures transformations (standard, min-max et inverse) est réalisée en se basant sur diverses métriques de performance : WCSS, indice de Davies-Bouldin, indice de Calinski-Harabasz, indice de Dunn, cohésion et séparation.

Évaluation des trois meilleurs groupements

Tableau 4-3 - Évaluation des trois meilleurs groupements K-moyennes

Transformation	WCSS	Indice de Davies-Bouldin	Indice de Calinski-Harabasz	Dunn Index	Cohésion	Séparation	Meilleur K
Standard	239,31	0,77	449,26	0,05	233,13	417,67	3
Min-max	34,16	0,85	378,86	0,12	89,34	142,54	3
Inverse	3,77	0,72	584,38	0,49	17,99	31,30	4

Le tableau (Tableau 4-3) montre que la transformation inverse offre les meilleurs résultats pour la segmentation, avec des groupes très compacts et bien séparés, comme l'indiquent ses faibles valeurs de WCSS et d'indice de Davies-Bouldin, ainsi que son indice de Calinski-Harabasz élevé. Cependant, la transformation standard présente également des performances robustes avec une bonne définition des groupes et des résultats équilibrés en termes de compacité et de séparation:

1. **Performance Équilibrée** : Bien que l'inverse ait les meilleurs scores sur plusieurs métriques, la transformation standard présente une performance équilibrée avec un bon compromis entre compacité (WCSS) et séparation (indice de Davies-Bouldin et de Calinski-Harabasz). Cela indique une robustesse et une stabilité des groupes, ce qui est important pour des analyses pratiques et opérationnelles.
2. **Cohérence des Groupes** : La transformation standard offre des valeurs de cohésion et de séparation élevées, suggérant des groupes bien définis et distincts, ce qui est crucial pour l'interprétation et l'utilisation des résultats de segmentation.
3. **Simplicité et Interprétabilité** : La standardisation est une méthode couramment utilisée et bien comprise, ce qui facilite l'interprétation des résultats et leur communication à un public plus large. De plus, elle ne modifie pas radicalement la distribution des données, conservant ainsi leur structure intrinsèque.
4. **Robustesse** : La transformation standard est moins sensible aux variations extrêmes des données comparée aux transformations comme l'inverse, qui peuvent être influencées par des valeurs aberrantes. Bien que la transformation inverse montre des performances supérieures dans certaines métriques, la transformation standard est choisie pour son équilibre, sa robustesse et sa simplicité, offrant une solution de segmentation fiable et interprétable.

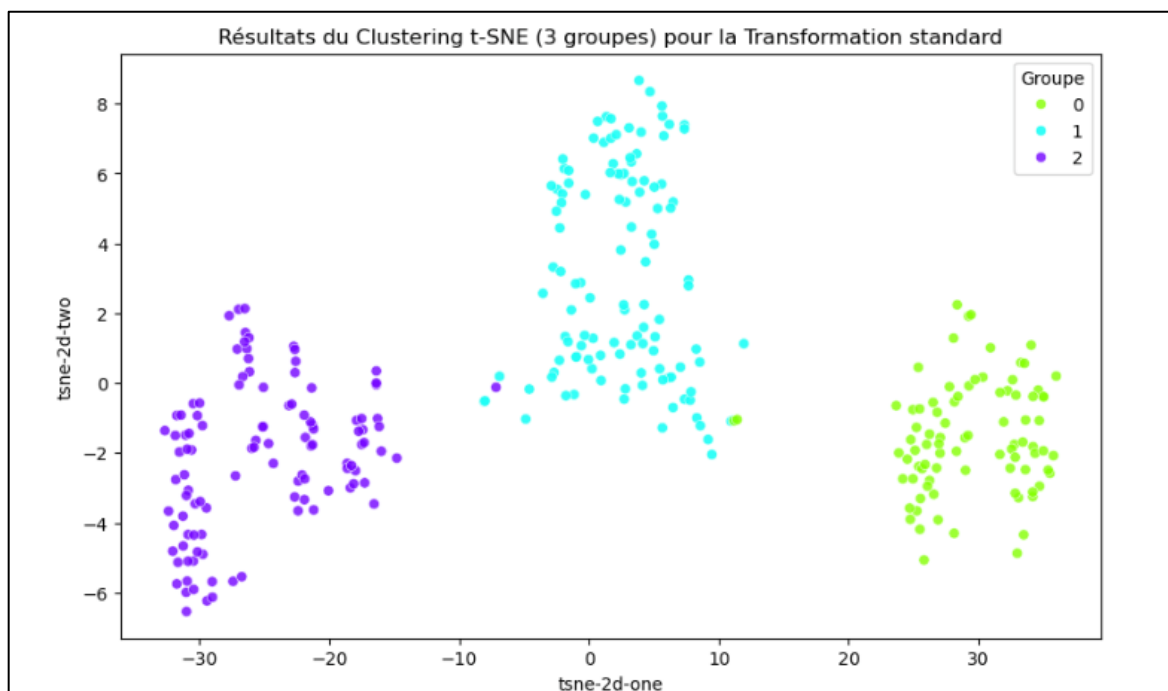


Figure 4-20 - Groupes pour la transformation standard K-moyennes

La visualisation (à l'aide de la technique t-SNE « *t-Distributed Stochastic Neighbor Embedding* ») montre clairement la séparation des trois groupes dans l'espace bidimensionnel (Figure 4-20). Chaque groupe de points de la même couleur représente des données qui sont similaires selon les vecteurs des heures, et distinctes des autres groupes :

- **Groupe 0 (Vert) :** Les points sont regroupés de manière serrée, indiquant une forte similarité entre les données au sein de ce groupe.
- **Groupe 1 (Cyan) :** Bien que légèrement plus dispersés que le groupe 0, les points montrent également une bonne cohésion interne.
- **Groupe 2 (Violet) :** Ce groupe présente une structure distincte avec une bonne séparation des autres groupes.

La bonne séparation et la cohésion des groupes visibles sur le graphique confirment que la transformation standard et l'algorithme K-moyennes ont réussi à identifier des groupes significatifs dans les données de l'année 2020.

Analyse du regroupement final

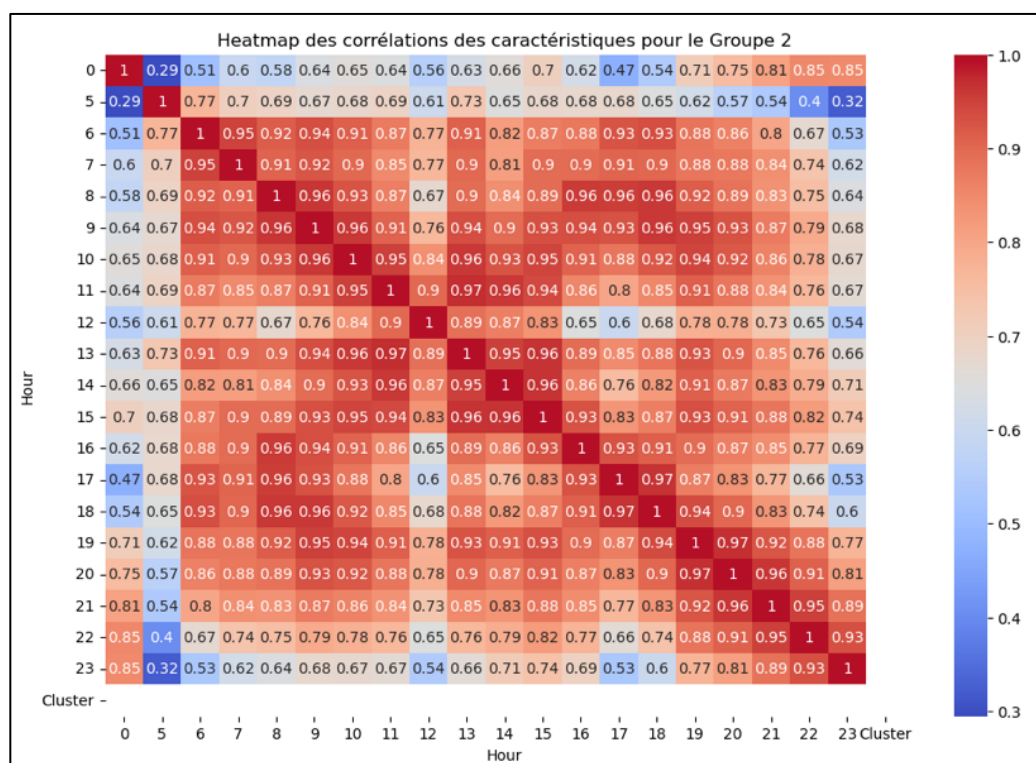


Figure 4-21 - Corrélations pour le groupe 0, transformation standard (K-moyennes)

Dans la carte de chaleur (Figure 4-21) pour le premier groupe, les corrélations positives élevées entre les heures consécutives indiquent que les valeurs suivent des tendances journalières similaires. Par exemple, les corrélations élevées entre 1h et 2h (0.78) ou 5h et 6h (0.88) montrent que les activités ou les mesures prises à ces heures sont fortement liées. Cela suggère une continuité dans les comportements ou les processus au cours de ces périodes.

Les corrélations négatives entre certaines heures, comme entre 0h et 23h (-0.26), indiquent des périodes de haute et basse activité. Cela peut être crucial pour la planification et l'allocation des ressources, en s'assurant que les périodes de pointe bénéficient d'une attention et de ressources appropriées, tandis que les périodes de faible activité peuvent être optimisées. Elles peuvent être utilisées pour ajuster les horaires de travail du personnel en fonction des périodes de forte et faible activité pour une utilisation optimale des ressources humaines; et permettent de segmenter efficacement la journée pour des analyses plus approfondies et des décisions opérationnelles éclairées. Une analyse similaire peut-être appliquée pour les deux autres groupes.

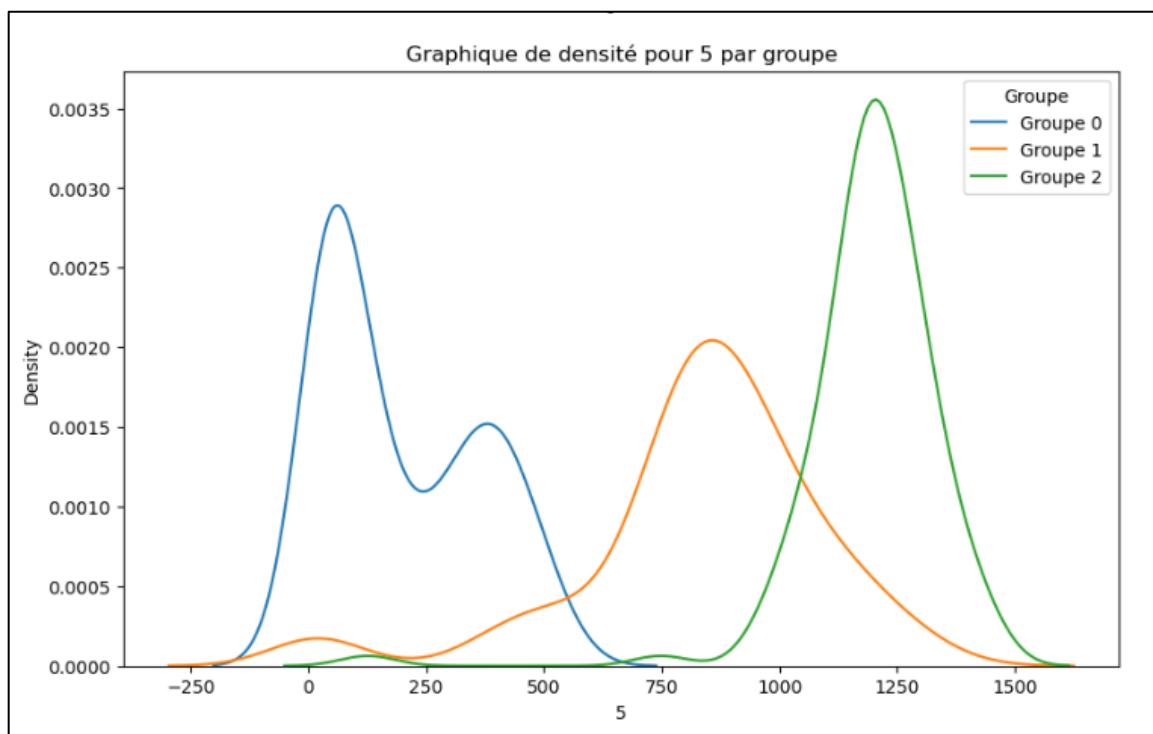


Figure 4-22 – Densité pour 5 heures, transformation standard (K-moyennes)

En comparant les groupes (Figure 4-22), on observe que le Groupe 1 montre une distribution plus concentrée, suggérant que les entités de ce groupe sont plus homogènes dans leur comportement à 5 heures. Les Groupes 0 et 2 présentent une plus grande variabilité dans leurs valeurs, indiquant des comportements plus diversifiés à cette heure. De plus, les pics de densité des trois groupes sont décalés, montrant des valeurs moyennes différentes entre les groupes.

Les différences dans les distributions des groupes à 5 heures peuvent être utilisées pour segmenter les entités en fonction de leurs comportements spécifiques à cette heure :

- Le Groupe 0 montre une grande variabilité des valeurs à 5 heures, indiquant que les entités de ce groupe présentent des comportements diversifiés. Pour ce groupe, une stratégie flexible et adaptable est recommandée. Il est crucial de mettre en place des processus flexibles qui peuvent s'adapter à la diversité des activités. Cela pourrait inclure des horaires de travail flexibles ou des systèmes de production ajustables.
- Le Groupe 1 présente une homogénéité élevée, indiquant que les entités de ce groupe ont des comportements ou des valeurs similaires à 5 heures. Une stratégie standardisée et

optimisée peut être efficace pour ce groupe. La mise en place de processus standardisés maximiserait l'efficacité, en se basant sur les comportements prévisibles de ce groupe.

- Le Groupe 2 montre une variabilité intermédiaire, ce qui suggère que les entités de ce groupe ont des comportements modérément variés à 5 heures. Une approche équilibrée combinant standardisation et flexibilité est recommandée. La combinaison de processus standardisés pour les comportements prévisibles avec des capacités d'ajustement pour gérer les variations modérées serait bénéfique. Enfin, effectuer des réévaluations régulières des comportements permettrait d'ajuster les stratégies en fonction des changements observés dans les patrons d'activités, assurant ainsi une réponse continue et appropriée aux besoins des usagers

Cette analyse peut s'appliquer pour chaque heure et des stratégies d'ajustement des horaires peuvent être trouvées.

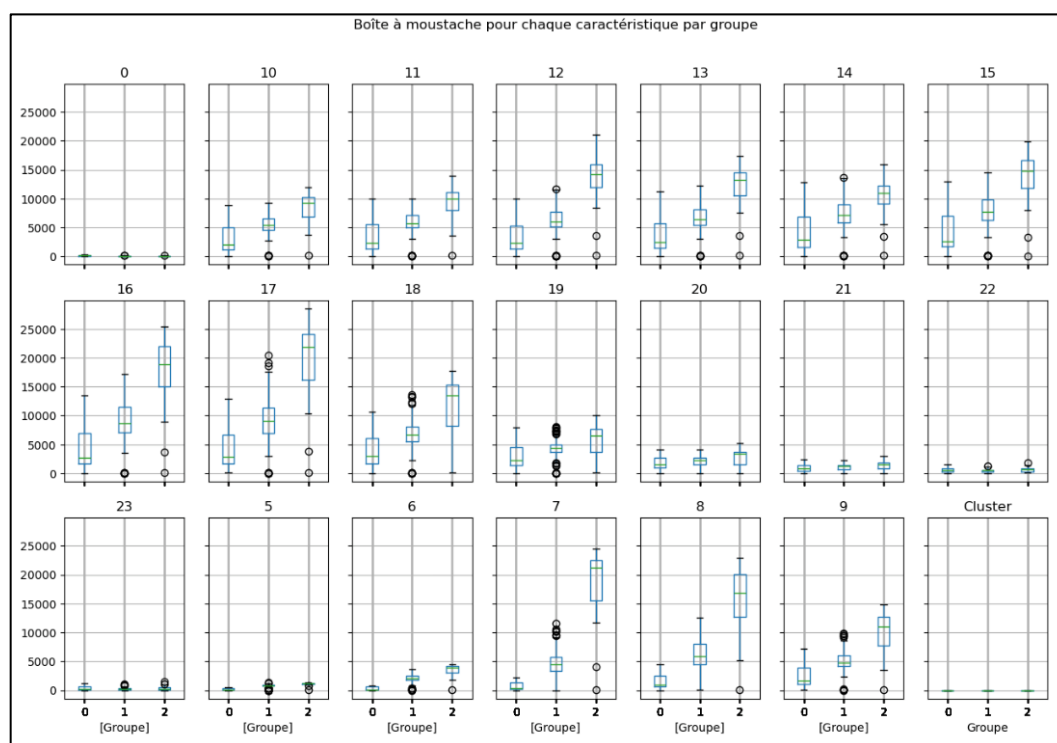


Figure 4-23 - Box plot des groupes obtenus K-moyennes - Standard

Les box plots ci-dessus (Figure 4-23) montrent la distribution des valeurs pour chaque heure, segmentées par les trois groupes identifiés. Les entités du Groupe 0 présentent une variabilité importante tout au long des heures, comme le montre l'étendue des boîtes et la présence de

nombreux points aberrants. Cela indique des comportements hétérogènes, avec des valeurs qui varient considérablement d'une entité à l'autre à chaque heure. En revanche, le Groupe 1 montre des distributions plus concentrées pour la plupart des heures, avec des boîtes plus serrées et moins de points aberrants. Cela suggère une plus grande homogénéité dans les comportements des entités à chaque heure, avec des valeurs plus cohérentes entre les différentes entités. Le Groupe 2, quant à lui, présente une variabilité intermédiaire. Les boîtes sont plus larges que celles du Groupe 1 mais moins que celles du Groupe 0, avec quelques points aberrants. Cela indique des comportements modérément variés parmi les entités à chaque heure. Les box plots révèlent que le Groupe 1 est le plus homogène en termes de valeurs horaires, tandis que le Groupe 0 est le plus hétérogène. Le Groupe 2 se situe entre les deux en termes de variabilité.

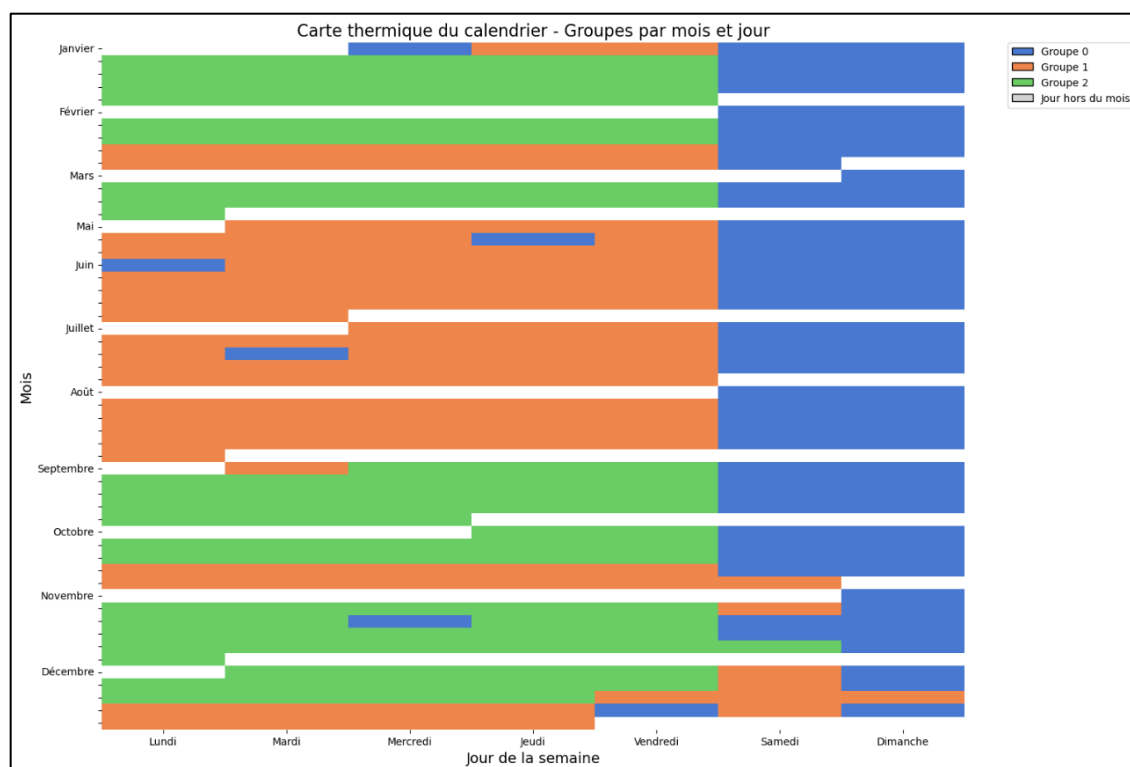


Figure 4-24 - Calendrier des groupes K-moyennes, standard

Le calendrier (Figure 4-24) cartographique des groupes illustre la répartition des groupes tout au long de l'année 2020. Chaque jour de l'année est coloré selon le groupe prédominant : le vert pour le Groupe 0, l'orange pour le Groupe 1 et le bleu pour le Groupe 2. Cette visualisation permet d'identifier les tendances saisonnières et hebdomadaires des groupes.

Le Groupe 0 (vert) est majoritairement présent en début et en fin d'année, avec des périodes ici ponctuelles tout au long de l'année. Cela indique que les comportements associés à ce groupe sont plus fréquents pendant ces périodes. Le Groupe 1 (orange) domine pendant les mois de février, juin, juillet et novembre, montrant des pics saisonniers spécifiques et suggérant des besoins de planification accrus durant ces périodes. Le Groupe 2 (bleu) est prévalent pendant les mois de mars, mai et décembre, indiquant des périodes spécifiques où les comportements liés à ce groupe sont plus communs.

Pour le Groupe 0, les ressources doivent être planifiées pour répondre à la grande variabilité des comportements, en particulier au début et à la fin de l'année. Pour le Groupe 1, la période de juin à juillet et en novembre nécessitera une allocation de ressources spécifique pour répondre aux besoins accrus. Des stratégies de gestion des pics d'activité doivent être mises en place pour assurer l'efficacité opérationnelle pendant ces mois. Pour le Groupe 2, une planification rigoureuse pendant les mois de mars, mai et décembre est cruciale.

Analyse statistique approfondie des résultats du groupement final

Les statistiques pour le Groupe 0 fournissent une analyse détaillée des heures de la journée, des jours et des mois. Les heures avec les moyennes les plus élevées, comme les heures 14 à 17, indiquent des périodes de forte activité. Il peut être bénéfique de prévoir davantage de ressources pendant ces heures pour gérer l'augmentation du volume de passagers. En revanche, les heures 0, 5 et 6 montrent des moyennes relativement basses, ce qui suggère des périodes de faible activité.

Les écarts-types et les IQR (Interquartile Range) élevés pour des heures comme 14 à 17 montrent une grande variabilité, indiquant que des ajustements dynamiques des ressources peuvent être nécessaires pour répondre à la variabilité de la demande. Les mesures d'asymétrie et de kurtosis, qui montrent la distribution des données, indiquent une distribution biaisée vers la droite avec des valeurs extrêmes, ce qui aide à mieux réaliser les prévisions et à identifier les périodes potentielles de surcharge.

Enfin, les statistiques sur les jours et les mois montrent des tendances plus larges. Par exemple, une moyenne plus élevée pendant certains jours ou mois peut indiquer des périodes de pointe saisonnières qui nécessitent une planification à long terme.

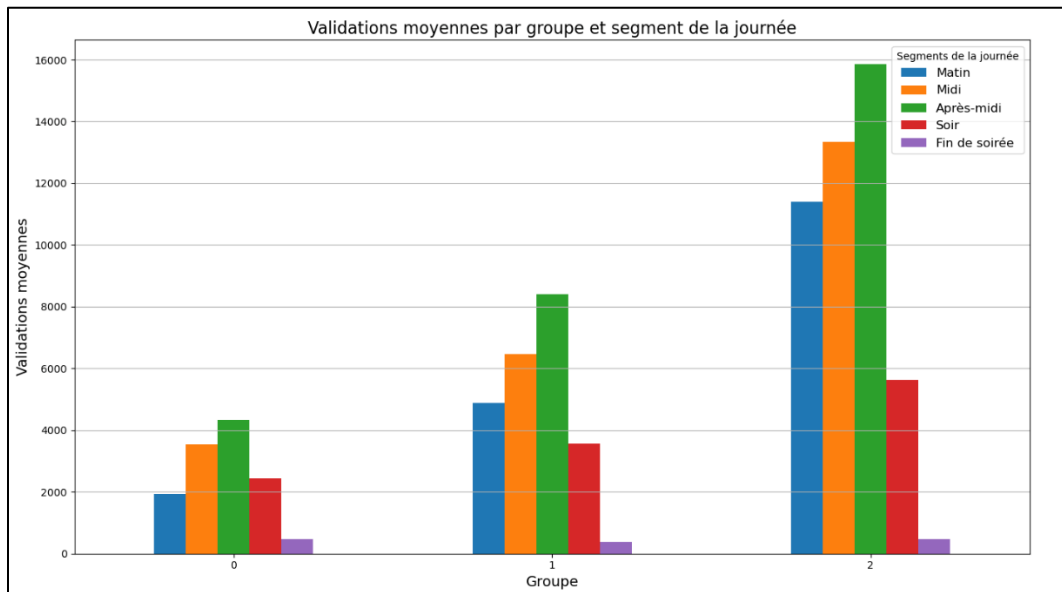


Figure 4-25 - Moyenne des validations par heure par segment de la journée par groupe - K-moyennes, standard

Les tendances observées dans la figure 39 impliquent plusieurs stratégies pour la planification des horaires. Pour le Groupe 0, l'allocation des ressources devrait se concentrer sur l'après-midi et le midi, avec une couverture adéquate le matin et le soir, et des ressources minimales en fin de soirée. Pour le Groupe 1, l'accent devrait être mis sur l'après-midi, suivi du midi et du soir, avec des ressources suffisantes le matin et minimales en fin de soirée. Le Groupe 2 nécessite le plus de ressources, en particulier l'après-midi, mais aussi une allocation significative le matin, le midi et le soir, avec une certaine attention même en fin de soirée en raison de son activité relativement élevée.

En général, l'après-midi doit être priorisé pour l'allocation des ressources à travers tous les groupes en raison du nombre élevé de validations. Le midi et le matin nécessitent également des ressources substantielles, surtout pour les Groupes 1 et 2. Le soir demande des ressources adéquates, bien que moins que l'après-midi et le midi, tandis que la fin de soirée peut être minimalement dotée en personnel pour les Groupes 0 et 1, mais nécessite encore une attention pour le Groupe 2. Cette allocation stratégique basée sur les tendances de validation optimisera la gestion des ressources et assurera une opération efficace.

4.2.2 La segmentation hiérarchique

Sélection des 10 meilleurs résultats

Tableau 4-4 - 10 meilleurs résultats de la segmentation hiérarchique

Transformation	K	Métrique	Méthode	Score de silhouette	Cophénétique	K Optimal (Écart)
Standardisation	8	Cosinus	Moyenne	0,72	0,79	20
Min-max	2	Cosinus	Moyenne	0,72	0,75	19
Logarithmique	2	Cosinus	Simple	0,92	0,70	18
Racine Carrée	2	Cosinus	Moyenne	0,72	0,57	19
Box-Cox	2	Cosinus	Moyenne	0,90	0,46	20
Inverse	2	Manhattan	Simple	0,95	0,95	7
Quantile uniforme	6	Cosinus	Complète	0,66	0,72	20
Quantile normal	20	Cosinus	Moyenne	0,82	0,92	19
Yeo Johnson	3	Cosinus	Complète	0,74	0,81	19
Différenciation	2	Euclidienne	Moyenne	0,69	0,91	19
None	2	Cosinus	Moyenne	0,73	0,63	20

Les trois meilleures transformations sélectionnées pour l'analyse – « standard », quantile uniforme, et quantile normal - ont démontré des performances optimales selon des critères clés tels que le score de silhouette, la corrélation cophénétique et l'K Optimal (Écart). La transformation standard s'est distinguée avec un score de silhouette de 0.729460 et une corrélation cophénétique de

0.793240, indiquant une forte cohésion interne et une bonne séparation des groupes avec 8 groupes optimaux. La transformation quantile uniforme, avec un score de silhouette de 0.663093 et une corrélation cophénétique de 0.723270, a montré un bon équilibre entre la cohésion interne et la séparation des groupes, particulièrement efficace pour les distributions de données non normales. Enfin, la transformation quantile normal a obtenu un score de silhouette exceptionnel de 0.829890 et une corrélation cophénétique de 0.921471, produisant des groupes bien définis et cohérents avec 20 groupes, ce qui est crucial pour des analyses nécessitant une granularité fine et des connaissances détaillées. Ces transformations offrent des solutions robustes et fiables pour améliorer la précision des résultats de segmentation, chaque transformation étant adaptée à différents types de données et besoins d'analyse.

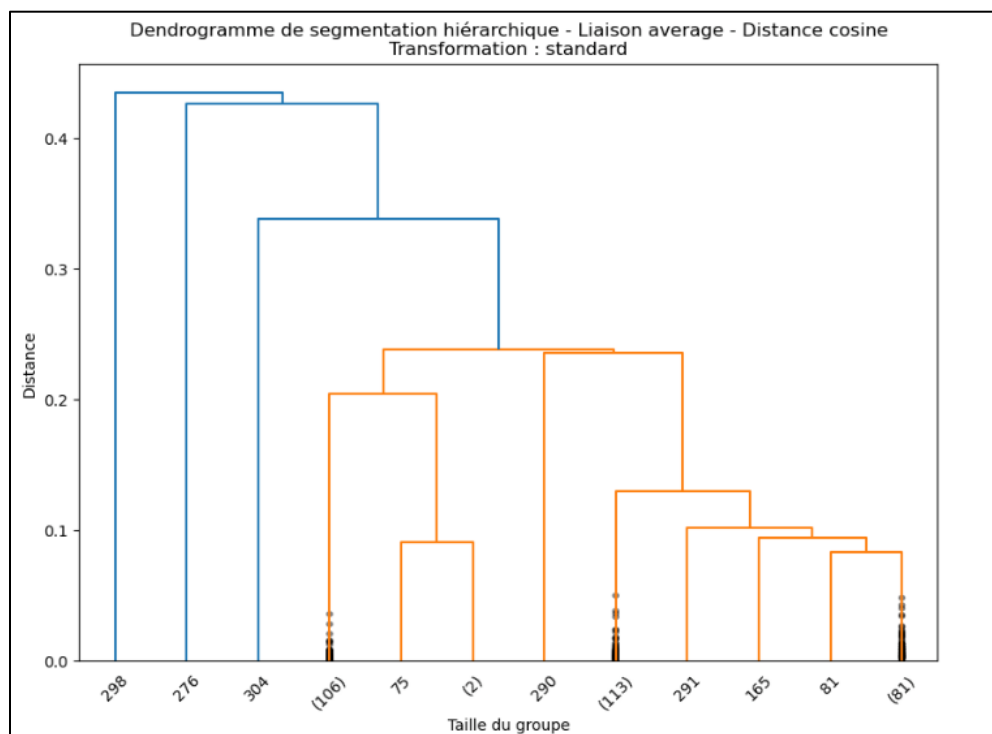


Figure 4-26 - Dendrogramme de la transformation standard

Les trois dendrogrammes (Figure 4-26, Figure 4-27 et Figure 4-28) montrent des structures hiérarchiques distinctes mais cohérentes, chaque transformation offrant une perspective unique sur la manière dont les données peuvent être segmentées en 6 groupes. La transformation standard révèle une distribution équilibrée des groupes, tandis que les transformations quantile uniforme et quantile normale montrent des structures de fusion légèrement différentes mais tout aussi efficaces pour identifier des sous-groupes distincts.

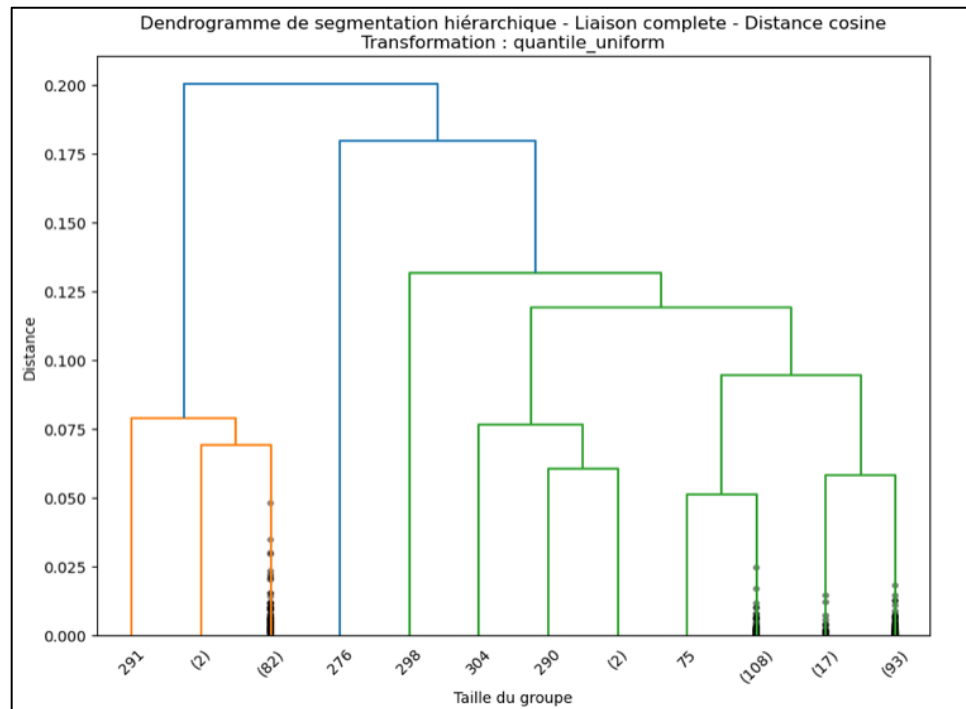


Figure 4-27 - Dendrogramme de la transformation quantile uniforme

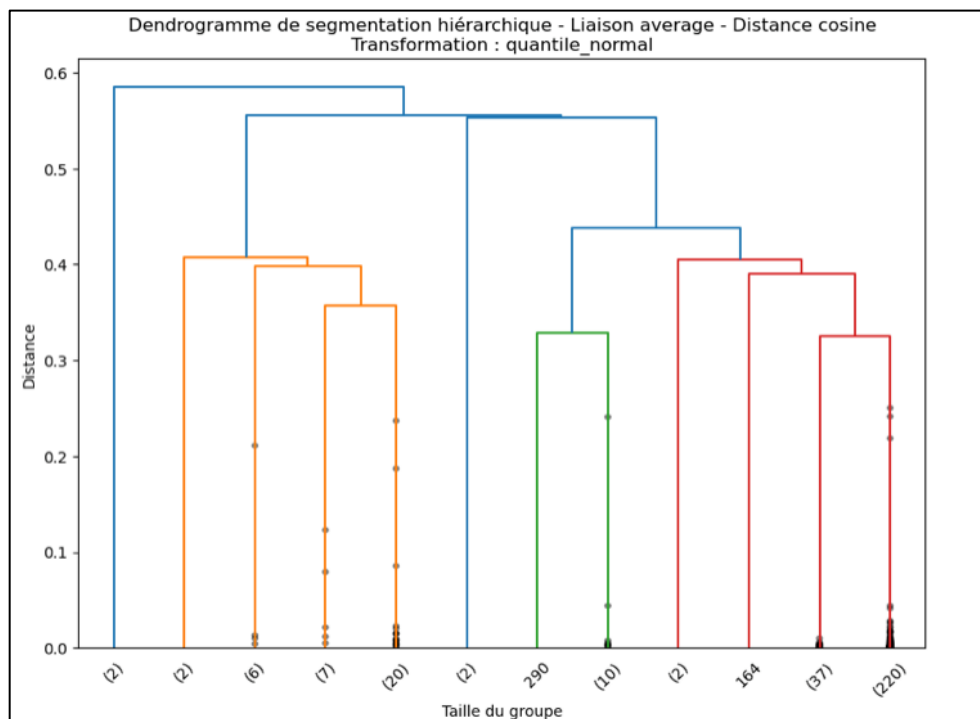


Figure 4-28 - Dendrogramme transformation quantile normale

Évaluation des trois meilleurs groupements

Tableau 4-5 - Trois meilleurs résultats de la segmentation hiérarchique

Transformation	K	Métrique	Méthode	Score de silhouette	Cophénétique	K Optimal (Écart)
Standard	8	Cosinus	Moyenne	0.72	0.79	20
Quantile uniforme	6	Cosinus	Complète	0.66	0.72	20
Quantile normal	20	Cosinus	Moyenne	0.82	0.92	19

Tableau 4-6- Évaluation des trois meilleurs groupements

Davies-Bouldin	Calinski-Harabasz	Dunn	Cohésion	Séparation	Score de silhouette	Transformation	K Optimal (Écart)
0,51	566,14	0,32	968,03	5798,68	0,24	Standard	20
0,79	94,24	0,983	56,775	140,51	0,29	Quantile uniforme	20
0,70	114,73	1,32	415,73	1211,38	0,65	Quantile normal	19

L'analyse des groupes (Tableau 4-5 et Tableau 4-6) basée sur plusieurs critères tels que l'indice de Davies-Bouldin, l'indice de Calinski-Harabasz, l'indice de Dunn, la cohésion, la séparation et le score de silhouette révèle que chaque transformation présente des avantages spécifiques. Cependant, en comparant ces critères, la transformation standard se distingue par son excellent indice de Calinski-Harabasz (566.146680) et sa très bonne séparation des groupes (5798.683034). Bien que le score de silhouette de 0.243154 soit modéré, les avantages en termes de cohésion modérée (968.031838) et de séparation excellente en font une option équilibrée et robuste. En

conséquence, pour une segmentation horaire efficace et pratique, il est recommandé de prioriser les groupes issus de la transformation standard, car ils offrent une bonne discrimination entre les périodes horaires tout en maintenant une cohésion interne adéquate.

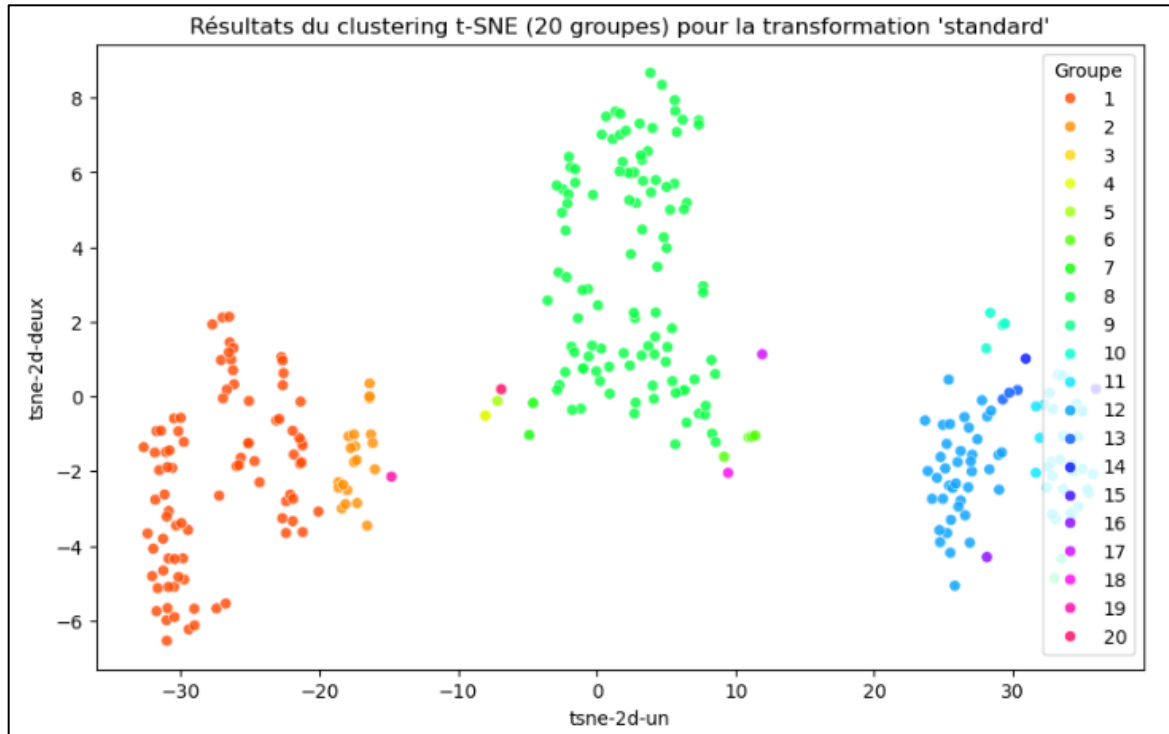


Figure 4-29 - Groupes de la transformation finale standard

L'analyse des groupes (Figure 4-29) issus de la transformation standard, illustrée par le graphique t-SNE, montre une séparation claire et une cohésion interne forte entre les périodes horaires. Cette segmentation permet d'identifier des groupes distincts d'heures avec des comportements similaires, facilitant ainsi la planification et l'optimisation des horaires. Par exemple, les groupes bien définis peuvent révéler des périodes de pointe, des moments de faible activité ou des segments horaires nécessitant des ressources spécifiques. En utilisant ces informations, les gestionnaires peuvent allouer les ressources de manière plus efficace, ajuster les stratégies opérationnelles et améliorer la répartition des charges de travail tout au long de la journée. Cette approche permet de répondre aux besoins spécifiques de chaque période horaire, optimisant ainsi la performance globale et l'efficacité des opérations.

Analyse du groupement final

La carte de chaleur des corrélations pour le Groupe 1 (

Figure 4-30) offre des enseignements précieux sur les similarités et les différences entre les différentes heures de la journée. En utilisant ces informations, il est possible d'améliorer la segmentation horaire en regroupant les périodes homogènes et en différenciant les périodes avec des comportements distincts. Par exemple, les heures de l'après-midi (13h à 15h) forment un bloc avec des corrélations positives élevées. Les heures de nuit (23h) montrent également des corrélations négatives avec les heures du matin (par exemple, 8h à 9h).

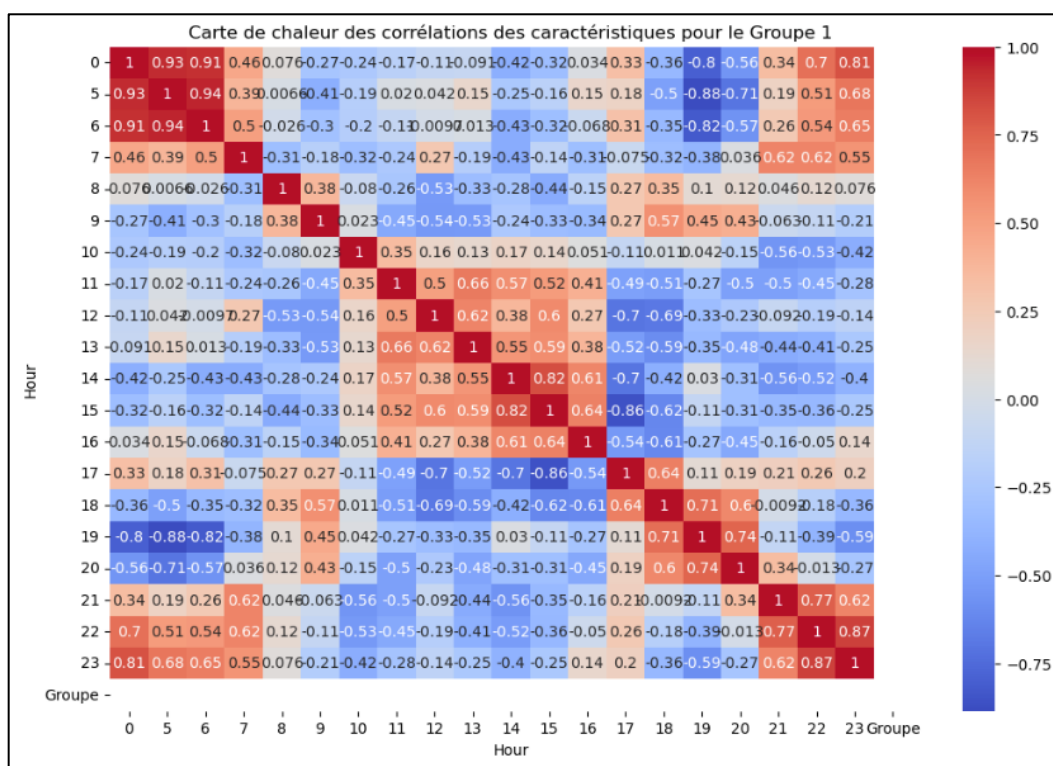


Figure 4-30 - Carte des corrélations du groupe 1 de la segmentation hiérarchique - transformation standard

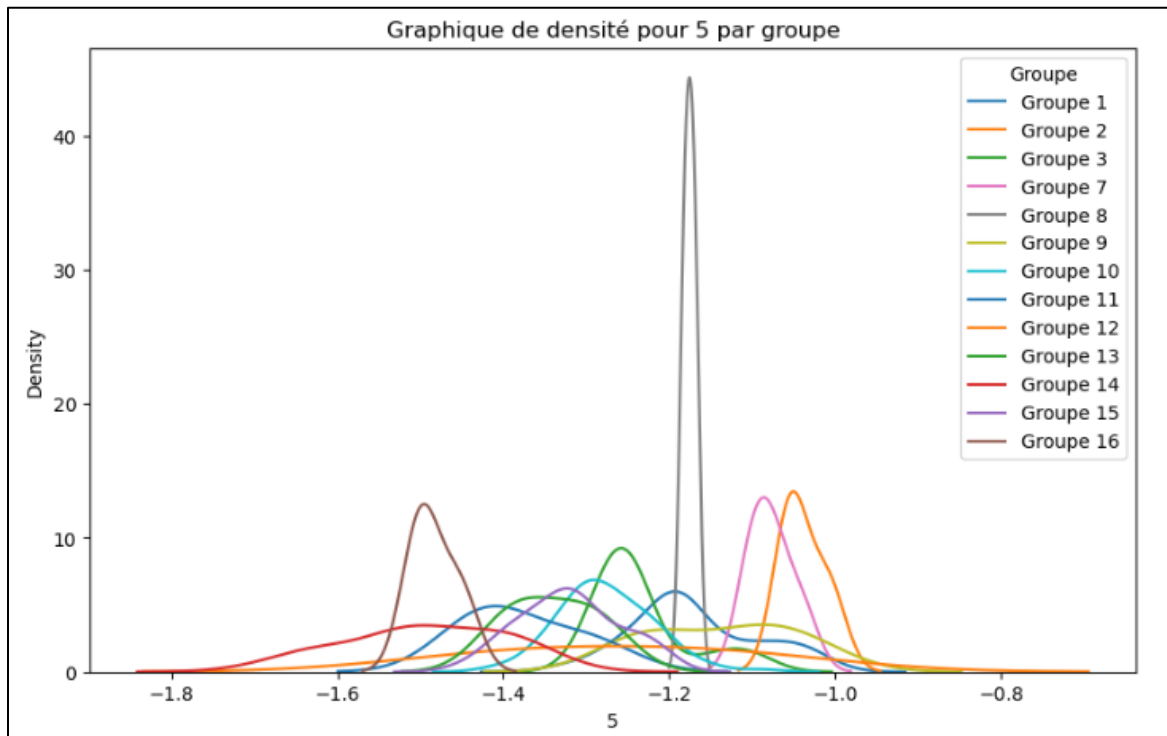


Figure 4-31 - Densité à 5 heures du matin pour la segmentation hiérarchique, transformation standard

La figure ci-dessus (Figure 4-31) de densité montre la distribution des valeurs pour différentes caractéristiques au sein de plusieurs groupes. Une observation clé est que le groupe 7 (comme 12) présente une densité très élevée concentrée autour de la valeur -1.2, indiquant une homogénéité élevée et une forte concentration des valeurs dans ce groupe. En revanche, les groupes 1, 2, 6, 8, 9, 10, 11, et 13 montrent des distributions plus larges et moins concentrées, suggérant une plus grande variabilité interne.

Ces résultats ont des implications importantes pour la segmentation des horaires. Le groupe 7, avec sa haute densité et sa faible variabilité, pourrait représenter des périodes horaires très spécifiques où les comportements ou les activités sont très homogènes, comme les heures de pointe ou les périodes de forte activité. En identifiant ces périodes, il devient possible de planifier les ressources de manière plus rigide et structurée, répondant précisément aux besoins uniformes de ces heures.

En revanche, les groupes avec plus de variabilité nécessitent une gestion plus flexible. Les périodes horaires représentées par ces groupes pourraient nécessiter une allocation de ressources adaptative et diversifiée pour gérer efficacement la diversité des activités ou des comportements. Cela pourrait

inclure des stratégies telles que l'augmentation de la flexibilité des horaires de travail ou la diversification des ressources disponibles.

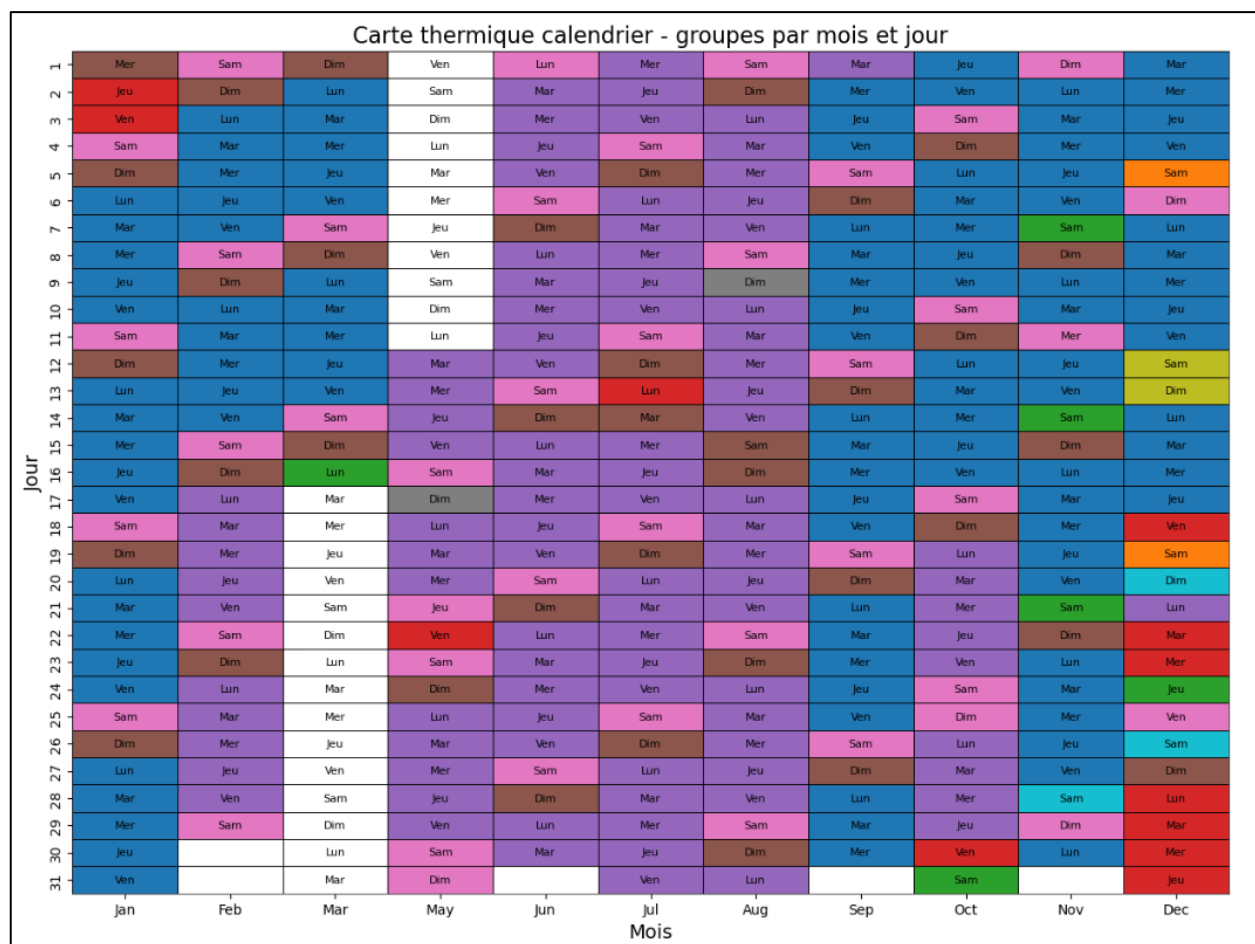


Figure 4-32 - Calendrier résultat de la segmentation hiérarchique

La carte thermique (Figure 4-32) calendaire révèle des motifs distincts de groupes par jour et par mois, illustrant des variations dans les comportements horaires tout au long de l'année. Les jours et mois avec des groupes homogènes indiquent des périodes stables où les activités sont similaires, permettant une planification cohérente et standardisée des horaires. En revanche, les périodes avec une grande variabilité de groupes nécessitent une gestion plus flexible pour s'adapter aux changements fréquents. Les fins de semaine, par exemple, montrent souvent des groupes distincts, reflétant des comportements différents des jours de semaine. De plus, les variations saisonnières évidentes, comme la transition de l'été à l'automne, suggèrent la nécessité d'ajuster les stratégies de planification en fonction des changements de comportement au fil de l'année.

Analyse statistique approfondie des résultats du groupement final

L'analyse des statistiques du Groupe 8 révèle des périodes distinctes de forte et faible activité tout au long de la journée, avec des implications importantes pour la gestion des ressources et la planification horaire. Les heures de la nuit, de 0h à 5h, montrent une activité relativement faible et une faible variabilité, permettant une planification stable et prévisible. En revanche, l'activité augmente significativement à partir de 6h, atteignant un pic autour de 17h, avec une grande variabilité pendant cette période de forte activité. Cela nécessite une gestion flexible et l'allocation de ressources supplémentaires pour s'adapter aux fluctuations. Après 18h, l'activité diminue, avec des niveaux plus bas et une variabilité modérée à élevée pendant les heures de soirée, permettant une optimisation des ressources. La répartition uniforme des observations sur les jours du mois et les mois de l'année permet une planification saisonnière cohérente, avec des ajustements mineurs en fonction des variations mensuelles pour optimiser les opérations tout au long de l'année.

4.2.3 Comparaison des résultats du K-moyennes et de la segmentation hiérarchique

La comparaison des cartes thermiques calendaires des groupes obtenus par le K-moyennes et la segmentation hiérarchique révèle des différences significatives dans la répartition des groupes au cours de l'année. Le K-moyennes tend à créer des groupes de tailles similaires, montrant des variations distinctes entre les jours de la semaine et les mois de l'année, avec des regroupements clairs durant les fins de semaine et certaines périodes de l'année. En revanche, la segmentation hiérarchique, qui peut générer des groupes de tailles variées, présente une distribution différente des mêmes jours et mois, capturant ainsi d'autres dimensions des données. Ces divergences suggèrent que chaque méthode met en évidence des aspects uniques des tendances d'activité, offrant des perspectives complémentaires. Pour la planification des horaires, intégrer les résultats des deux méthodes permet d'optimiser la gestion des ressources en tenant compte des variations saisonnières et hebdomadaires identifiées. Une approche combinée offre une vue plus robuste et complète des patrons d'activité, facilitant une planification plus efficace et informée.

4.2.4 La méthode DTW combiné au K-moyennes et segmentation hiérarchique

Les deux méthodes, DTW + K-moyennes et DTW + segmentation hiérarchique, offrent des perspectives distinctes pour la planification des horaires. La méthode DTW + K-moyennes, avec sa plus grande variabilité et flexibilité dans la répartition des groupes, est idéale pour des environnements dynamiques nécessitant des ajustements fréquents des horaires en fonction des changements rapides des tendances d'activité. En revanche, la méthode DTW + segmentation hiérarchique, caractérisée par une stabilité et une homogénéité des groupes sur de longues périodes, convient mieux à des environnements stables où la planification à long terme est essentielle. Cette méthode permet une gestion prévisible avec moins de réajustements fréquents. Ainsi, le choix de la méthode dépendra des besoins spécifiques de l'organisation, en privilégiant DTW + K-moyennes pour une approche réactive et flexible, et DTW + segmentation hiérarchique pour une planification stable et durable (

Figure 4-33 et Figure 4-34).

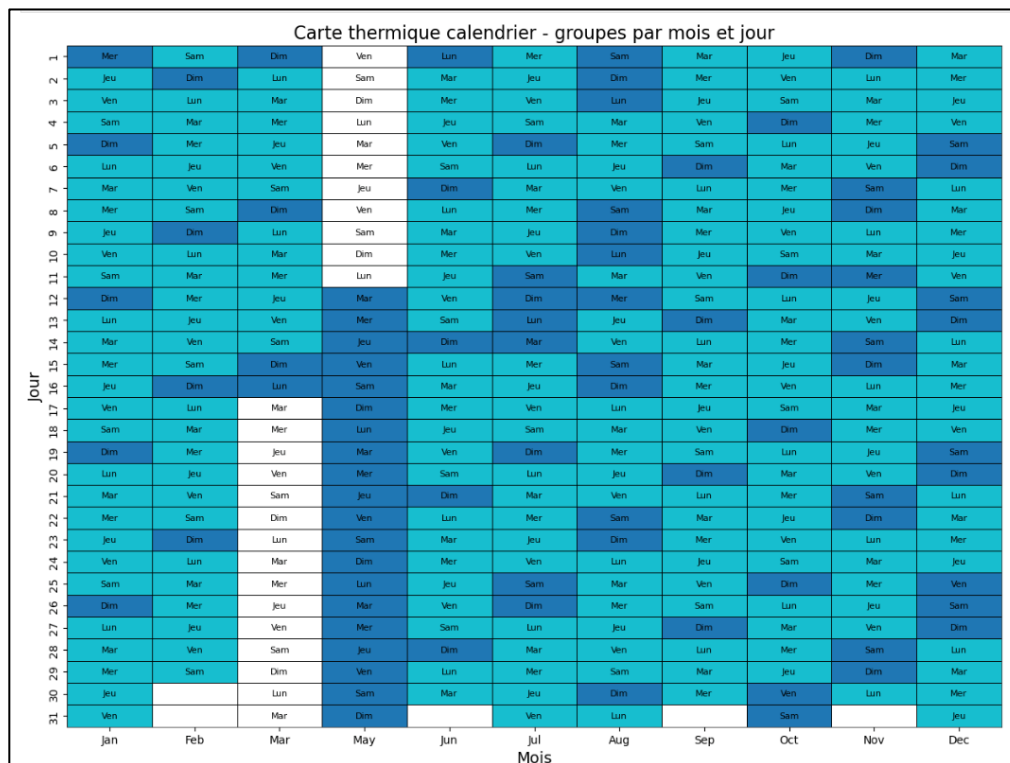


Figure 4-33 - DTW combiné au K-moyennes



Figure 4-34 - DTW combiné à la segmentation hiérarchique

4.2.5 Agrégation demi-horaire

L'agrégation des données de validation des passagers à intervalles de 30 minutes ou vecteur "demi-heure", permet une analyse plus granulaire des tendances d'utilisation du réseau de transport. En divisant chaque heure en deux segments de 30 minutes, il est possible de détecter des variations plus fines dans les flux de passagers, ce qui offre une vue détaillée des périodes de pointe et des creux d'activité. Cette méthode permet également de comparer les comportements des passagers de manière plus précise au cours de la journée, ce qui est essentiel pour une planification optimale des ressources et une gestion efficace du réseau de transport (Tableau 4-7).

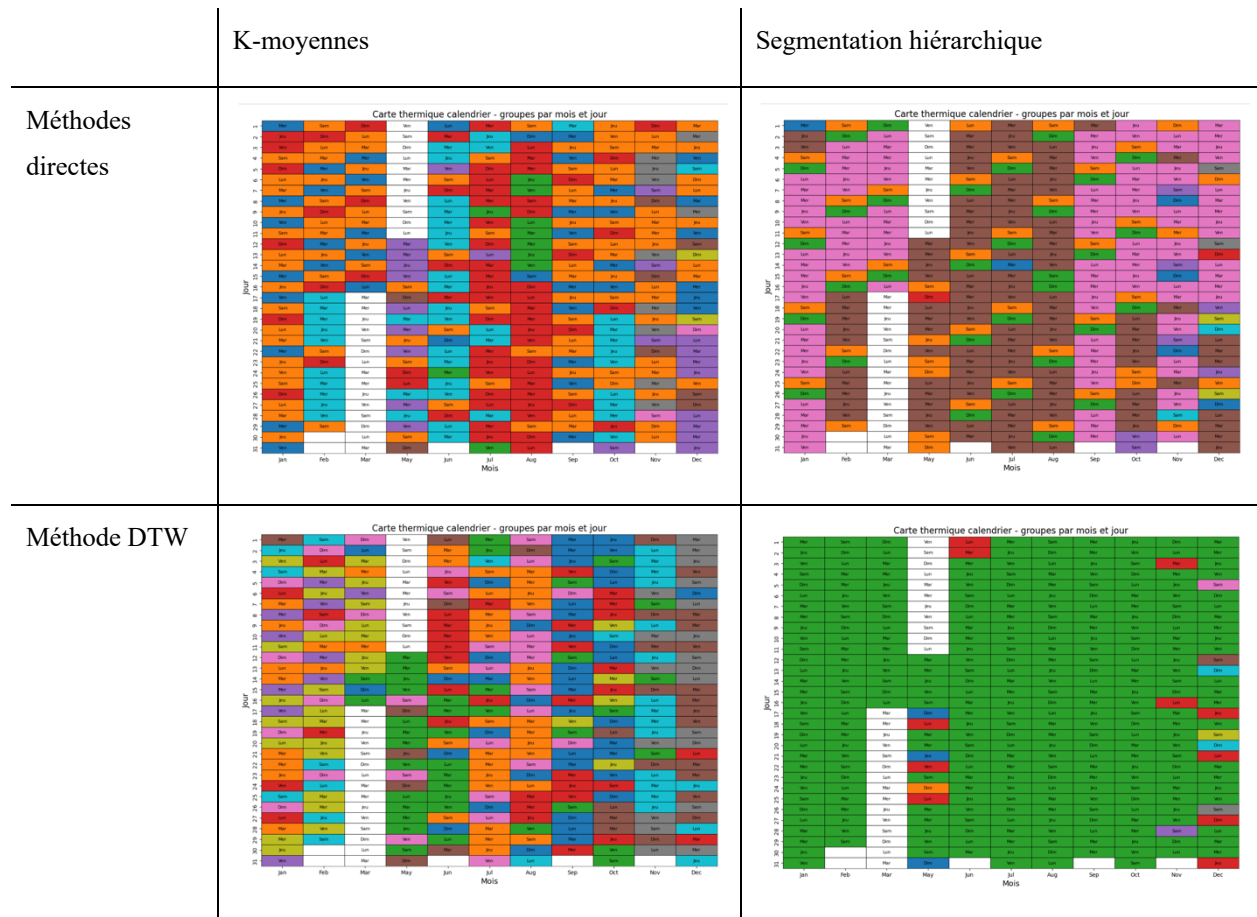
Tableau 4-7 - Vecteur "demi-heure"

demi-heure	00:00 00:30	05:00 05:30	05:30 06:00	06:00 06:30	06:30 07:00	07:00 07:30	07:30 08:00	08:00 08:30	08:30 09:00	09:00 09:30	09:30 10:00	10:00 10:30	10:30 11:00	11:00 11:30	11:30 12:00	12:00 12:30	12:30 13:00	13:00 13:30	13:30 14:00	14:00 14:30	14:30 15:00
Date																					
2020-01-01	110.0	46.0	50.0	34.0	50.0	80.0	138.0	168.0	150.0	228.0	210.0	310.0	310.0	440.0	458.0	538.0	562.0	484.0	612.0	670.0	898.0
2020-01-02	36.0	262.0	484.0	756.0	1172.0	1986.0	2586.0	3134.0	3228.0	2608.0	2770.0	2746.0	3130.0	3286.0	3706.0	3792.0	3514.0	3508.0	4470.0	5216.0	5736.0
2020-01-03	52.0	294.0	558.0	760.0	1280.0	2042.0	2844.0	3634.0	3314.0	2874.0	3110.0	3216.0	3798.0	4066.0	3966.0	4438.0	4112.0	4212.0	5060.0	5846.0	5822.0
2020-01-04	144.0	194.0	308.0	306.0	388.0	568.0	820.0	1194.0	1476.0	2012.0	2642.0	2850.0	3158.0	3504.0	3508.0	3546.0	3420.0	3256.0	4356.0	5172.0	5348.0
2020-01-05	114.0	16.0	52.0	8.0	80.0	160.0	242.0	350.0	536.0	606.0	846.0	840.0	810.0	1100.0	1190.0	962.0	1046.0	1214.0	1290.0	1244.0	1402.0

Résultats de l'agrégation demi-horaire

L'analyse des groupes avec différentes méthodes d'agrégation demi-horaire montre des motifs saisonniers et hebdomadaires distincts. K-moyennes et DTW combiné avec K-moyennes révèlent des variabilités accrues en été et pendant les vacances, ainsi que des pics d'activité les jours de semaine, nécessitant une augmentation des ressources. La segmentation hiérarchique et DTW combinés avec la segmentation hiérarchique indiquent une forte homogénéité avec une répartition stable des groupes, suggérant une allocation cohérente des ressources, avec des ajustements ponctuels pour les périodes spécifiques (Tableau 4-8).

Tableau 4-8 - Résultats des 4 méthodes



Comparaison des résultats de l'agrégation horaire et demi-horaire

L'agrégation demi-horaire, par rapport à l'agrégation horaire, offre une granularité supérieure qui permet de capturer des variations plus fines dans les comportements de validation des usagers. Cette approche révèle des motifs saisonniers et hebdomadaires plus détaillés, facilitant une identification précise des périodes de pointe et de faible activité. Les méthodes de segmentation appliquées à cette agrégation montrent des différences marquées dans la répartition des groupes, avec une segmentation plus granulaire qui optimise la gestion des ressources. En somme, l'agrégation demi-horaire permet une planification plus réactive et efficace, ajustant les ressources de manière précise pour répondre aux variations dynamiques des habitudes de déplacement.

4.3 Segmentation sur le trimestre d'automne 2020

Le choix du trimestre d'automne 2020 pour cette analyse de segmentation s'explique par plusieurs raisons. Premièrement, l'automne offre une période de transition entre les saisons estivales et hivernales, ce qui permet d'observer des variations saisonnières significatives dans les données de validation. Deuxièmement, cette période inclut plusieurs jours fériés et variations de comportement des utilisateurs, fournissant ainsi un ensemble de données diversifié et représentatif. Enfin, l'année 2020 a été marquée par des événements exceptionnels tels que la pandémie de COVID-19, qui ont influencé les schémas de mobilité. Analyser les données de l'automne 2020 permet donc de capturer ces dynamiques particulières et d'étudier les variations horaires de manière plus précise, car les horaires de validation peuvent révéler des tendances et des comportements spécifiques qui sont cruciaux pour une planification efficace et une meilleure compréhension de l'utilisation des réseaux de transport en période de changement.

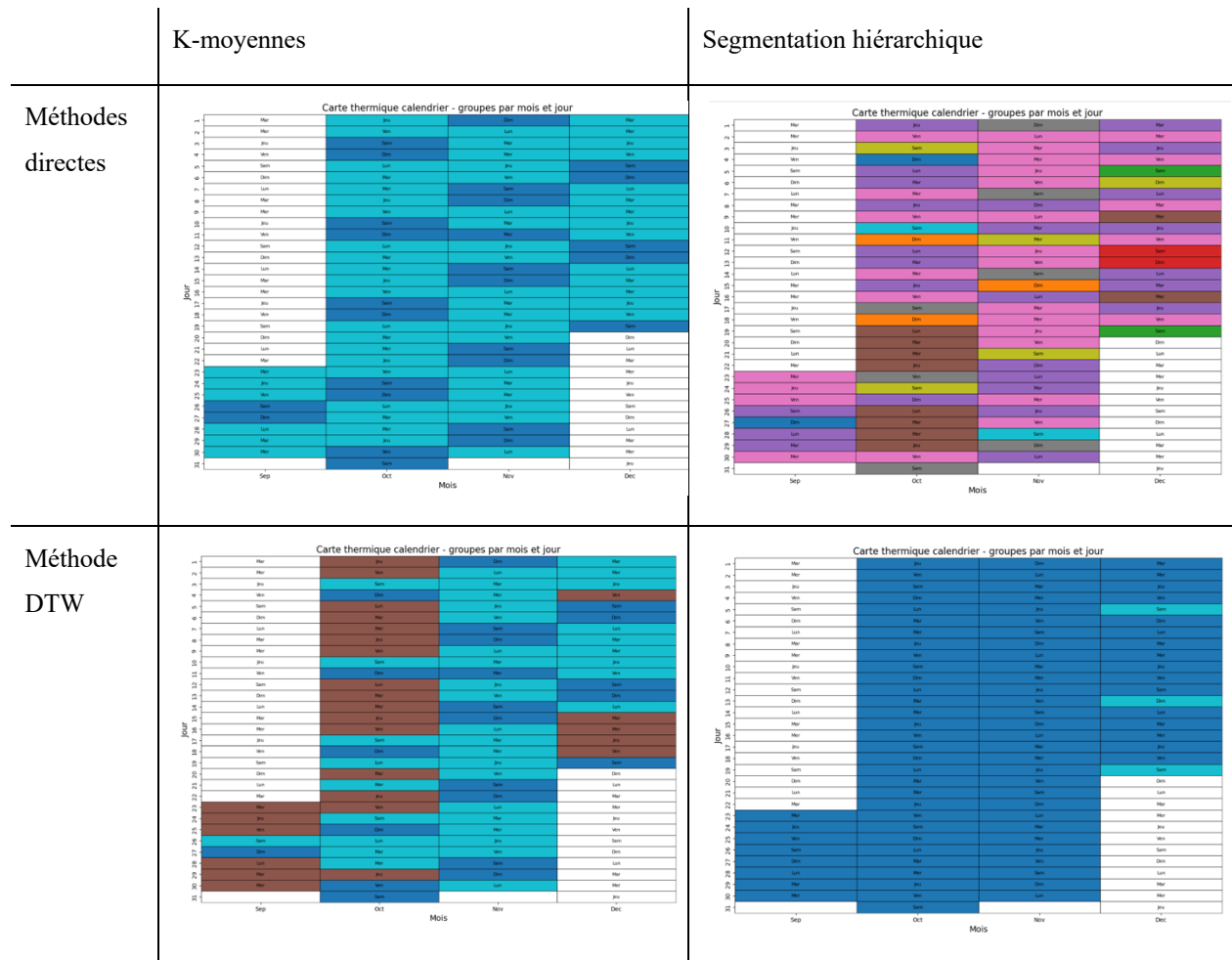
Pour les sections suivantes de ce chapitre, nous omettrons de présenter les résultats détaillés de l'analyse et ne présenterons que les résultats des groupements finaux sélectionnés, sachant que toutes les étapes présentées précédemment ont été suivies. Le cas échéant, nous commenterons sur les forces et faiblesses des différentes approches en relation avec les données utilisées.

4.3.1 L'ensemble du réseau

Tableau 4-9 - Validations par demi-heure pour le réseau, trimestre d'automne 2020

Carte	Vehicule	sens	courses	Code_Produit	Type_Transaction	Ligne	Code_arret	Lib_arret	Datetime	Nb_Passager	Date	Hour	demi-heure	Season
0	39	0	1466	1	Compostage Magnétique	Ligne A	TAULNA1	Les Aulnaies	2020-10-10 19:00:00	1	2020-10-10	19	19:00 - 19:30	Autumn
0	39	0	1466	1	Compostage Magnétique	Ligne A	TBOLIE1	Boliere	2020-10-12 09:00:00	1	2020-10-12	9	09:00 - 09:30	Autumn
0	39	0	1466	1	Compostage Magnétique	Ligne A	TCSMAR1	Saint-Marceau	2020-10-06 09:30:00	1	2020-10-06	9	09:30 - 10:00	Autumn
0	39	0	1466	1	Compostage Magnétique	Ligne A	TINDIE1	L'Indien	2020-10-21 15:15:00	1	2020-10-21	15	15:00 - 15:30	Autumn
0	39	0	1466	1	Compostage Magnétique	Ligne A	TINDIE1	L'Indien	2020-10-22 13:15:00	1	2020-10-22	13	13:00 - 13:30	Autumn

Tableau 4-10 - Résultats des 4 méthodes - réseau, automne



La méthode K-moyennes a montré des performances moindres en termes de cohésion et de séparation. La segmentation hiérarchique a présenté une bonne compacité et séparation avec un Indice de Davies-Bouldin de 0.256 et un Dunn Index de 1.912. La méthode K-moyennes avec

DTW a offert une meilleure cohésion et séparation que le K-moyennes classique, mais a été surpassée par les méthodes hiérarchiques. Enfin, la méthode DTW avec Segmentation hiérarchique s'est révélée la plus efficace, affichant un Score de silhouette de 0.911, une corrélation cophénétique de 0.942, et un bon équilibre entre les indices Davies-Bouldin et Calinski-Harabasz. En conclusion, DTW avec Segmentation hiérarchique est la méthode la plus performante pour segmenter les validations horaires sur cette période (Tableau 4-10).

4.3.2 Par mode

Une segmentation par mode après l'analyse réseau est crucial car elle permet de capter les comportements spécifiques et les variations des flux de passagers. En identifiant les modes principaux de déplacement, cette méthode isole des périodes de forte affluence ou des motifs de déplacement particuliers qui peuvent être masqués dans une analyse globale. Cela fournit une base solide pour ajuster les horaires de manière plus précise et pertinente, assurant ainsi une meilleure répartition des ressources et une amélioration de l'expérience des usagers.

Tramways

Tableau 4-11 - Résultats des 4 méthodes - tramways, automne

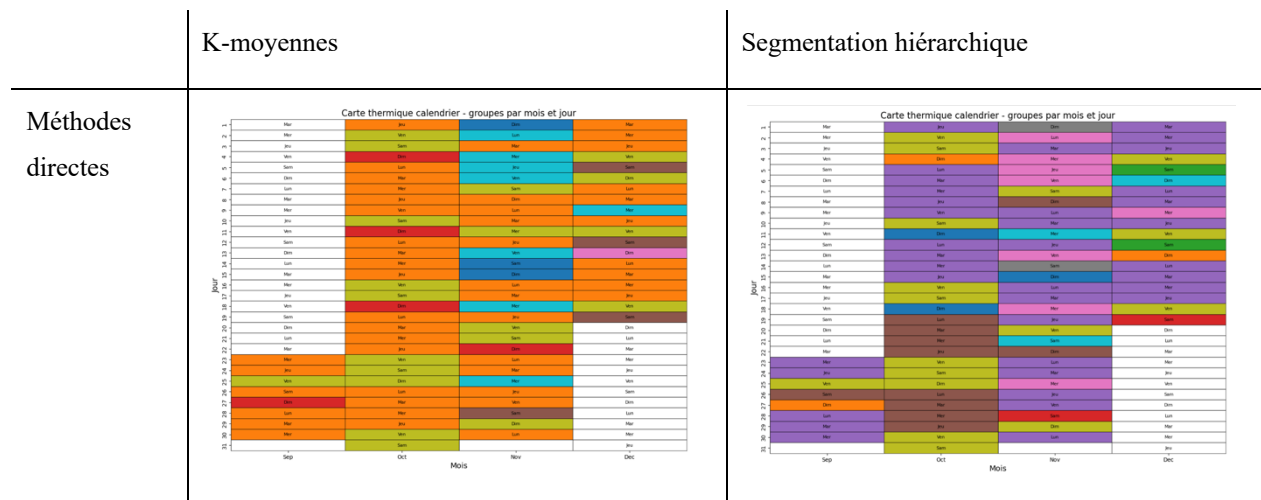
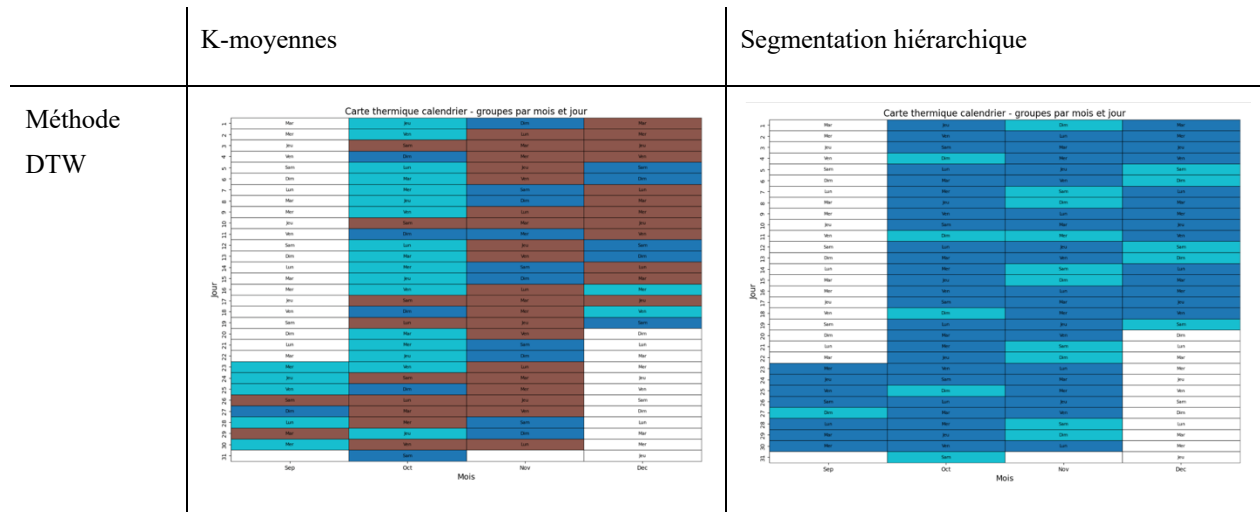


Tableau 4-11 - Résultats des 4 méthodes - tramways, automne (suite et fin)



Les quatre méthodes (Tableau 4-11) de segmentation appliquée aux données des tramways pour l'automne 2020 révèlent des caractéristiques distinctes et des performances variées, fournissant des informations cruciales pour l'ajustement des horaires. La méthode K-moyennes sans transformation produit trois groupes avec une cohésion élevée (245191.95) et une bonne séparation (571097.75), indiquant des groupes bien distincts. Cette approche pourrait être utile pour identifier des périodes de la journée avec des similarités dans les flux de passagers, facilitant ainsi l'optimisation des fréquences des tramways.

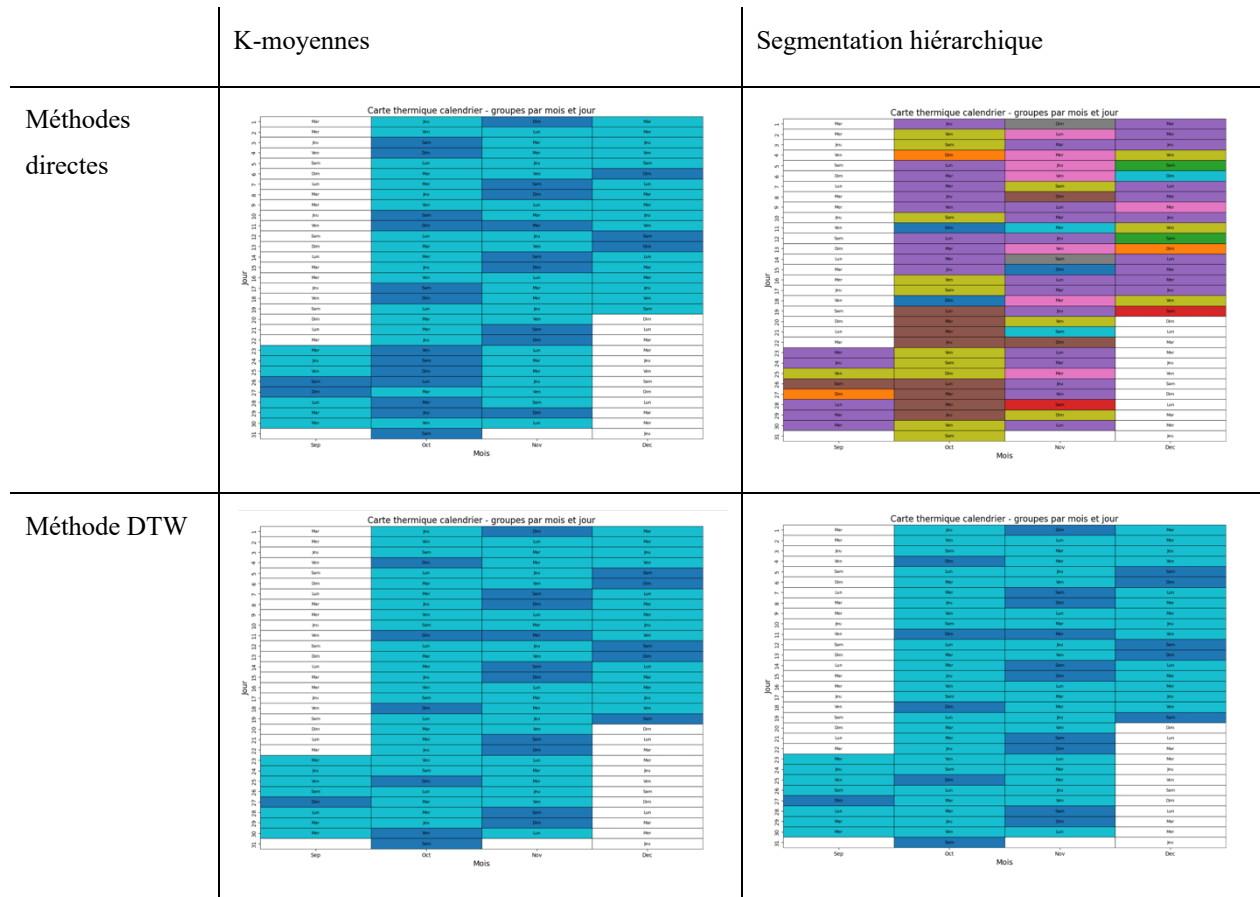
La segmentation hiérarchique avec transformation quantile normal identifie 19 groupes optimaux, avec une excellente séparation (338.14) et cohésion (62.47), ainsi qu'un score de silhouette de 0.470, démontrant des groupes bien formés. Cette méthode permet une analyse plus granulaire des flux, idéale pour ajuster les horaires en fonction des variations subtiles et des pics d'achalandage.

La combinaison DTW + K-moyennes, après transformation Box-Cox, génère trois groupes, présentant une bonne séparation (194186.88) et une cohésion raisonnable (79326.46). Cette technique est particulièrement efficace pour capturer des similarités temporelles complexes, aidant à aligner les horaires de tramway sur des motifs de déplacement récurrents mais non linéaires.

Enfin, la méthode DTW + segmentation hiérarchique, avec transformation racine carrée, forme deux groupes avec un score de silhouette élevé (0.743), une bonne cohésion (82683.05) et une corrélation cophénétique satisfaisante (0.838), indiquant des groupes bien définis et cohérents. Cette approche permet de distinguer clairement les périodes de forte et de faible affluence, simplifiant la mise en place d'horaires différenciés selon les besoins de la journée.

Bus

Tableau 4-12 - Résultats des 4 méthodes - bus, automne



Les résultats du Tableau 4-12 montrent que K-moyennes avec DTW et la segmentation hiérarchique avec DTW produisent des groupes bien définis, avec de meilleures séparations et cohésions par rapport aux méthodes sans DTW. En comparaison avec les tramways, où les groupes étaient également bien définis avec les mêmes méthodes, les bus montrent une variabilité temporelle plus prononcée, justifiant l'utilisation de DTW pour capturer ces variations.

Les résultats montrent que la segmentation hiérarchique fournit un plus grand nombre de groupes en raison de son approche progressive. Elle commence par traiter chaque point de données comme un groupe distinct, puis fusionne les points et groupes successivement pour former des groupes plus grands, ce qui permet une analyse plus fine et la détection de sous-groupes distincts. Cette méthode révèle plus de variations et de comportements spécifiques. Contrairement à K-moyennes, qui impose un nombre fixe de groupes dès le départ, la segmentation hiérarchique s'adapte mieux à la structure réelle des données, offrant une vision plus détaillée et nuancée des schémas. Cette

capacité à créer des groupes plus nombreux et plus petits est particulièrement utile pour identifier des modèles saisonniers et horaires, fournissant des éléments précieux pour l'établissement des horaires.

4.4 Segmentation par ligne

Il s'agit de réaliser des analyses temporelles séparées pour chaque ligne de bus et de tram et de comprendre les patrons d'achalandage et les cycles de demande propres à chaque ligne. Cette différenciation est importante car les bus et les trams desservent souvent des itinéraires différents, ont des capacités variées, et sont soumis à des contraintes opérationnelles et des comportements de passagers distincts. En identifiant les heures de pointe spécifiques à chaque ligne, on peut optimiser les horaires pour mieux répondre à la demande des passagers.

4.4.1 Lignes de tram A et B

Pour la ligne de tram A, en utilisant les K-moyennes avec la transformation Box-Cox, nous avons identifié deux groupes principaux avec un WCSS de 50,703,924.13 et un indice de Davies-Bouldin de 0.443. Le premier groupe, représentant les périodes de forte affluence, affiche une moyenne de 2,115.79 passagers et une distribution étendue, tandis que le second groupe, correspondant aux périodes de faible utilisation, montre une moyenne de 378.24 passagers. L'analyse hiérarchique a proposé un K optimal de 30, suggérant une sur-segmentation, avec des indices de performance inférieurs, notamment un score de silhouette de 0.233. En combinant K-moyennes avec DTW, les résultats sont restés cohérents, confirmant la robustesse des groupes identifiés.

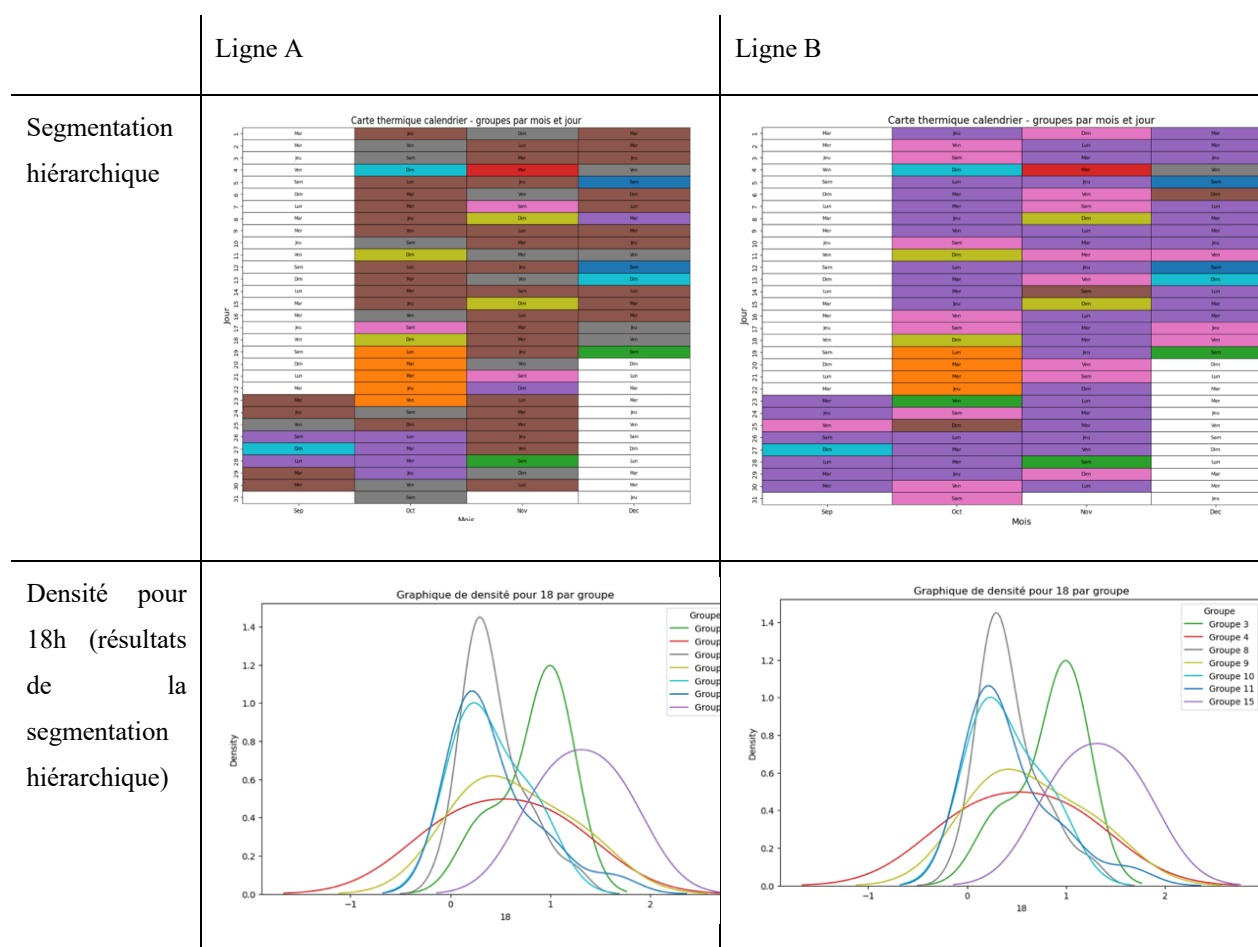
Les résultats de la ligne B ont montré des différences notables par rapport à la ligne A. En utilisant K-moyennes avec la transformation inverse, trois groupes principaux ont été identifiés, avec un WCSS de 37.97 et un indice de Davies-Bouldin de 0.926. Le premier groupe, correspondant aux périodes de forte utilisation, présente une moyenne de 1,241.11 passagers. Le second groupe, avec une moyenne de 169.18 passagers, et le troisième groupe, avec 561.74 passagers, représentent respectivement les périodes de faible et d'utilisation intermédiaire. L'analyse hiérarchique a également suggéré une sur-segmentation avec un K optimal de 28 et un score de silhouette de 0.269. Cependant, l'intégration de DTW dans la segmentation hiérarchique a amélioré les indices de performance, particulièrement la cohésion et la séparation, indiquant une meilleure définition des groupes.

En comparant les résultats des lignes de tram A et B, il est évident que les schémas d'utilisation et les besoins de service varient considérablement entre les deux lignes. La ligne A, avec deux groupes distincts, montre une différenciation plus nette entre les périodes de forte et de faible utilisation. En revanche, la ligne B présente une segmentation plus complexe avec trois groupes, indiquant des variations plus subtiles dans les schémas de demande. Les méthodes de transformation et les techniques de segmentation utilisées ont prouvé leur efficacité dans la détection des patterns de trafic, bien que la sur-segmentation observée dans l'analyse hiérarchique suggère la nécessité de raffiner les approches pour certains contextes.

Dans le tableau ci-dessous, l'analyse de densité à 18 heures pour la ligne de tram A montre une répartition distincte des passagers entre plusieurs groupes. Les différentes courbes de densité illustrent la concentration des passagers pour chaque groupe à cette heure spécifique. On observe que certains groupes, comme le Groupe 3, affichent des pics de densité élevés, indiquant des périodes de forte affluence. D'autres groupes, comme le Groupe 17, présentent des densités plus basses, suggérant une utilisation plus dispersée. Cette distribution hétérogène reflète des variations dans les comportements de déplacement des usagers à 18 heures, un moment crucial de la journée correspondant souvent à la fin des horaires de travail et au début des déplacements domicile-travail.

Pour la ligne de tram B, le graphique de densité à 18 heures révèle également des variations significatives entre les groupes. Par exemple, le Groupe 15 montre un pic de densité prononcé, indiquant une haute fréquentation à cette heure, tandis que d'autres groupes comme le Groupe 6 présentent des densités plus uniformes. Cette analyse met en évidence des différences dans les schémas de déplacement des passagers, reflétant potentiellement des facteurs tels que la localisation des arrêts, la proximité de zones résidentielles ou commerciales, et la fréquence des services.

Tableau 4-13 - Résultats Tram A et B



Les analyses de densité à 18 heures pour les lignes de tram A et B révèlent des schémas d'utilisation spécifiques qui ont des implications importantes pour la gestion des services de transport. La présence de pics de densité élevés dans certains groupes suggère des périodes de forte demande, nécessitant une augmentation de la fréquence des services ou le déploiement de véhicules supplémentaires pour éviter la surcharge et améliorer le confort des passagers. À l'inverse, les groupes présentant des densités plus uniformes ou faibles indiquent des périodes où les ressources peuvent être redistribuées.

4.5 Segmentation par arrêt

La segmentation par arrêt est cruciale pour comprendre les dynamiques spécifiques de chaque point du réseau de tramway. Chaque arrêt peut avoir des schémas d'utilisation différents en fonction de sa localisation, des activités environnantes, ou des horaires spécifiques de la journée. En regroupant les arrêts avec des profils de validation similaires, il devient possible d'identifier des tendances

générales, telles que les heures de pointe ou les périodes de faible affluence. Cela permet non seulement d'optimiser les fréquences des trams ou bus en fonction de la demande réelle, mais aussi de mieux planifier les opérations, de distribuer les ressources de manière plus efficace et de réagir aux besoins des usagers avec une plus grande précision.

4.5.1 Les arrêts de tram

L'analyse par K-moyennes a révélé deux principaux groupes parmi les arrêts les plus fréquentés du tramway d'Orléans, indiquant deux schémas distincts d'utilisation des services. Le premier groupe montre des pics de validation importants durant les heures de pointe, notamment le matin et en fin de soirée, tandis que le second groupe présente une distribution plus homogène des validations tout au long de la journée. La segmentation hiérarchique a également identifié deux groupes principaux, avec un troisième groupe plus petit montrant des comportements de validation plus extrêmes. Le Groupe 1, similaire à celui identifié par K-moyennes, regroupe des périodes avec des validations stables, tandis que le Groupe 2 montre des périodes de validation très faibles ou même nulles, suggérant des heures creuses. Le Groupe 3, bien que moins représenté, montre des variations importantes autour de certaines heures, potentiellement liées à des événements spécifiques ou des perturbations du service.

Les deux méthodes de segmentation, K-moyennes et hiérarchique, ont révélé des schémas similaires, en particulier dans l'identification de deux principaux groupes, ce qui renforce la robustesse des conclusions. Cependant, la segmentation hiérarchique a permis de mettre en lumière un troisième groupe avec des comportements plus extrêmes, ce qui pourrait indiquer des besoins spécifiques en termes de gestion des horaires. Cette différence suggère que si K-moyennes est efficace pour une première segmentation des horaires, la segmentation hiérarchique peut offrir des connaissances supplémentaires pour des ajustements plus fins, notamment en ce qui concerne les cas atypiques.

L'analyse réalisée avec l'approche K-moyennes combinée avec la méthode DTW a révélé que les transformations telles que quantile normal, Yeo Johnson et quantile uniforme produisent les meilleurs résultats. Ces transformations ont permis d'obtenir des scores de silhouette élevés, indiquant une bonne séparation entre les groupes. La configuration finale a montré deux groupes principaux, ce qui suggère une dichotomie marquée dans les comportements des usagers aux différentes heures de la journée. Le meilleur modèle, utilisant la transformation quantile normal, a

présenté un WCSS (« *Within-Cluster Sum of Squares* ») faible, ce qui indique une forte cohésion au sein des groupes. Contrairement à l'approche K-moyennes, la segmentation hiérarchique a identifié jusqu'à 16 groupes distincts avec la transformation « différentiation », mettant en évidence une complexité beaucoup plus grande dans la variabilité des comportements horaires des usagers. Cette approche a également permis d'atteindre un score de silhouette exceptionnel (0.9859) et une corrélation cophénétique élevée, soulignant une bonne représentation des distances originales dans l'arbre hiérarchique.

4.6 Synthèse des résultats obtenus des différentes approches

Les méthodes employées dans le cadre de ce travail combinent des techniques de segmentation et de transformation des données pour l'analyse des flux de passagers du réseau d'Orléans durant l'année 2020. Le K-moyennes segmente les jours selon des profils de validation similaires, permettant de distinguer efficacement les différentes typologies de jours, telles que les jours de semaine et les fins de semaine. La segmentation hiérarchique permet de détecter les relations hiérarchiques entre les groupes identifiés. La méthode DTW, en ajustant les différences temporelles entre les séries de données, améliore la précision des groupes, en particulier dans le contexte de variations horaires. Enfin, les transformations par quantile, normalisent les distributions et atténuent l'impact des valeurs aberrantes, garantissant ainsi une analyse robuste fournissant ainsi un cadre analytique pour l'ajustement potentiel des horaires en fonction des besoins identifiés.

4.6.1 Comparaison des résultats des quatre algorithmes

Les cartes thermiques calendaires des groupes obtenus révèlent des différences significatives dans la structure et la stabilité des groupes tout au long de l'année. Les méthodes K-moyennes et segmentation hiérarchique montrent une répartition plus régulière des groupes, bien que K-moyennes se distingue par une plus grande sensibilité aux changements à court terme. En revanche, l'approche DTW + K-moyennes démontre une flexibilité accrue, capable de s'adapter à des variations temporelles non linéaires, ce qui souligne sa pertinence élevée dans des environnements dynamiques.

Par ailleurs, l'association DTW + segmentation hiérarchique produit des groupes plus stables et homogènes, offrant un avantage notable pour la planification à long terme. Dans le cadre de la planification des horaires, DTW + K-moyennes est techniquement recommandé en raison de sa

capacité à traiter des séquences temporelles complexes et à effectuer des ajustements fréquents. À l'inverse, DTW + segmentation hiérarchique est privilégié pour des analyses nécessitant une stabilité à long terme et une compréhension hiérarchique des données.

Enfin, les méthodes classiques telles que K-moyennes et segmentation hiérarchique conservent leur utilité pour des analyses moins complexes. Toutefois, leur capacité à prendre en compte des variations temporelles reste limitée.

CHAPITRE 5 CONCLUSION

5.1 Contributions

Ce travail a permis d'explorer l'utilisation des techniques de segmentation et de transformation des données pour ajuster les horaires en transport collectif, en particulier dans un contexte post-pandémique où les habitudes de déplacement ont été profondément modifiées. En utilisant des méthodes avancées de segmentation, telles que la segmentation hiérarchique utilisant la distance DTW (Dynamic Time Warping), le travail a pu identifier avec précision les périodes de forte et de faible affluence. En effet, l'identification des groupes homogènes de passagers selon les heures de la journée offre la possibilité de moduler les ressources en fonction des moments critiques, optimisant ainsi l'usage des véhicules et réduisant les temps d'attente. D'un point de vue méthodologique, le travail a exploré plusieurs niveaux de granularité du réseau et fenêtres temporels combinées à différentes techniques de segmentation et de transformation permettant une comparaison des résultats selon les méthodes.

D'un point de vue méthodologique, le travail s'est distingué par l'exploration de plusieurs niveaux de granularité, tant au niveau du réseau que des fenêtres temporelles, combinée à une variété de techniques de segmentation et de transformation des données. Cette approche a permis une comparaison rigoureuse des résultats obtenus avec différentes méthodes, assurant une évaluation complète de leur efficacité. Les différentes techniques de transformation appliquées, telles que la standardisation et les transformations quantiles (normal et uniforme), ont démontré leur capacité à normaliser les données de manière efficace, facilitant ainsi leur analyse comparative.

Plusieurs critères d'évaluation ont été systématiquement appliqués aux groupes identifiés. À chaque étape, un score composite, sans a priori sur le poids des critères individuels, a été utilisé pour évaluer la pertinence des groupes, tester leur flexibilité, et mesurer leur sensibilité à une multitude de configurations. Cette démarche a permis d'attester de la robustesse des résultats obtenus. Les transformations de normalisation se sont révélées particulièrement performantes, et les méthodes de segmentation hiérarchique combinées au DTW ont été jugées plus adaptées pour capturer les variations les plus fines, indépendamment de l'objet d'étude considéré.

Les évaluations détaillées des groupes finaux offrent un niveau d'analyse encore plus approfondi, fournissant des indications précises à l'heure près sur les ajustements d'horaires à effectuer. Les diagrammes de densité des groupes et les matrices de corrélation élaborées dans ce cadre fournissent les informations nécessaires pour identifier les variations intra-groupes, déterminer leur intensité, et établir les rapprochements horaires à mettre en œuvre pour une journée ou un ensemble de jours spécifique. Ainsi, ce travail propose une méthode efficace pour moduler les ressources en fonction des moments critiques, optimisant l'utilisation des véhicules tout en minimisant les temps d'attente pour les usagers.

Enfin, la méthodologie développée dans ce travail se distingue par son exhaustivité et sa granularité, ce qui lui confère une grande polyvalence. Elle pourrait être appliquée à d'autres réseaux de transport, quel que soit le contexte ou la qualité des données disponibles, grâce à sa capacité à s'adapter à différentes réalités locales et à ses exigences en termes de robustesse des analyses.

5.2 Limitations

Cependant, plusieurs limitations doivent être soulignées. Les analyses fondées sur des données historiques, bien qu'informatives, ne captent pas les fluctuations en temps réel des flux de passagers, ce qui pourrait limiter la réactivité des ajustements horaires en cas de perturbations imprévues. Cette limitation pourrait réduire l'efficacité des ajustements horaires dans des situations imprévues, telles que des perturbations majeures ou des événements non planifiés.

La complexité des transformations appliquées aux données, bien qu'efficace pour l'analyse, peut poser des défis en termes d'interprétation pour les décideurs non spécialisés. Cette complexité pourrait entraver l'implémentation opérationnelle des résultats, nécessitant ainsi une formation spécifique ou des outils d'interprétation simplifiés pour faciliter l'adoption des recommandations.

Enfin, l'extension de ces techniques à d'autres réseaux ou contextes urbains pourrait nécessiter des ajustements significatifs pour tenir compte des particularités locales. Cette nécessité d'adaptation pourrait limiter la généralisation des méthodes développées, réduisant leur applicabilité universelle sans ajustements supplémentaires.

5.3 Perspectives

Les perspectives de recherche future sont nombreuses. L'intégration de données en temps réel pourrait permettre d'ajuster les horaires de manière plus réactive, répondant ainsi aux variations quotidiennes et imprévues de la demande. Une autre perspective pourrait consister à intégrer des paramètres environnementaux, tels que les conditions météorologiques ou les grands événements publics, dans le processus de segmentation pour affiner encore davantage la précision des ajustements. De plus, la flexibilité des méthodes développées pourrait être testée sur d'autres réseaux ou contextes pour valider leur robustesse et leur adaptabilité.

En somme, bien que des défis subsistent, les contributions de ce travail offrent une base prometteuse pour l'amélioration des horaires dans les réseaux de transport collectif.

RÉFÉRENCES

- Abanda, A., Mori, U., & Lozano, J. A. (2019). A review on distance based time series classification. *Data Mining and Knowledge Discovery*, 33(2), 378-412
- Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes*, 39(3), 399-404
- Agard, B., Partovi Nia, V., & Trépanier, M. (2013). *Assessing public transport travel behaviour from smart card data with advanced data mining techniques*. World conference on transport research (vol. 13, p. 15-18).
- Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering—a decade review. *Information systems*, 53, 16-38.
- Ali, A., Kim, J., & Lee, S. (2016). Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering*, 20, 1532-1539.
- Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., & Hickman, M. (2018). Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 87, 123-137.
- Andreoni, A., & Postorino, M. N. (2006). Time series models to forecast air transport demand: a study about a regional airport. *IFAC Proceedings Volumes*, 39(12), 101-106.
- Arbex, R. O., & da Cunha, C. B. (2015). Efficient transit network design and frequencies setting multi-objective optimization by alternating objective genetic algorithm. *Transportation Research Part B: Methodological*, 81, 355-376.
- Badoe, D. A. (2007). Forecasting travel demand with alternatively structured models of trip frequency. *Transportation Planning and Technology*, 30(5), 455-475.
- Bagchi, M., & White, P. R. (2005). The potential of public transport smart card data. *Transport Policy*, 12(5), 464-474.
- Bagnall, A. J., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*. <https://doi.org/10.1007/S10618-016-0483-9>
- Banister, D. (2008). The sustainable mobility paradigm. *Transport Policy*. <https://doi.org/10.1016/J.TRANPOL.2007.10.005>
- Bhaskar, A., & Chung, E. (2014). Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, 16(3), 1537-1548.
- Bourbonnais, P.-L., Morency, C., Trépanier, M., & Martel-Poliquin, É. (2021). Transit network design using a genetic algorithm with integrated road network and disaggregated O-D demand data. *Transportation*, 48, 95-130.
- Cats, O. (2022). Identifying Human Mobility Patterns using Smart Card Data. *ArXiv*. <https://doi.org/10.48550/ARXIV.2208.05352>

- Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research Part A: General*, 18(5-6), 439-453.
- Ceder, A. (2015). Public Transit Planning and Operation: Modeling, Practice and Behavior. <https://doi.org/10.1201/B18689>
- Chen, C.-F., Chang, Y.-H., & Chang, Y.-W. (2009). Seasonal ARIMA forecasting of inbound air travel arrivals to Taiwan. *Transportmetrica*, 5(2), 125-140.
- Chu, K. K. A., & Chapleau, R. (2008). Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record*. <https://doi.org/10.3141/2063-08>
- Cipriani, E., Gori, S., & Petrelli, M. (2012). Transit network design: A procedure and an application to a large urban area. *Transportation Research Part C: Emerging Technologies*, 20(1), 3-14.
- Cyprich, O., Konečný, V., & Kilianova, K. (2013). Short-term passenger demand forecasting using univariate time series theory. *PROMET-Traffic&Transportation*, 25(6), 533-541.
- Cyril, A., Mulangi, R. H., & George, V. (2019). Bus passenger demand modelling using time-series techniquesbig data analytics. *The Open Transportation Journal*, 13(1).
- de Dios Ortúzar, J., & Willumsen, L. G. (2011). *Modelling transport*. John Wiley & sons.
- Decouvelaere, R., Trépanier, M., & Agard, B. (2022). Modulated spatiotemporal clustering of smart card users. *Public Transport*, 1-24.
- Delbosc, A., & Currie, G. (2011). The spatial context of transport disadvantage, social exclusion and well-being. *Journal of Transport Geography*. <https://doi.org/10.1016/J.JTRANGE.2011.04.005>
- Deschaintres, E., Morency, C., & Trépanier, M. (2018). Analyzing Interpersonal and Intrapersonal Variability of Transit Use with Smart Card Data.
- Deschaintres, E., Morency, C., & Trépanier, M. (2019). Analyzing Transit User Behavior with 51 Weeks of Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*. <https://doi.org/10.1177/0361198119834917>
- Deschaintres, E., Morency, C., & Trépanier, M. (2022). Cross-analysis of the variability of travel behaviors using one-day trip diaries and longitudinal data. <https://doi.org/10.1016/J.TRA.2022.07.013>
- Egu, O., & Bonnel, P. (2020). Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon. *Travel behaviour and society*. <https://doi.org/10.1016/J.TBS.2019.12.003>
- Erhardt, G. D., Roy, S., Cooper, D., Sana, B., Chen, M., & Castiglione, J. (2019). Do transportation network companies decrease or increase congestion? *Science advances*, 5(5), eaau2670.
- Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*. <https://doi.org/10.1145/2379776.2379788>
- Frondel, M., & Vance, C. (2011). Rarely enjoyed? A count data analysis of ridership in Germany's public transport. *Transport Policy*, 18(2), 425-433.

- Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*. <https://doi.org/10.1016/J.ENGAPPAI.2010.09.007>
- Furth, P. G., & Wilson, N. H. (1981). Setting frequencies on bus routes: Theory and practice. *Transportation Research Record*, 818(1981), 1-7.
- Ghaemi, M. S., Agard, B., Trépanier, M., & Nia, V. P. (2017). A visual segmentation method for temporal smart card data. <https://doi.org/10.1080/23249935.2016.1273273>
- Ghosh, B., Basu, B., & O'Mahony, M. (2009). Multivariate short-term traffic flow forecasting using time-series analysis. *IEEE transactions on intelligent transportation systems*, 10(2), 246-254.
- Hall, P. (2007). Urban land, housing, and transportation: the global challenge. *World Resour*, 1998, 99.
- Hammerle, M., Haynes, M., & McNeil, S. (2005). Use of automatic vehicle location and passenger count data to evaluate bus operations: experience of the Chicago Transit Authority, Illinois. *Transportation Research Record*, 1903(1), 27-34.
- Han, J., Kamber, M., & Pei, J. (2006). *Data Mining: Concepts and Techniques*, 3rd edition.
- He, L., Agard, B., & Trépanier, M. (2017). Comparing Time Series Segmentation Methods for the Analysis of Transportation Patterns with Smart Card Data.
- He, L., Agard, B., & Trépanier, M. (2020). A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. *Transportmetrica*. <https://doi.org/10.1080/23249935.2018.1479722>
- Jha, S. B., Jha, J. K., & Tiwari, M. K. (2019). A multi-objective meta-heuristic approach for transit network design and frequency setting problem in a bus transit system. *Computers & Industrial Engineering*, 130, 166-186.
- Kamarianakis, Y., & Prastacos, P. (2005). Space-time modeling of traffic flow. *Computers & Geosciences*, 31(2), 119-133.
- Kechagiopoulos, P. N., & Beligiannis, G. N. (2014). Solving the urban transit routing problem using a particle swarm optimization based algorithm. *Applied Soft Computing*, 21, 654-676.
- Kenyon, S., Lyons, G., & Rafferty, J. (2002). Transport and social exclusion: investigating the possibility of promoting inclusion through virtual mobility. *Journal of Transport Geography*. [https://doi.org/10.1016/S0966-6923\(02\)00012-1](https://doi.org/10.1016/S0966-6923(02)00012-1)
- Kieu, L.-M., Ngoduy, D., Malleson, N., & Chung, E. (2019). A stochastic schedule-following simulation model of bus routes. *Transportmetrica B: Transport Dynamics*, 7(1), 1588-1610.
- Kumar, S. V., & Vanajakshi, L. (2015). Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review*, 7(3), 1-9.
- Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C-emerging Technologies*. <https://doi.org/10.1016/J.TRC.2014.05.012>

- Lee, R. K.-W., & Kam, T. S. (2014). Time-Series Data Mining in Transportation: A Case Study on Singapore Public Train Commuter Travel Patterns. *International journal of engineering and technology*. <https://doi.org/10.7763/IJET.2014.V6.737>
- Li, T., Sun, D., Jing, P., & Yang, K. (2018). Smart card data mining of public transport destination: A literature review. *Information*, 9(1), 18.
- Lucas, K. (2012). Transport and social exclusion: Where are we now? *Transport Policy*. <https://doi.org/10.1016/J.TRANPOL.2012.01.013>
- Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y.-J. (2017). Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, 58, 135-145.
- Ma, X., Wu, Y. J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C-emerging Technologies*. <https://doi.org/10.1016/J.TRC.2013.07.010>
- Mariñas-Collado, I., Sipols, A. E., Santos-Martín, M. T., & Frutos-Bernal, E. (2022). Clustering and forecasting urban bus passenger demand with a combination of time series models. *Mathematics*, 10(15), 2670.
- Menon, A. K., & Lee, Y. (2017). Predicting short-term public transport demand via inhomogeneous Poisson processes. (p. 2207-2210).
- Menon, A. K., & Lee, Y. (2017). Predicting short-term public transport demand via inhomogeneous Poisson processes. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (p. 2207-2210).
- Mohamed, K., Côme, E., Oukhellou, L., & Verleysen, M. (2016). Clustering smart card data for urban mobility analysis. *IEEE Transactions on intelligent transportation systems*, 18(3), 712-728.
- Morency, C., Trépanier, M., & Agard, B. (2006). Analysing the Variability of Transit Users Behaviour with Smart Card Data. <https://doi.org/10.1109/ITSC.2006.1706716>
- Morency, C., Trépanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3), 193-203.
- Munizaga, M., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smartcard data. <https://doi.org/10.1016/J.TRC.2014.03.008>
- Munizaga, M., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C-emerging Technologies*. <https://doi.org/10.1016/J.TRC.2012.01.007>
- Nikolić, M., & Teodorović, D. (2013). Transit network design by bee colony optimization. *Expert Systems with Applications*, 40(15), 5945-5955.
- Park, J. Y., Kim, D.-J., & Lim, Y. (2008). Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea. *Transportation Research Record*. <https://doi.org/10.3141/2063-01>
- Patnaik, J., Chien, S., & Bladikas, A. (2006). Using data mining techniques on APC data to develop effective bus scheduling plans. *Journal of Systemics, Cybernetics and Informatics*, 4(1), 86-90.

- Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4), 557-568.
- Rainville Jr, W. S. (1947). *Bus Scheduling Manual: Traffic Checking and Schedule Preparation*. United States. Department of Transportation. Urban Mass Transportation
- Richer, C., Côme, E., Mahrsi, M. K. E., & Oukhellou, L. (2018). La mobilité intermodale par les données billettiques. Analyses spatio-temporelles du réseau bus-métro de Rennes Métropole. *Cybergeo: European Journal of Geography*. <https://doi.org/10.4000/CYBERGEO.29132>
- Robinson, S., Narayanan, B., Toh, N., & Pereira, F. (2014). Methods for pre-processing smartcard data to improve data quality. *Transportation Research Part C: Emerging Technologies*, 49, 43-58.
- Shaw, J. (2019). Review of *The Geography of Transport Systems* (4th ed.), by J.-P. Rodrigue, C. Comtois, & B. Slack. *Journal of Transport Geography*, 74, 102563. Routledge. <https://doi.org/10.1016/j.jtrangeo.2019.102563>
- Sun, L., Jin, J. G., Lee, D.-H., Axhausen, K. W., & Erath, A. (2014). Demand-driven timetable design for metro services. *Transportation Research Part C: Emerging Technologies*, 46, 284-299.
- Thacker, S., Adshead, D., Fay, M., Hallegatte, S., Harvey, M., Meller, H., . . . Hall, J. W. (2019). Infrastructure for sustainable development. *Nature Sustainability*, 2(4), 324-331.
- Torti, A., Galvani, M., Menafoglio, A., Secchi, P., & Vantini, S. (2022). A general bi-clustering algorithm for object data with an application to the analysis of a Lombardy railway line. <https://doi.org/10.1016/J.IJAR.2021.12.003>
- Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *J. Intell. Transp. Syst.* <https://doi.org/10.1080/15472450601122256>
- Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, 3-19.
- Wang, Y., Ram, S., Currim, F., Dantas, E., & Sabóia, L. A. (2016). A big data approach for smart transportation management on bus network. *2016 IEEE International Smart Cities Conference (ISC2)*. <https://doi.org/10.1109/ISC2.2016.7580839>
- Williams, B. M., & Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6), 664-672.
- Zhao, J., Qu, Q., Zhang, F., Xu, C.-Z., & Liu, S. (2017). Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*. <https://doi.org/10.1109/TITS.2017.2679179>
- Zhou, C., Dai, P., Wang, F., & Zhang, Z. (2016). Predicting the passenger demand on bus services for mobile users. *Pervasive and Mobile Computing*, 25, 48-66.

ANNEXE A – ANALYSE DESCRIPTIVE

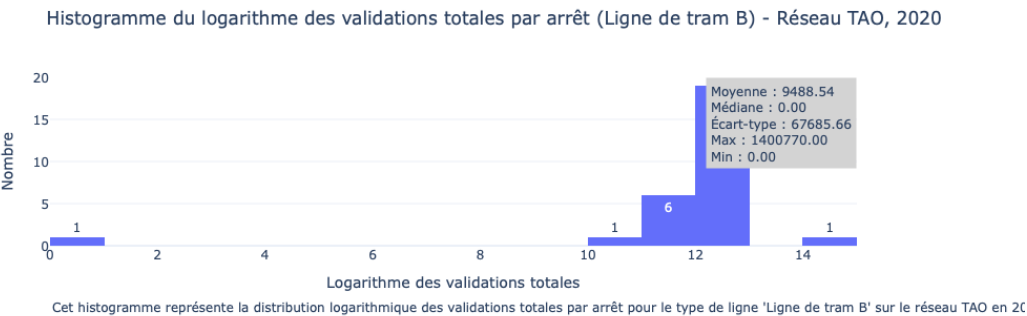


Figure A.1 - Histogramme des validations de la ligne B du Tram

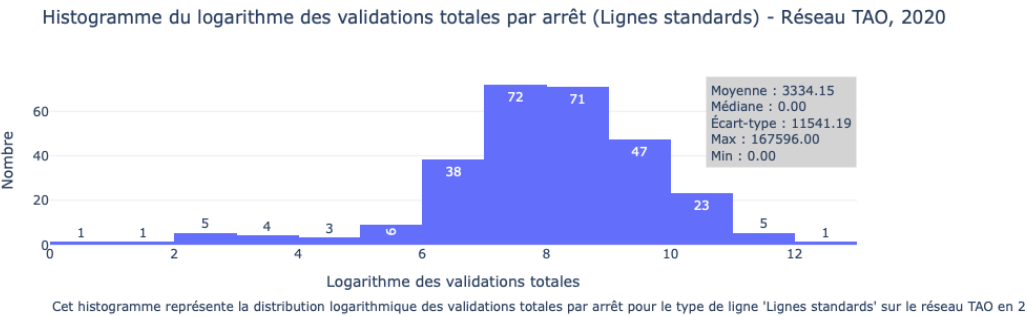


Figure A.2 - Histogramme des validations des lignes de Bus standards

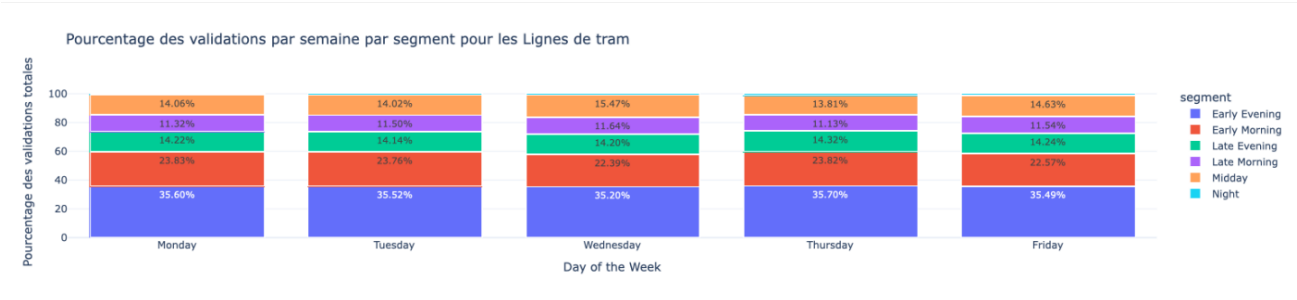


Figure A.3 - Distribution du pourcentage de validations par segment de la journée des lignes de tram

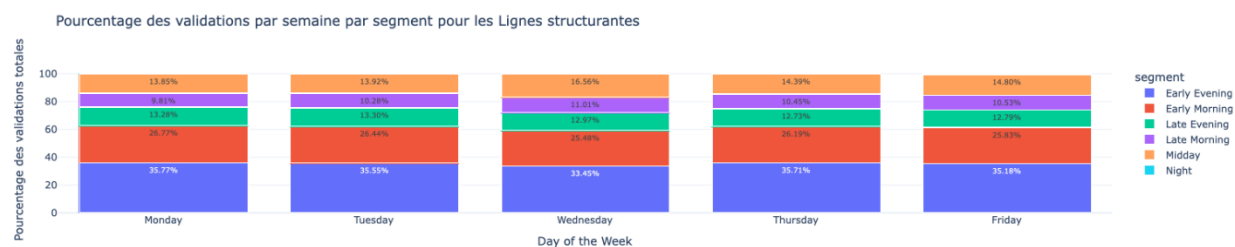


Figure A.4 - Distribution en pourcentage des validations par segment de la journée des lignes structurantes

20 arrêts les plus fréquentés sur l'ensemble du réseau TAO en 2020

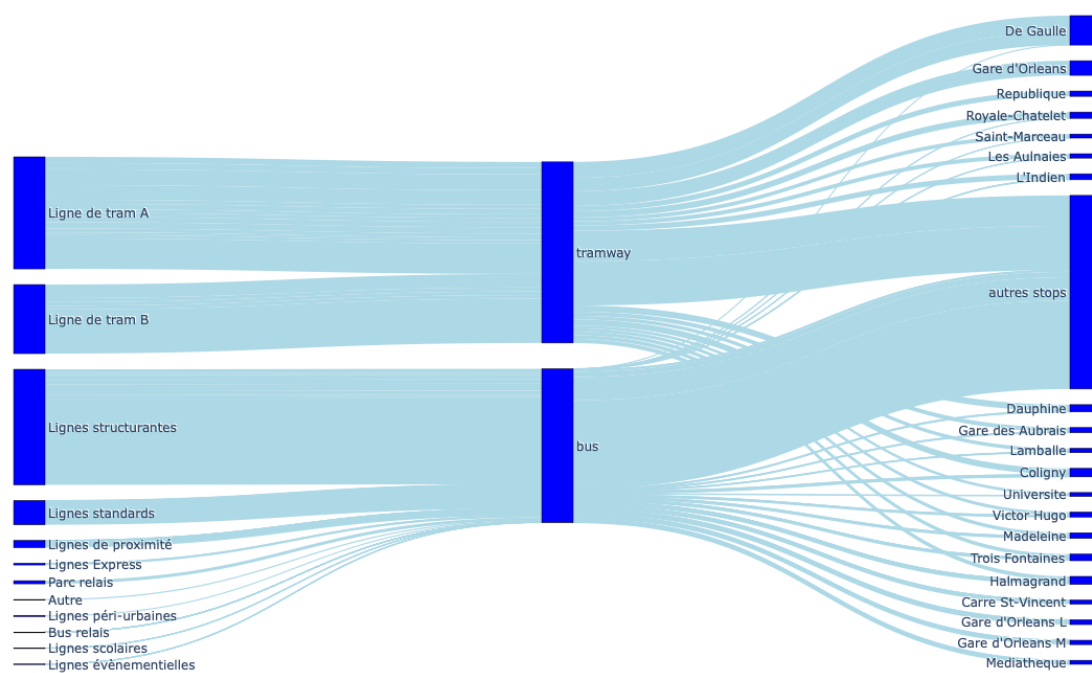


Figure A.5 - Les 20 arrêts les plus fréquentés du réseau TAO en 2020

ANNEXE B – RÉSULTATS DE LA SEGMENTATION

Tableau B.1 - Évaluation détaillée des groupes du K-moyennes, réseau

Cluster	Mean	Median	Mode	Range	Standard Deviation	IQR	Skewness	Kurtosis	25th Percentile	50th Percentile	75th Percentile	Number of Observations	Center of Cluster
0	2353.298837	1370.0	52	13442	2686.751099	2613.5	1.677850	2.288425	482.0	1370.0	3095.5	1720	2353.298837
1	4587.640517	4495.0	0	20382	3658.076308	5907.0	0.652442	0.063126	1085.0	4495.0	6992.0	2320	4587.640517
2	9158.407407	9182.0	4	28442	7216.213794	12847.5	0.384958	-0.885394	1633.5	9182.0	14481.0	2160	9158.407407

Tableau B.2 - Analyse détaillée du groupe 0, K-moyennes sur le réseau

Groupe 0

Moyenne Heure	Médiane	Mode	Range	Écart-type	IQR	Skewness	Kurtosis	25th Percentile	50th Percentile	75th Percentile	Nombre d'observations	Centre du groupe
0 129.279070	124.0	10.0	306	80.910066	137.0	0.074607	-0.970623	59.0	124.0	196.0	86	129.279070
5 197.093023	132.0	46.0	536	164.045232	303.0	0.517159	-1.270927	50.5	132.0	353.5	86	197.093023
6 300.813953	98.0	64.0	832	279.622835	537.5	0.541820	-1.467873	66.0	98.0	603.5	86	300.813953
7 773.883721	416.0	304.0	2152	600.635691	1046.5	0.727558	-0.952919	304.5	416.0	1351.0	86	773.883721
8 1548.441860	930.0	382.0	4504	1181.484506	1852.5	0.814854	-0.648583	640.0	930.0	2492.5	86	1548.441860
9 2456.441860	1597.0	546.0	7188	1860.805933	2941.5	0.795445	-0.634539	1039.0	1597.0	3980.5	86	2456.441860
10 3081.232558	2076.0	774.0	8786	2349.131848	3729.5	0.768659	-0.719013	1216.0	2076.0	4945.5	86	3081.232558
11 3410.813953	2276.0	44.0	9880	2593.085401	4205.0	0.817296	-0.631180	1312.5	2276.0	5517.5	86	3410.813953
12 3422.046512	2277.0	36.0	9888	2651.447267	3915.5	0.924683	-0.360595	1363.5	2277.0	5279.0	86	3422.046512
13 3651.302326	2504.0	2504.0	11162	2847.997666	4268.0	0.979884	-0.080327	1408.0	2504.0	5676.0	86	3651.302326
14 4289.790698	2872.0	1464.0	12770	3472.175043	5234.0	0.968954	-0.266107	1635.0	2872.0	6869.0	86	4289.790698
15 4378.325581	2637.0	1574.0	12904	3536.728643	5270.0	0.938442	-0.378439	1749.0	2637.0	7019.0	86	4378.325581
16 4361.767442	2708.0	1876.0	13382	3465.532828	5193.0	0.915943	-0.405538	1744.0	2708.0	6937.0	86	4361.767442
17 4333.465116	2881.0	74.0	12784	3423.739125	4936.5	0.946365	-0.300594	1661.5	2881.0	6598.0	86	4333.465116
18 3966.790698	2912.0	3108.0	10630	2904.456961	4433.5	0.852254	-0.544436	1676.5	2912.0	6110.0	86	3966.790698
19 3066.604651	2326.0	60.0	7946	2118.686920	3243.0	0.747334	-0.687791	1355.5	2326.0	4598.5	86	3066.604651
20 1820.860465	1559.0	3482.0	4138	1097.042464	1663.5	0.471043	-0.910800	994.0	1559.0	2657.5	86	1820.860465
21 940.488372	806.0	728.0	2446	593.503174	903.0	0.534491	-0.675719	482.5	806.0	1385.5	86	940.488372
22 564.069767	511.0	122.0	1610	370.674609	564.5	0.578724	-0.385107	257.0	511.0	821.5	86	564.069767
23 372.465116	268.0	44.0	1186	312.953396	471.0	0.914169	-0.220913	146.5	268.0	617.5	86	372.465116