

Louvain School of Management

Optimisation de la gestion des flux de visiteurs dans un parc à thème

Cas d'application Pairi Daiza

Auteur : Noé Drion du Chapois
Promoteur : Daniele Catanzaro
Année académique 2024-2025
Master 120 en Ingénieur de gestion Business Analytics
Horaire de jour

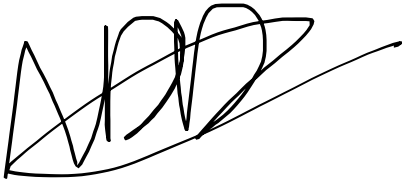
Déclaration

Durant la préparation de ce mémoire, l'auteur a utilisé CHATGPT, MISTRALAI et SCRIBENS pour les raisons suivantes :

- 1) RAISON : aide pour de la relecture et détection de fautes d'orthographe ainsi que pour trouver des synonymes et de la reformulation.
- 2) Après avoir utilisé CHATGPT, MISTRALAI et SCRIBENS, l'auteur a soigneusement révisé et édité le contenu produit par l'outil. Nous assumons la pleine responsabilité du contenu final présenté dans ce mémoire.

En signant cette déclaration, nous affirmons que le contenu de ce mémoire reflète notre travail original, enrichi par l'utilisation responsable de l'IA.

Noé Drion du Chapois

A handwritten signature in black ink, consisting of stylized, overlapping loops and a long horizontal stroke extending to the right.

Remerciements

Je désirerais d'abord remercier le Professeur Daniele Catanzaro pour sa disponibilité et son aide dans la construction de ce mémoire.

Je souhaiterais ensuite remercier AGILOS, l'entreprise de consultance dans laquelle j'ai effectué mon stage de fin d'études et qui m'a fourni tout l'encadrement et les outils nécessaires à la réalisation de ce beau projet.

Je tiens également à remercier Pairi Daiza pour m'avoir permis d'accéder aux données qui m'ont permis de développer cette solution.

Merci également à ma famille ainsi qu'à mes proches pour leur compréhension et leur patience tout au long de mes études.

Table des matières

Introduction.....	1
1. Contexte théorique	3
1.1. Impact stratégique de la gestion des flux.....	3
1.2. Dualité réactive vs proactive	4
1.3. Stratégies de réorientation	4
1.4. L'apprentissage machine appliqué à la prédiction d'affluence	5
1.4.1. Régression linéaire.....	5
1.4.2. Arbres de décisions	6
1.5. Optimisation et évaluation des modèles	8
1.6. Modèle de prédiction d'affluence : cas d'application.....	8
1.7. Optimisation : modélisation des flux	9
1.8. Problème de conception de réseaux capacitaires.....	10
1.9. Optimisation des flux : cas d'application	11
2. Approche.....	13
3. Machine Learning	13
3.1. Méthodologie.....	13
3.1.1. Collecte des données.....	13
3.1.2. Analyse descriptive des données	14
3.1.3. Transformation et enrichissement des données	19
3.1.4. Prédiction du nombre de tickets.....	20
3.2. Résultats	21
3.3. Vers une stratégie d'optimisation	24
4. Optimisation.....	25
4.1. Méthodologie.....	25
4.1.1. Optimisation du nombre de visiteurs	25

4.1.2. Présentation du modèle de base	25
4.2. Résultats	28
4.2.1. Scénarios spécifiques	28
4.2.2. Effets sur les données historiques	29
4.2.3. Variations du modèle	32
5. Discussion	36
5.1. Interprétation des résultats	36
5.2. Limites méthodologiques	37
5.3. Perspectives	38
Conclusion	39
Références	41
6. Annexes	45
Annexe 1 : Réaction de la demandes dans les restaurants aux conditions météorologiques	45
Annexe 2 : Réaction de la demandes dans les restaurants aux conditions météorologiques	50
Annexe 3 : Tableau de description des variables pour le modèle prédictif	53
Annexe 4 : Flux d'optimisation Microsoft Fabric	55
Annexe 5 : Carte Power BI du niveau de saturation des zones	55
Annexe 6 : Tableau des notations	56

Table des figures

Figure 1: Schéma d'un arbre de décision Source : Jiang et al. 2020	6
Figure 2 : Évolution du nombre de commandes par année.....	14
Figure 3 : Comportement du nombre de commandes par mois	15
Figure 4 : Comportement du nombre de commandes par jour de la semaine.....	15
Figure 5 : Comportement du nombre de commandes par tranche horaire.....	15
Figure 6 : Nombre de commandes par restaurant	16
Figure 7 : Comportement du nombre moyen de commandes par type de temps.....	16
Figure 8 : Comportement du nombre moyen de commandes par température.....	16
Figure 9 : Comportement du nombre moyen de commandes par type de jours	17
Figure 10 : Graphique combiné du nombre de commandes et nombre de visiteurs.....	17
Figure 11 : Comportement spécifique du restaurant F.....	18
Figure 12 : Comportement spécifique du restaurant I	18
Figure 13 : Matrice de corrélation de Pearson	19
Figure 14 : Importance des variables explicatives dans la prédiction de XGBoost.....	23
Figure 15 : Visualisation des prédictions finales	24
Figure 16 : Réseau des points de restauration.....	28
Figure 17: Capacité de dépassement maximale de chaque restaurant	29
Figure 18 : Nombre de périodes de dépassement pour le modèle de base	30
Figure 19 : Réorientations proposées par le modèle de base ($K_{\max} = 1$).....	30
Figure 20 : Amélioration du dépassement par restaurant avec le modèle de base	31
Figure 21 : Nombre de périodes de dépassement pour le modèle de multi-réorientation	33
Figure 22 : Nombre de périodes de dépassement pour le modèle avec graphe complet	34
Figure 23 : Taux d'occupation maximal avec lissage	35
Figure 24 : Réaction du restaurant A aux conditions météorologiques	45
Figure 25 : Réaction du restaurant B aux conditions météorologiques	45
Figure 26 : Réaction du restaurant C aux conditions météorologiques	46
Figure 27 : Réaction du restaurant D aux conditions météorologiques	46
Figure 28 : Réaction du restaurant E aux conditions météorologiques	47
Figure 29 : Réaction du restaurant F aux conditions météorologiques.....	47
Figure 30 : Réaction du restaurant G aux conditions météorologiques	48
Figure 31 : Réaction du restaurant H aux conditions météorologiques	48

Figure 32 : Réaction du restaurant I aux conditions météorologiques.....	49
Figure 33 : Comportement de chaque restaurant à la pluie	49
Figure 34 : Comportement spécifique du restaurant A.....	50
Figure 35 : Comportement spécifique du restaurant B	50
Figure 36 : Comportement spécifique du restaurant C	51
Figure 37 : Comportement spécifique du restaurant D.....	51
Figure 38 : Comportement spécifique du restaurant E	52
Figure 39 : Comportement spécifique du restaurant G.....	52
Figure 40 : Comportement spécifique du restaurant H.....	53

Introduction

Pairi Daiza est l'un des plus grands et populaires zoos d'Europe. Élu meilleur zoo d'Europe à plusieurs reprises, Pairi Daiza a accueilli, en 2024, au sein de ses 75 hectares, plus de 2.65 millions de visiteurs. Avec plus de 800 espèces animales différentes et plus de 7 500 animaux, le parc attire des visiteurs du monde entier. Le zoo est constitué de neuf mondes immersifs qui permettent un voyage à travers les cinq continents. Pairi Daiza contient de nombreux restaurants ainsi que des points de restauration à emporter.

Toutefois, la gestion d'un tel espace accueillant un nombre aussi important de visiteurs peut être particulièrement complexe. Gérer le flux de foules au sein du parc est un défi organisationnel majeur. Le concept de parc animalier provoque une grande dispersion des visiteurs au sein du parc, chacun choisissant son parcours, comme il le souhaite pour voir les animaux qu'il désire. Cependant, plusieurs points névralgiques peuvent se créer, dus à des événements ou à des points de restauration au sein du parc. Ces points névralgiques accueillent un nombre très différent de personnes en fonction du jour, de l'heure, des événements au sein du parc, de la météo, etc.

La gestion du flux de visiteurs représente un défi majeur dans les parcs à thème. Ce concept a une grande importance dans l'expérience des visiteurs et, par conséquent, la satisfaction du consommateur. De plus, une expérience fluide permet une meilleure rentabilité économique pour le parc. Aux heures des repas, les points de restauration peuvent rapidement se retrouver saturés. C'est dans ce contexte que la gestion du flux de visiteurs est particulièrement pertinente pendant ces périodes. Ce défi n'est pas uniquement présent dans les parcs à thème, mais également au sein de toute structure proposant des points de vente.

Pour répondre à cet enjeu, plusieurs questions se posent. La première concerne la manière de prédire avec précision l'affluence des visiteurs dans les points de restauration d'un parc animalier comme Pairi Daiza. Ensuite, comment peut-on exploiter ces prédictions pour réaliser une suggestion opérationnelle des réorientations optimales de visiteurs ?

Pour répondre à ces questions, ce mémoire explorera une solution en deux temps pour optimiser l'affluence des visiteurs au sein des points de restauration de Pairi Daiza. Pour optimiser cette gestion du flux de visiteurs, une combinaison d'algorithmes prédictifs et d'algorithmes d'optimisation sera utilisée. Une telle combinaison permettra d'anticiper l'affluence et de

répartir intelligemment les visiteurs de manière proactive. Cela mettra en avant une approche proactive à la différence d'une approche réactive. Pour cela, ce mémoire se découpera en deux grandes parties, la première dédiée à l'apprentissage machine et une seconde à l'optimisation des flux. Cette approche s'inscrit dans un cadre méthodologique reconnu appelé *predict-then-optimize*, qui utilise les résultats de prédictions comme données d'entrées pour un modèle d'optimisation, afin de procéder à des suggestions opérationnels.

Le premier enjeu sera de construire un modèle de prédiction d'affluence. Le but sera d'avoir un horizon temporel d'une heure pour permettre aux gestionnaires du parc d'anticiper les pics de fréquentation et d'adapter en conséquence les ressources disponibles. Pour la réalisation de ce modèle, différentes méthodes d'apprentissage machine seront analysées avec pour but la création d'un modèle fiable et robuste. Afin d'alimenter ces modèles d'apprentissage machine, plusieurs éléments clés seront ajoutés aux données des tickets des points de vente, tels que la météo, les dates des congés scolaires, l'affluence au sein du parc, etc.

Après avoir déterminé l'affluence des restaurants dans un horizon temporel d'une heure, la seconde étape consistera en la modélisation d'un algorithme capable d'optimiser dynamiquement les flux de visiteurs. Pour cela, un ensemble de contraintes mathématiques d'optimisation linéaire sera modélisé et implémenté. Cela proposera une suggestion opérationnelle de solutions de répartition des visiteurs vers les restaurants les moins saturés. Ensuite, le parc disposera de plusieurs méthodes pour appliquer cette suggestion opérationnelle concrète.

Le but de ce mémoire est de réaliser une suggestion opérationnelle proposant une stratégie optimale de redirection des visiteurs. Ce projet pourrait non seulement permettre une meilleure expérience du client, mais également améliorer l'utilisation des ressources au sein des points de restauration, et par conséquent, accroître la rentabilité.

1. Contexte théorique

1.1. Impact stratégique de la gestion des flux

La gestion des flux de visiteurs est un enjeu crucial dans les parcs à thème, elle impacte directement l'expérience du client et par la même occasion, la compétitivité du parc (Ma et al., 2013). L'une des sources majeures de l'insatisfaction des visiteurs est la congestion (Yuan et al., 2018). La gestion des flux permet de limiter la congestion en optimisant le parcours de personnes pour assurer l'expérience la plus fluide et agréable possible. Cependant, l'affluence dans les parcs à thème est souvent bien complexe. Cela est dû à la multitude de possibilités que les visiteurs ont pour se déplacer et par conséquent de la flexibilité de leurs itinéraires (Yuan et al., 2018). Le contrôle et l'estimation de l'affluence future sont rendus complexes par le nombre accru d'activités dans le parc (Kataoka et al., 2005). Or, la gestion des flux de visiteurs est fondamentale pour l'efficacité opérationnelle des parcs à thème, car cela a des répercussions sur la satisfaction des clients et les revenus du parc (Clavé, 2023b). En effet, ces meilleures performances économiques sont dues à plusieurs facteurs tels que la satisfaction du visiteur, ce qui conduit à une augmentation de la répétition de leurs visites. Une meilleure répartition des visiteurs engendre également une augmentation des ventes intra-muros, ainsi qu'une meilleure allocation des ressources qui conduit à une réduction du coût opérationnel. La satisfaction du visiteur joue également dans la recommandation du parc à son entourage, ce qui a également tendance à attirer de nouveaux visiteurs (Clavé, 2023a). Elle est donc un élément clé, dans ces infrastructures, qui est liée à plusieurs concepts tels que : la réduction du temps d'attente, la fluidité dans les déplacements, une bonne répartition dans les zones de restauration ou des commerces, ainsi que l'optimisation dans l'utilisation des attractions. Tous ces éléments ont un lien direct avec la gestion du flux de visiteurs au sein du parc (Clavé, 2023b).

La perception de la congestion peut également varier d'une personne à une autre. Elle sera beaucoup plus importante chez une personne qui recherche une expérience calme et relaxante que chez une personne qui s'attend à une ambiance animée. C'est pourquoi, en fonction des zones visitées, la perception de l'attente peut varier de manière importante. Dans un parc à thème, la foule peut donc être plus facilement acceptée dans certaines zones ou certains contextes (Jin & Paerce, 2011). L'acceptation dépend aussi de la réputation de l'endroit, lorsqu'un site est réputé pour être fortement fréquenté, les visiteurs sont moins susceptibles d'être déçus, car leurs attentes ont été influencées en connaissance de cause (Clavé, 2023c).

1.2. Dualité réactive vs proactive

Deux approches sont à distinguer : la coordination prospective et réactive. La coordination prospective a pour but de prédire et donc de prévenir l'affluence de visiteurs, c'est une approche proactive. Tandis que la coordination réactive consiste à subir, et ensuite, réagir face à la congestion des foules. La coordination réactive se base sur la situation actuelle en ignorant les projections futures. Le concept d'oscillation de foule est important dans ce contexte. Une oscillation de foule survient lorsque la gestion du flux de visiteurs se base sur des données actuelles. Dans le cas où les visiteurs recevraient une information sur la congestion actuelle, ils choisiraient de se rediriger, créant un effet de balancier entraînant un déplacement de foule. L'oscillation de foule montre que la coordination réactive mène à la congestion de nouvelles zones en laissant les zones auparavant bondées relativement vides. (Brown et al., 2013; Kataoka et al., 2005)

Afin de surmonter la limite de l'approche réactive, il est essentiel de se tourner vers une approche prospective. Cette méthode permet d'anticiper la distribution spatio-temporelle des visiteurs, ce qui a pour but d'optimiser l'expérience au sein du parc. La prédiction de l'affluence repose sur l'analyse de données historiques ainsi que la définition d'un modèle prédictif (Kim, 2024). Elle est un élément clé de la coordination prospective. Cependant, dans le cas où la demande serait sous-estimée, cela provoquerait de l'insatisfaction chez le visiteur due au manque de services ou d'infrastructures qui lui sont proposés (Kim, 2024). Une étude de Kim a montré qu'à l'aide de données historiques sur l'affluence au sein du parc, et de variables telles que des facteurs météorologiques ou les caractéristiques du jour, il est possible d'estimer le nombre de visiteurs d'un parc à thème. Cette étude a démontré une précision correcte avec un R^2 à 67,3 % avec des erreurs principales dues à des facteurs imprévus qui n'ont pas été pris en compte dans le modèle.

1.3. Stratégies de réorientation

Plusieurs solutions sont envisageables pour gérer efficacement les foules dans les sites touristiques. La signalisation dynamique permet d'orienter le client dans une direction moins bondée. Une signalétique interactive peut aider à répartir les flux de personnes plus équitablement dans le parc. Une répartition proactive des visiteurs peut également être appliquée en attirant des clients dans des zones moins populaires. C'est le cas en proposant des recommandations personnalisées ou des offres intéressantes en temps réel, afin d'attirer le

visiteur vers un endroit moins bondé. L'aménagement spatial des espaces a également une importance essentielle. Un aménagement efficace peut être constitué de panneaux directionnels interactifs, de chemins correctement définis, mais également d'éléments architecturaux comme des sculptures. Une technique particulièrement utilisée à Disney avec leurs éléments architecturaux appelés « wienies ». Les « wienies » sont des éléments de décor qui ont pour but d'agir comme des aimants visuels. Le plus connu est le Château de la Belle au Bois Dormant, qui a tendance à attirer les visiteurs de la 'Main Street' à la place centrale (Clavé, 2023c).

1.4. L'apprentissage machine appliqué à la prédiction d'affluence

L'apprentissage machine est une branche de la science informatique, qui a pour but de construire des modèles capables de réaliser des prédictions à partir de données d'entraînement. Ces algorithmes peuvent être classés en trois catégories : supervisés et non-supervisés et semi-supervisés. Les algorithmes d'apprentissage supervisé sont utilisés sur des ensembles de données étiquetées, ceux-ci contiennent donc des entrées et des sorties désirées. Leur objectif est d'apprendre sur base de ces données et de prédire une sortie pour de nouvelles entrées aux labels inconnus. En contraste, les algorithmes non supervisés fonctionnent avec des données non étiquetées, visant à découvrir des structures et patterns intrinsèques dans les données. L'apprentissage semi-supervisé, quant à lui, combine les avantages des deux méthodes en utilisant des données étiquetées ainsi que non étiquetée. Cette dernière méthode est largement utilisée lorsque l'obtention de données étiquetées est coûteuse ou difficile à obtenir. Le choix du type d'algorithme utilisé se fait en fonction de la nature du problème et des données disponibles (Jiang et al., 2020 ; Géron, 2019, chap. 2). L'apprentissage machine est de plus en plus utilisé pour prédire la demande dans divers secteurs (Zhou et al., 2025). Dans le cas de ce mémoire, les données récupérées sont étiquetées, c'est pourquoi la suite de cette littérature cible des algorithmes d'apprentissage machine supervisés. Les algorithmes d'apprentissage machine, basés sur des données supervisées, traités dans cette étude, se divisent en 2 différents groupes : la régression linéaire et les arbres de décision.

1.4.1. Régression linéaire

La régression linéaire est une méthode qui permet d'étudier les relations linéaires entre une variable cible et une ou plusieurs variables explicatives. Le principe de ce modèle repose sur

une relation linéaire entre variables. Mathématiquement, le modèle est représenté par cette équation :

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_p \cdot x_p + \epsilon$$

Où y est la variable cible, x_i correspond aux variables explicatives, β_i les coefficients estimés par le modèle et ϵ l'erreur. L'un des avantages de ce modèle est sa simplicité et sa compréhensibilité. En effet, chaque coefficient β_i explique l'influence de la variable explicative i sur la variable cible. La simplicité de ce modèle a des limites, car cette équation suppose : les relations entre variables linéaires, chaque observation indépendante, une homoscedasticité¹, ainsi que les erreurs qui suivent une distribution normale. Dans des environnements complexes, ces conditions ne sont pas respectées, ce qui montre les limites de ce modèle. Dans le cas où il y a de nombreuses variables, parfois liées entre-elles, des modèles comme la régression Ridge ou Lasso peuvent permettre d'éviter des erreurs d'ajustement. (Xin et al., 2009, chap. 1, 2)

1.4.2. Arbres de décisions

Les arbres de décisions sont des méthodes d'apprentissage supervisé qui sont adaptées aux prédictions impliquant plusieurs variables. Les modèles basant sur des arbres de décision sont capables de modéliser des problèmes de classification et de régression. Un arbre de décision est une structure qui partitionne l'espace de données en régions distinctes via une série de règles de décision binaire. (Breiman et al., 1984)

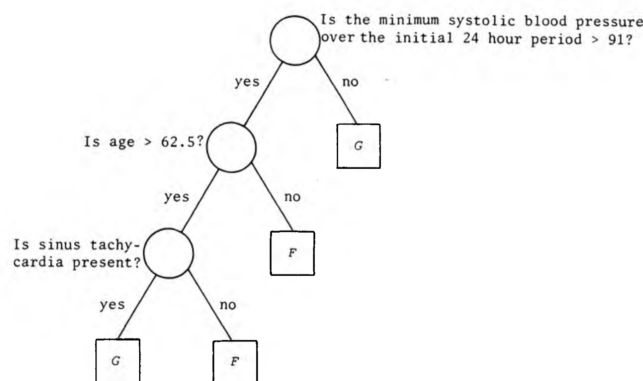


Figure 1: Schéma d'un arbre de décision Source : Jiang et al. 2020

¹ Homoscedasticité : les erreurs du modèle ont une dispersion constante quelle que soit la valeur des variables.

La figure 1 illustre un arbre de décision qui montre la division des données en régions distinctes non-chevauchantes. Un arbre de décision se développe à partir d'un nœud racine et se développe grâce à des décisions binaires à chaque nœud suivant. Un nœud terminal, qui n'est plus divisé, est appelé une feuille. Chaque feuille contient une prédiction finale qui est calculée comme étant la moyenne des observations présentes dans la feuille. La structure de l'arbre est créée avec comme objectif de minimiser l'erreur de prédiction. La profondeur d'un arbre est le nombre maximal de divisions effectuées avant la feuille finale. Plus un arbre est profond, plus il capte de relations complexes entre variables. Cependant, lorsque l'arbre de décision considère des relations trop complexes, le modèle se spécialise sur les données utilisées lors de la création de l'arbre et n'est pas généralisable sur de nouvelles données : c'est le surentraînement (Jiang et al. 2020).

Plusieurs techniques sont utilisées pour contrôler le surapprentissage, la première est l'étalement. Cela consiste à supprimer les dernières feuilles de l'arbre afin d'avoir un meilleur équilibre entre généralisation et précision. La validation croisée est également une méthode qui permet de limiter le surapprentissage. Pour cela, les données sont divisées en plusieurs sous-ensembles et le modèle est testé et entraîné à plusieurs reprises sur des ensembles distincts, pour un résultat plus robuste. (Xin et al., 2009, chap. 4 ; Hastie et al., 2009, chap. 7)

De nombreux modèles plus avancés se basent sur des arbres de décision : forêt aléatoire (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM). Pour améliorer la précision et le risque de surapprentissage, RF réalise plusieurs arbres de décision et combine les résultats obtenus. Chacun des arbres est créé à partir d'un échantillon aléatoire (Breiman, 2001). XGBoost est un modèle qui se base sur la descente de gradient et qui est populaire grâce à sa précision et son efficacité. Pour cette méthode, les arbres ne sont plus construits aléatoirement, mais en suivant l'approche de la descente de gradient : chaque arbre est créé en minimisant les erreurs des arbres précédents (Chen et al., 2016). LightGBM est une autre méthode du boosting qui apporte beaucoup d'importance à l'efficacité en termes de mémoire. Son efficacité est notamment due à deux techniques novatrices : le Gradient-based One-Side Sampling (GOSS) et l'Exclusive Feature Bundling (EFB). La première optimise la sélection des instances de données et la seconde réduit le nombre de caractéristiques à traiter (Ke et al., 2017).

1.5. Optimisation et évaluation des modèles

L'optimisation des hyperparamètres est une étape cruciale dans l'apprentissage machine. Le but de cette optimisation est de trouver la combinaison d'hyperparamètres, qui minimise l'erreur de prédiction du modèle. C'est une étape qui peut rapidement être longue et coûteuse en ressources. Pour cela, il est possible d'utiliser des outils d'optimisation automatisés, tels qu'Optuna. Ce framework d'optimisation utilise une approche « define-by-run », ce qui signifie que la recherche s'effectue dynamiquement en fonction des résultats des essais précédents (Akiba et al., 2019). Optuna a été introduit via une étude récente réalisée par Akiba et al. (2019) dans laquelle l'efficacité de la méthode est démontrée.

Pour évaluer la performance des modèles, les métriques couramment utilisées sont : l'écart quadratique moyen (RMSE), l'erreur absolue moyenne (MAE) et le coefficient de détermination (R^2). La métrique RMSE est la racine carrée de la moyenne des carrés des erreurs. Cette mesure statistique est grandement sensible aux erreurs extrêmes. MAE est moins sensible aux extrêmes, car elle se calcule comme la moyenne des valeurs absolues des erreurs (Hodson, 2022). Enfin, R^2 montre à quel point le modèle explique les variations de données, il vaut 1 si le modèle les explique à 100% et à 0 s'il n'explique rien. (Xin et al., 2009, chap. 2)

$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$
--	--	---

Où y_i représente la i^e observation, $\hat{y}_i$ la valeur prédite par le modèle pour la i^e observation, $\bar{y}$ est la moyenne des valeurs réelles et n le nombre total d'observations.

1.6. Modèle de prédiction d'affluence : cas d'application

Dans la littérature, les algorithmes d'apprentissage automatique comme XGBoost, LightGBM ou RF ont fait leurs preuves dans diverses applications. Une étude récente réalisée par Zhou et al. (2025), a montré la performance de XGBoost face aux modèles traditionnels, tels RF. Cette étude avait pour but de prédire le volume de patients externes dans les hôpitaux. Ce travail explique comment Zhou et al. ont optimisé leurs hyperparamètres, via des recherches croisées, et validé leurs résultats via des métriques telles que MAE, RMSE et R^2 . Les résultats de cette

étude montrent une excellente performance du modèle sur la prédiction du volume de patients externes dans les hôpitaux, avec un R^2 de 0,90.

Lorsqu'il convient de faire des prévisions de demandes, les algorithmes XGBoost et LightGBM sont fréquemment utilisés, en raison de leur efficacité ainsi que de leur précision (Hobor et al., 2025). Enfin, selon Kang et al (2022) la prédiction d'affluence de touriste est réalisable via des modèles linéaires, cependant, la réaction de la tendance touristique n'est pas toujours linéaire avec ses caractéristiques. C'est pourquoi, Kang et al. ont proposé dans leur article des extensions de XGBoost, pour réaliser la prédiction de tendances touristique. Ces modèles ont permis la prédiction des tendances touristiques à Pékin, avec un R^2 de 0,94. Ces travaux montrent la capacité des modèles à base d'arbres de décision de travailler dans des environnements complexes non-linéaire, avec pour but la prédiction d'une demande.

1.7. Optimisation : modélisation des flux

La théorie des graphes est une partie importante des mathématiques qui est traitée par Diestel dans son ouvrage de référence *Graph Theory*. Cette théorie étudie les structures composées de nœuds et d'arêtes : $G = (V, E)$, où les arcs E correspondent à un ensemble fini d'arêtes et V est un ensemble fini de nœuds. Une arête $e \in E$ relie deux sommets, x et y , noté (x, y) . Cela constitue la structure générale des graphes (Diestel, 2017, chap. 1). Dans un graphe, la notion de degré $d(v)$ d'un sommet $v \in V$ correspond au nombre d'arêtes incidentes à ce sommet. Un chemin est une suite de sommets $x_0, x_1, \dots, x_k$ où $x_i, x_{i+1} \in E$ et tous les x_i sont distincts, tandis qu'un cycle est un chemin fermé où $x_0 = x_k$ avec tous les autres sommets qui sont distincts. Un graphe connexe est un graphe tel que tout couple de sommets est relié par un chemin, tandis qu'un graphe complet est un graphe dans lequel tout couple de sommets est relié par une arête (Diestel, 2017, chap. 1).

Il existe deux types de graphes différents : les graphes orientés et les graphes non-orientés. Pour le premier type de graphes, les arêtes ont une direction spécifique qui est représentée par une paire ordonnée (i, j) . Tandis que dans le cas d'un graphe non-orienté, les relations entre les sommets sont symétriques et autorisent une circulation de flux dans les deux sens. Ces structures sont d'un grand intérêt dans le monde mathématique, car elles permettent de représenter un grand nombre de situations, telles que les réseaux de transports, les circuits électroniques, les réseaux sociaux, les foules, etc. (Diestel, 2017, chap. 1). Dans la recherche de parcours optimaux et d'optimisation, la notion d'arbres est également essentielle. Ce

concept fait référence à un graphe connexe qui ne contient aucun cycle, tandis qu'une forêt est un ensemble d'arbres disjoints (Diestel, 2017, chap. 1).

Diestel définit la théorie des flux tels que la circulation d'une quantité à travers les arêtes d'un graphe, dans un contexte de réseau. Tels qu'un flux $f : V^2 \rightarrow \mathbb{Z}$ satisfait l'antisymétrie ainsi que la loi de Kirchhoff :

Loi de Kirchhoff :

$$\sum_{y \in N(x)} f(x, y) = 0$$

Antisymétrie :

$$f(x, y) = -f(y, x)$$

L'antisymétrie signifie que le flux entre deux sommets est orienté. Ainsi, si une certaine quantité de flux circule de x vers y , alors le flux de y vers x est considéré comme son opposé. Tandis que la loi de Kirchhoff exprime la conservation du flux. Ce concept montre le fait qu'il n'y a ni création ni perte de flux au niveau des sommets (hors sources ou puits).

1.8. Problème de conception de réseaux capacitaires

Les problèmes de conception de réseaux capacitaires (CNDP) sont une branche massive de l'optimisation des réseaux. Ces modèles ont pour objectif la création de réseaux optimaux sous contraintes de capacités, tout en minimisant un coût ou un dépassement. Ces problèmes appartiennent à la classe des problèmes d'optimisation linéaire mixte (MILP). Une application fréquente des CNDP consiste à l'activation de certains arcs du graphe. Ces arcs sont soumis à des contraintes de capacités et sont généralement générateurs de coûts. La génération générique d'un CNDP peut être synthétisée comme suit.

Variables

$y_a \in \{0,1\}$: Activation de l'arc a (1 si activé, 0 sinon)

$x_a \geq 0$: Flux de l'arc a nombre de personne par heure

$c_a \geq 0$: Coût de l'arc a en euro

Objectif et contraintes

$$\min \sum_{a \in A} (c_a y_a + k_a x_a)$$

Où c_a correspond au flux d'activation de l'arc et k_a au coût unitaire du flux.

Contrainte de conservation des flux

$$\sum_{a \in \delta^+(i)} x_a - \sum_{a \in \delta^-(i)} x_a = d_i \quad \forall i \in N$$

Où d_i est la demande du nœud i et $\delta^+(i)$ et $\delta^-(i)$ sont les arcs entrant et sortant du nœud i .

Contrainte de capacité

$$x_a \leq Cap_a y_a \quad \forall a \in A$$

Cap_a étant la capacité maximale de l'arc a par heure. Cette contrainte a pour but l'activation des arcs.

Contraintes techniques

$$y_a \in \{0,1\}, \quad x_a \geq 0 \quad \forall a \in A$$

Pour une implémentation en temps réel, sur de grands réseaux, des méthodes de décomposition peuvent être utilisées pour améliorer l'efficacité du modèle. Cette synthèse prend pour hypothèse que chaque nœud du réseau a une capacité suffisante pour permettre la redirection des flux. Dans des conditions réelles, des contraintes physiques ou logistiques peuvent limiter cette capacité. Afin de limiter les coûts tout en respectant les contraintes de capacités, des redirections partielles peuvent être optimales par rapport à des redirections complètes.

Ce type de modèle est fortement utilisé en logistique, réseaux de télécommunication, réseaux énergétiques ou encore pour la gestion des foules. Cette synthèse a été réalisée via les travaux proposés par Khelifi et al. (2017), Holmberg et al. (1997) et Konak et al. (2004).

1.9. Optimisation des flux : cas d'application

L'optimisation des flux sous différents types de contraintes nécessite une approche rigoureuse et basée sur des concepts mathématiques bien définis. L'étude de Rajaram et Ahmadi réalisée en 2003 fournit un exemple d'optimisation linéaire des flux dans un parc à thème. L'objectif de cette étude est d'optimiser les flux de visiteurs entre les différentes attractions d'un parc à thème. Pour parvenir à un modèle fiable, la conservation des flux est un concept indispensable.

Pour modéliser cela, Rajaram et Ahmadi se sont assurés que le nombre total d'entrées F vers une attraction i à un instant t correspond aux flux générés par les autres attractions :

$$\sum_{s=1}^m F_{sit} = \sum_{s=1}^m P_{sit} \cdot S_{s(t-1)} \quad \forall i, t$$

- F_{sit} : le flux de visiteurs allant de s vers i , en temps t .
- P_{sit} : la fraction de visiteurs allant de s vers i en temps t .
- S_{st} : le nombre de visiteurs servis par s à la en temps t .

Après s'être assuré de la conservation du flux de visiteurs entre les attractions, l'article montre également la conservation des flux tout en considérant les nœuds sources, soit les entrées dans le parc. Pour cela, cette contrainte est modélisée :

$$Q_{it} = Q_{i(t-1)} + \sum_{s=1}^m F_{sit} + A_t \cdot P_{it} - S_{it} \quad \forall i, t$$

- Q_{it} : nombre de visiteurs dans la queue de l'attraction i en temps t
- A_t : nombre de visiteurs entrant dans le parc en temps t
- P_{it} : nombre de visiteurs commençant par l'attraction i en temps t

Dans ce contexte, l'une des contraintes les plus importantes permettant de limiter la capacité dans chacune des attractions est la contrainte de capacité modélisée comme cela :

$$S_{it} \leq \sum_{j=1}^n C_{ij} \cdot y_{ijt} \quad \forall i, t$$

- C_{ij} : la capacité maximale j de l'attraction i .
- y_{ijt} : variable binaire qui active la capacité j de l'attraction i au temps t .

D'autres contraintes sont modélisées autour de ces contraintes principales pour répondre aux contraintes spécifiques de ce cas d'application. L'utilisation de ce modèle montre un gain potentiel de 9 % sur le chiffre d'affaires annuel des boutiques du parc. Cela illustre l'importance de l'optimisation de la gestion des flux dans les parcs à thème (Rajaram et al., 2003).

2. Approche

L'approche utilisée pour répondre aux questions de recherche de ce mémoire est une approche quantitative qui est possible grâce aux données mises à disposition par Pairi Daiza. L'approche quantitative permet une analyse objective des données de sorte à avoir un travail précis. Le projet est constitué d'une première partie en apprentissage machine ainsi qu'une seconde en optimisation. Une approche quantitative est donc indispensable pour permettre l'intégration des modèles de prédiction et d'optimisation basés sur un grand nombre de données. Ensuite, les performances des modèles mathématiques peuvent être analysées via des métriques quantitatives, et les résultats du modèle d'optimisation linéaire évalués. La méthodologie de ce travail constitue une base solide et fiable qui a pour but de proposer une solution, fondée sur des données, proposant une suggestion opérationnelle de la gestion des flux de visiteurs vers les restaurants de Pairi Daiza.

3. Machine Learning

3.1. Méthodologie

3.1.1. Collecte des données

Les données utilisées pour ce projet proviennent majoritairement de quatre sources différentes. Premièrement, les données de tickets de vente de chacune des caisses enregistreuses des points de restauration de Pairi Daiza. Elles contiennent l'heure d'achat et le point de vente correspondant. Ces données sont récupérées toutes les heures. Deuxièmement, Pairi Daiza a mis à disposition pour cette étude, le nombre de visiteurs du parc, qui est récupéré tous les jours. Troisièmement, des données météo sont téléchargées via l'API de *meteoblue*. Les éléments récupérés sont la température et les précipitations à la coordonnée géographique du parc. Les prévisions météorologiques de la journée sont téléchargées tous les jours via cette API. Quatrièmement, les dates de congé scolaire de la Wallonie sont obtenues via les API de *nager.date*. Cela inclue tous les jours fériés applicables en Wallonie.

Les données des points de restauration de Pairi Daiza sont récupérées via des composants Copyjobs Fabric² et ajoutées dans un Lakehouse Fabric³. Les données concernant la météo et les congés scolaires sont téléchargés dans ce même Lakehouse via des Notebook Fabric⁴. Le rafraîchissement des données est automatisé et planifié via des Pipelines⁵ dans l’environnement Microsoft Fabric. Ces Copyjobs sont programmés en incrémentale, afin de limiter les coûts liés aux capacités Fabric. En effet, cet environnement de travail est facturé en fonction des capacités utilisées. Lors du premier chargement, les données historiques ont été récupérées à partir de 2020.

3.1.2. Analyse descriptive des données

L’analyse descriptive a pour objectif de comprendre les variations du nombre de commandes par rapport aux autres variables récupérées. Les données de restauration récupérées pour cette analyse descriptive concernent les tranches horaires de pleine affluence, soit entre 12h et 14h. Pour commencer, le nombre de commandes est représenté en fonction des variables temporelles, telles que l’année, le mois, le jour de la semaine ou la tranche horaire.

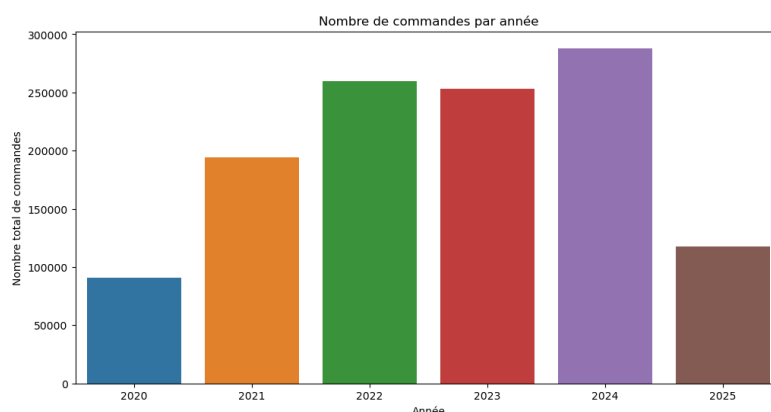


Figure 2 : Évolution du nombre de commandes par année

² Copyjobs Fabric : nom des composants Microsoft Fabric utilisés pour la récupération des données.

³ Lakehouse Fabric : environnement Microsoft Fabric centralisé pour le stockage et la gestion des données.

⁴ Notebook Fabric : interface dans l’environnement de développement Microsoft Fabric

⁵ Pipelines Fabric : chaînes automatisées dans l’environnement Fabric

En représentant le nombre de commandes par an dans les restaurants sélectionnés pour cette étude, on constate une tendance à la croissance depuis 2020. Le Covid-19 étant encore très présent en 2020, les données pour cette année sont nettement inférieures aux données de 2021. Ensuite, on constate une augmentation progressive du nombre de commandes jusqu'en 2024. Cela montre l'intérêt de cette étude qui a pour but de limiter la saturation des restaurants du parc soumis à une demande grandissante. L'année 2025 n'étant pas encore terminée, les données donnent un aperçu des résultats jusqu'en juin. (Voir figure 2)

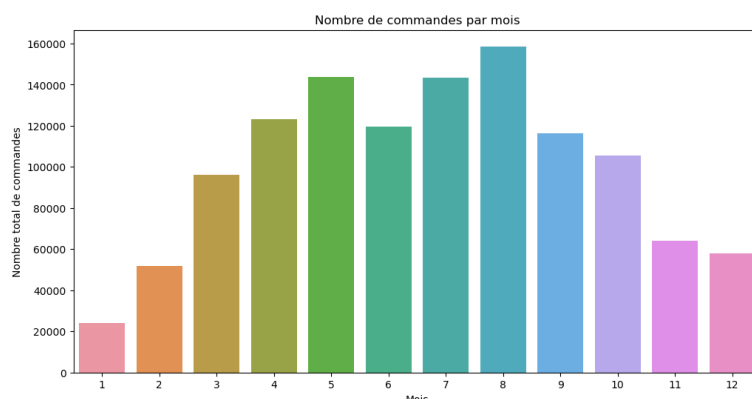


Figure 3 : Comportement du nombre de commandes par mois

Sur la figure 3, qui montre le nombre d'achats réalisés par mois, on peut voir un effet saisonnier, qui augmente depuis janvier pour atteindre un premier pic au mois de mai. Le second pic visible sur ce graphique comprend juillet et août, ce qui correspond aux périodes de congés scolaires. Ensuite, la demande baisse progressivement jusqu'en décembre.

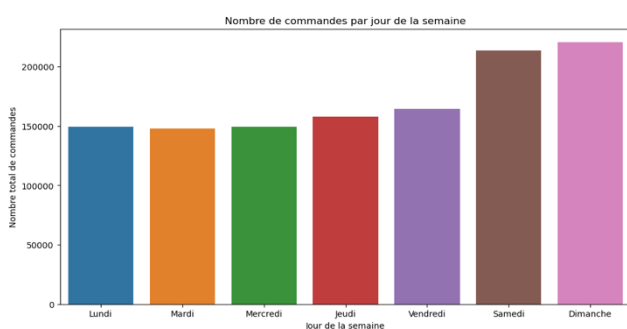


Figure 4 : Comportement du nombre de commandes par jour de la semaine

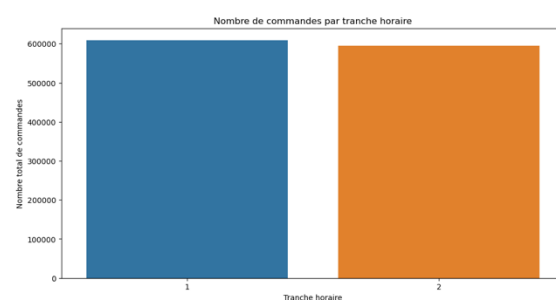


Figure 5 : Comportement du nombre de commandes par tranche horaire

Les données par jour montrent une demande constante en semaine ainsi qu'un nombre de commandes plus important le week-end (voir Figure 4). Concernant les tranches horaires de pointe, soit 12h-13h et 13h-14h, la demande semble être en moyenne similaire (voir Figure 5).

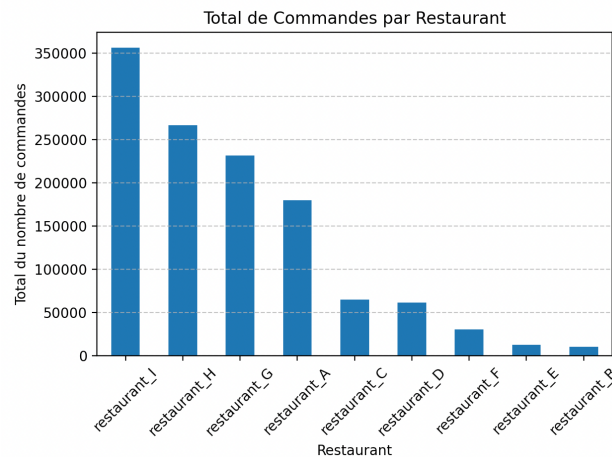


Figure 6 : Nombre de commandes par restaurant

Sur la figure 6, on peut constater que chaque restaurant a une demande différente. Cette disparité peut être due à plusieurs facteurs, tels que le nombre de jours d'ouverture de chaque restaurant, la capacité de chaque restaurant ou la possibilité de prendre son repas à emporter.

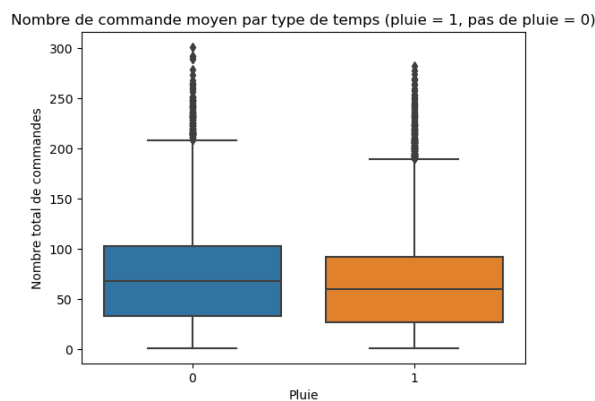


Figure 7 : Comportement du nombre moyen de commandes par type de temps

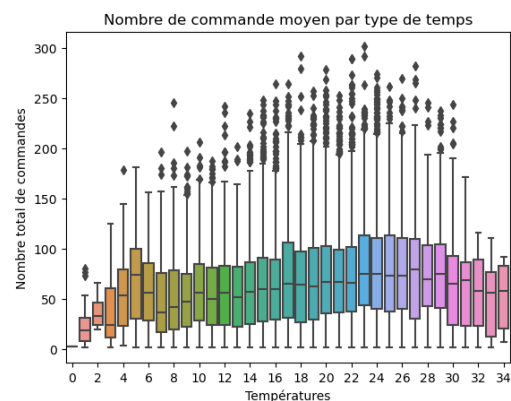


Figure 8 : Comportement du nombre moyen de commandes par température

La météo a également un impact sur le nombre de commandes de chacun des restaurants. Il est facilement remarquable, sur la figure 7 que le nombre total de commandes a tendance à être moins important lors des jours pluvieux. Sur la figure 8 on peut également constater que la température a également tendance à influencer le nombre de commandes des points de vente Horeca. Lors d'une analyse plus approfondie, il a été remarqué que ces tendances se rapprochent fortement du nombre de visiteurs dans le parc. Ce nombre de commandes croissant serait plus lié au nombre de visiteurs dans le parc, qui lui aussi augmente avec les beaux temps. Cependant, il a été jugé pertinent de conserver ces données, car les figures de l'annexe 1 nous montrent que chaque point de vente réagit différemment aux variations météorologiques.

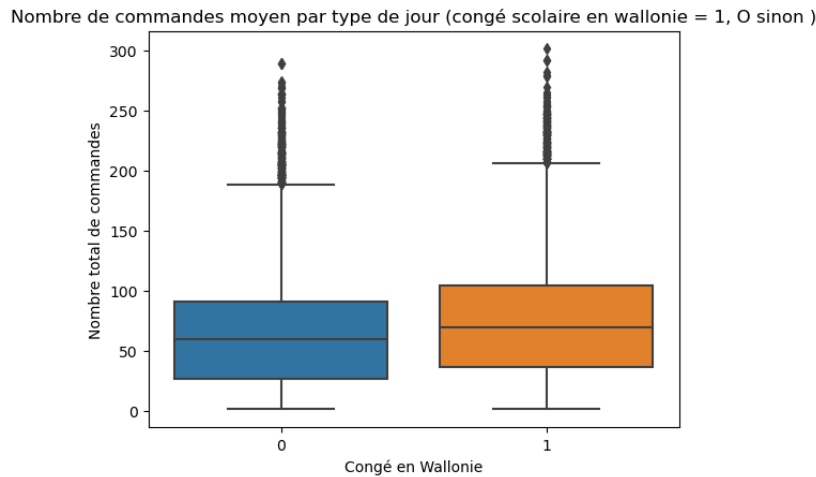


Figure 9 : Comportement du nombre moyen de commandes par type de jours

Les périodes de congé scolaire ainsi que les jours fériés ont un impact positif sur les ventes. Cette valeur est également corrélée au nombre de visiteurs dans le parc. Cependant, il a été constaté que le nombre moyen de commandes par visiteur varie selon le type de jour.

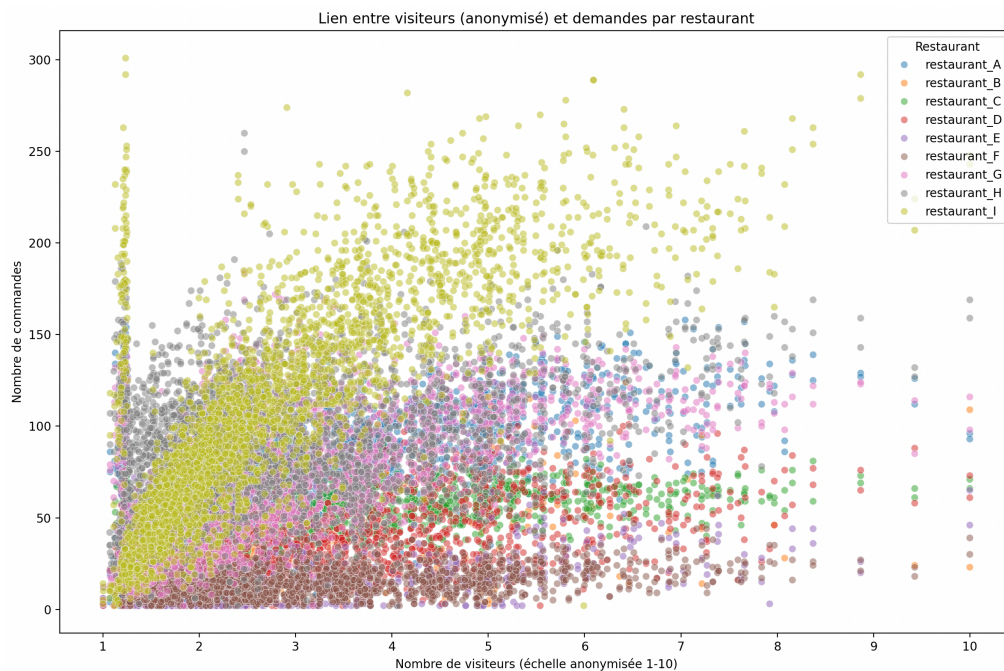


Figure 10 : Graphique combiné du nombre de commandes et nombre de visiteurs

La figure 10 montre la relation entre le nombre de visiteurs et le nombre de commandes dans chacun des points de restauration. Les couleurs de ce graphique montrent la différence des réactions qu'à chaque point de vente au nombre de visiteurs dans le parc. Les figures se trouvant en annexe 2 sont réalisées afin d'avoir une analyse plus précise de chacun des restaurants. À noter que les données ne suivant pas la tendance, au niveau de la valeur 1 sur l'axe des abscisses, sont liées à des événements spéciaux au sein du parc.

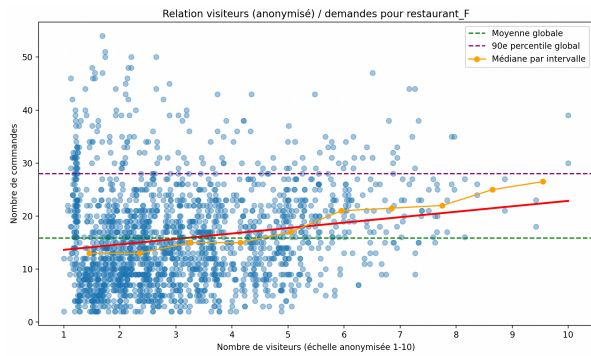


Figure 11 : Comportement spécifique du restaurant F

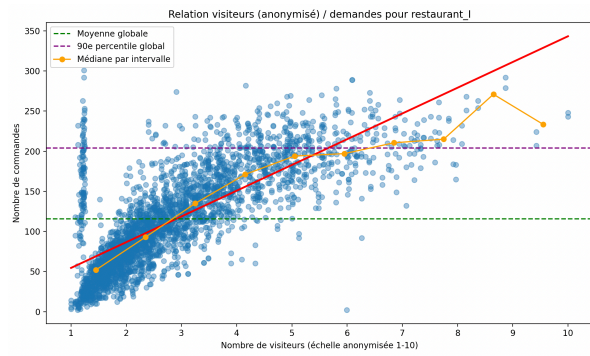


Figure 12 : Comportement spécifique du restaurant I

Sur les figures 11 et 12, on peut voir deux points de restauration qui montrent des réactions différentes à l'augmentation du nombre de visiteurs dans le parc. Pour le restaurant F, la demande a tendance à croître rapidement pour atteindre un maximum aux alentours de 30 commandes par heure. Tandis que le restaurant I montre une augmentation des commandes plus linéaire, jusqu'à atteindre un maximum vers 240 commandes par heure. Ces maximums sont observés pour tous les points de vente et considérés comme leur seuil de saturation. Cette information sera utile dans la partie optimisation de ce travail.

Enfin, pour comprendre les corrélations entre chacune de ces variables, la figure 13 montre la matrice de corrélation de Pearson qui montre les relations linéaires entre variables. On peut constater que la variable VenteTranche1, qui récupère les ventes déjà effectuées dans ce restaurant le même jour, est fortement corrélée à la variable cible Commandes (0,45). Cela étant attendu, car les ventes conservent une même tendance sur une même journée. Ensuite, la variable contenant le nombre de visiteurs montre également une forte corrélation avec notre variable cible (0,30). En effet, les figures 11 et 12 montrent clairement une corrélation entre le nombre de visiteurs du parc et le nombre de commandes effectuées. Ensuite, la température (0,17), la pluie (-0,09), les périodes de congé en Wallonie (0,10) et l'année (0,13) présentent également une corrélation intéressante. Une forte corrélation est également observée entre VenteTranche1 et tranche_heure, ce qui est lié au fait que le nombre de commandes avant l'heure de pointe est quasi nul.

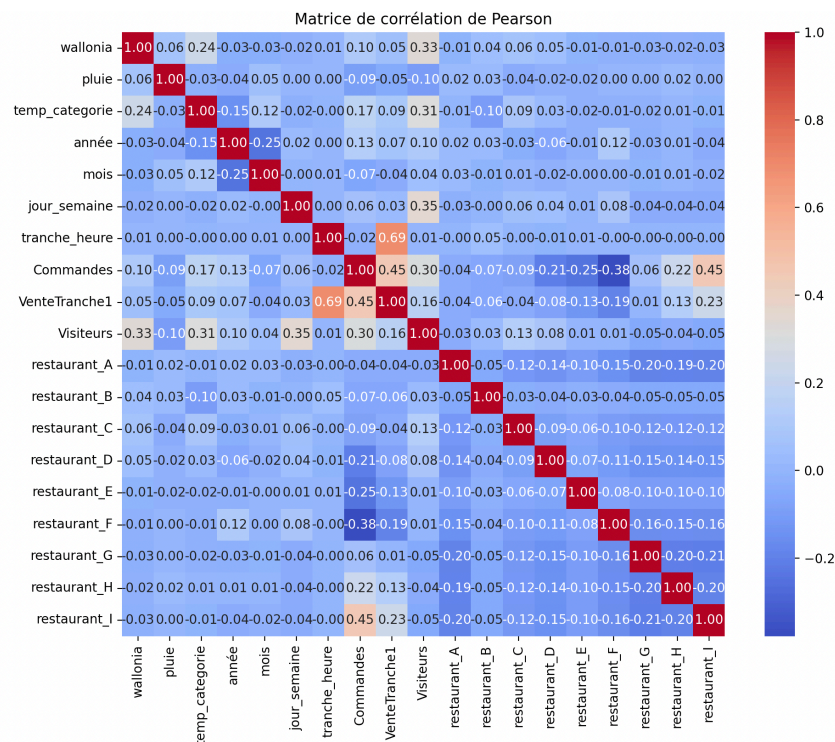


Figure 13 : Matrice de corrélation de Pearson

3.1.3. Transformation et enrichissement des données

Le but de cette préparation de données pour l'apprentissage machine est de préparer une table contenant toutes les variables explicatives prévues pour le modèle prédictif ainsi que la valeur à prédire. Toute cette préparation de données est réalisée dans des notebook Fabric. Ce qui permet de facilement connecter le Lakehouse dans lequel les données sont chargées.

Dans un premier temps, les données sur le nombre de visiteurs sont récupérées et regroupées par jour. Cela permet d'obtenir un tableau de données contenant : la date et le nombre de visiteurs du parc. Ensuite, les données concernant les différents restaurants sont également chargées et groupées par date, période ainsi que le restaurant. Dans le cadre de ce projet, les périodes sont des tranches d'une heure, entre 11h et 14h. Chaque instance correspond donc aux caractéristiques d'un restaurant pour un jour et une tranche d'une heure. Via les données de restauration, un tableau de données est créé contenant les colonnes : date, tranche horaire, restaurant et nombre d'achats. Les tranches horaires correspondent à 1 pour la tranche 12h-13h et à 2 pour la tranche 13h-14h. Ce qui permet de répertorier le nombre de commandes effectuées dans un restaurant, chaque jour, par tranche d'une heure.

Les données météorologiques sont également ajoutées à un tableau de données. Celui-ci contient deux colonnes : temps catégorie et pluie. La première contient la moyenne entre la

température minimale et maximale de la journée, et la deuxième contient une valeur booléenne (1 si les précipitations du jour sont supérieures à 0, et 0 sinon). Enfin, le dernier tableau de données est créé grâce aux données chargées des congés scolaires et jours fériés. Celui-ci contient une colonne Wallonie, qui reprend pour chaque jour la valeur 0 si c'est un jour classique, et 1 s'il correspond à un jour férié ou de congé scolaire.

L'ensemble des tableaux de données sont ensuite joints via la colonne date pour créer une table complète des variables explicatives disponibles pour chaque période. Afin de maximiser la pertinence de l'information contenue dans les données, de l'ingénierie de variable est réalisée pour plusieurs types de données. Les données de dates sont divisées en 3 colonnes : l'année, le numéro du mois (1 = janvier, 2 = février, ...) et le numéro du jour de la semaine (lundi = 0 et dimanche = 6). Ensuite, la colonne VenteTranche1 est créée, leur but étant de répertorier le nombre de commandes faites dans les tranches d'heures précédentes. VenteTranche1 correspond donc au nombre de ventes de la tranche précédente. Enfin, les restaurants sont encodés via une méthode d'encodage binaire, afin d'avoir des variables explicatives binaires et non-catégorielles. Après ces opérations, les données finales utilisées pour l'entraînement des modèles de régression sont disponibles en annexe 3.

3.1.4. Prédiction du nombre de tickets

Dans le cadre de ce projet, les prédictions du nombre d'achats effectués dans chacun des restaurants sont intéressantes pour l'heure qui suit le moment de prédiction. La variable cible est donc un nombre de commandes. La prédiction d'un entier comme dans ce cas nécessite un modèle d'apprentissage machine de régression. Pour l'entraînement et le test des modèles, plusieurs jeux de données sont créés : entraînement (80%) et test (20%). Plusieurs modèles de régression sont testés afin de sélectionner le meilleur pour ce cas d'étude. Ensuite, la librairie Optuna permet de sélectionner la meilleure configuration d'hyperparamètre. Pour évaluer la performance des modèles, plusieurs métriques sont utilisées : l'écart quadratique moyen (RMSE), l'erreur absolue moyenne (MAE) et le coefficient de détermination (R^2). De plus, afin de s'assurer de l'absence de surapprentissage, la fin de cette partie se consacre à une validation croisée.

Une fois le modèle à utiliser validé, sa version finale est créée et entraînée sur l'intégralité des données historiques pour être évaluée sur de nouvelles entrées. Après avoir entraîné le modèle final, la table des données à prédire doit être construite de la même manière que la table d'entraînement. Comme mentionné dans l'introduction de ce travail, l'objectif est de réaliser

cette prédiction avec un horizon d'une heure. La table des variables explicatives nécessaires est donc construite et donnée au modèle une heure avant la période concernée, ce processus étant entièrement automatisé dans un environnement Microsoft Fabric (annexe 4). Dans le cas où les prédictions réalisées montrent de bonnes performances, il est intéressant de les utiliser afin de procéder à des suggestions opérationnelles.

3.2. Résultats

Cette partie est dédiée à l'analyse des résultats des modèles d'apprentissage machine utilisés afin de prédire le nombre de tickets dans un restaurant, pour une tranche d'heure précise. Les données préparées sont utilisées afin d'entraîner plusieurs modèles de régression pour définir le plus adapté pour ce projet. Les modèles de régression testés sont la régression linéaire, la forêt aléatoire, XGBoost ainsi que LightGBM.

Enfin, l'analyse de plusieurs métriques permet de choisir le meilleur modèle pour ce projet. Les métriques étudiées sont : MAE, RMSE, R^2 , le temps d'entraînement et le temps de prédiction. Les deux dernières informations temporelles sont très importantes, car l'environnement de travail de ce projet, Microsoft Fabric, facture à la capacité utilisée. Ce qui signifie qu'en cas de performances très similaires de deux modèles, il est préférable de privilégier celui nécessitant le moins de ressources.

Model	MAE	RMSE	R^2	Temps d'entraînement	Temps de prédiction
Régression Linéaire	14.89	20.22	0.825	32.34ms	14.53ms
Random Forest	10.44	14.88	0.905	4.72s	87.42ms
XGBoost	10.29	14.6	0.909	276.49ms	2.35ms
LightGBM	10.47	14.72	0.907	763.15ms	4.64ms

La régression linéaire est un modèle simple et interprétable qui capture les relations linéaires dans les données, c'est pourquoi ses résultats montrent une moins bonne performance. Le modèle montre une MAE de 14.89 ainsi qu'une RMSE de 20.22. Cependant, le temps de

prédiction est extrêmement faible, ce qui pourrait avoir son importance dans un environnement comme Microsoft Fabric.

Les autres modèles testés se basent sur des arbres et sont donc plus performants pour trouver des relations non-linéaires entre variables. C'est la raison pour laquelle les résultats des trois autres modèles sont nettement meilleurs. La forêt aléatoire a montré une RMSE de 14.88 ainsi qu'une MAE de 10.44 et un R^2 de 0.905. Quant aux temps d'entraînement et de prédiction, ils sont relativement plus longs que pour la régression linéaire. Les deux modèles qui performant le mieux sont XGBoost ainsi que LightGBM, avec une moyenne des erreurs absolues de 10.29 pour XGBoost et 10.47 pour LightGBM, et respectivement 14.6 et 14.72 de RMSE. Ces deux derniers modèles montrent un R^2 presque similaire, avec 0.909 pour XGBoost et 0.907 pour LightGBM. Les temps de prédiction sont relativement plus courts que du résultat de la régression linéaire, avec 2.35ms pour XGBoost et 4.64ms pour LightGBM. Concernant le temps d'entraînement, il est nettement plus court que celui de la forêt aléatoire, avec 276.49ms pour le premier et 763.15ms pour le second.

Le modèle sélectionné pour continuer, grâce à ses excellentes performances, est XGBoost. Il est connu pour être particulièrement efficace sur des ensembles de données de petite à moyenne taille. Robuste face aux valeurs aberrantes, c'est un modèle qui montre une bonne efficacité computationnelle. Dans le cas de cette étude, les résultats montrent, en plus d'une bonne précision, un temps d'entraînement et de prédiction relativement court.

Ensuite, la librairie Optuna permet de déterminer les hyperparamètres du modèle. Les paramètres principaux optimisés sont le nombre d'arbres (*n_estimators*), profondeur maximale de chaque arbre (*max_depth*), le taux d'apprentissage (*learning_rate*), la fraction des échantillons utilisés pour entraîner chaque arbre (*subsample*), la fraction des caractéristiques utilisées pour entraîner chaque arbre (*colsample_bytree*), le gain minimal requis pour qu'un nœud soit divisé (*gamma*). La configuration la plus performante trouvée via Optuna est : *n_estimators* : 845, *max_depth* : 7, *learning_rate* : 0.039, *subsample* : 0.94, *colsample_bytree* : 0.91 et *gamma* : 4

Après l'obtention de ces résultats, il est crucial de prendre des mesures afin d'éviter le surapprentissage du modèle. Il peut conduire à une mauvaise généralisation, où le modèle serait fortement influencé par le bruit. Pour cela, une validation croisée ainsi qu'une évaluation sur un jeu de test sont réalisées. Pour procéder à une validation croisée, les données sont divisées en 5 ensembles, le modèle étant entraîné sur 4 d'entre eux et testé sur le cinquième. Ce

processus est réalisé 5 fois, afin de tester les résultats obtenus pour chacun des ensembles. Les résultats obtenus montrent une stabilité des métriques peu importe les données utilisées : une MAE de 10.03 ± 0.12 , un RMSE de 14.38 ± 0.25 et un R^2 de 0.912 ± 0.004 .

Ensuite, ces résultats sont comparés avec le test sur un jeu indépendant. Pour cela, l'intégralité des données se divise en 2 sets : entraînement (80%) et test (20%). Les résultats obtenus sur ce dernier test avec les paramètres optimaux sont extrêmement similaires aux résultats de validation croisée : une MAE de 9.880, un RMSE de 14.004 et un R^2 de 0.916.

En conclusion, on peut assumer que le modèle ne souffre pas de surentraînement et généralise correctement sur de nouvelles données. De plus, les résultats des prédictions sont relativement précis, avec un R^2 de 0.916, il peut être envisageable d'exploiter les prédictions pour réaliser des suggestions opérationnelles au parc. Cependant, il est important de mentionner de potentielles sources de biais ou limitations qui peuvent entraver une bonne prédiction. C'est le cas d'un changement soudain du comportement des visiteurs ou des événements imprévus. Ce modèle se base également sur les hypothèses de conformité des visiteurs qui stipule que les visiteurs continueront à se comporter de la même manière que dans les données d'entraînement.

Un avantage du modèle XGBoost est sa simplicité à obtenir l'importance de chaque variable dans la prédiction du modèle. La figure 14 montre l'importance relative de chacune des variables explicatives, à l'exception du type de restaurant. On constate l'importance de récupérer les ventes précédemment réalisées en début de journée pour obtenir une prédiction de qualité. Le nombre de visiteurs et la tranche horaire concernée sont également indispensables au modèle, suivis des caractéristiques temporelles, météorologiques ainsi que des jours fériés.

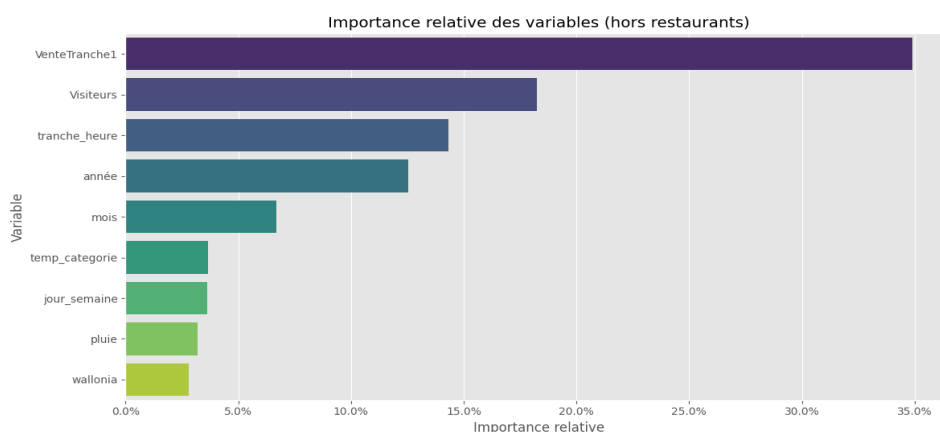


Figure 14 : Importance des variables explicatives dans la prédiction de XGBoost

La figure 15 montre la comparaison des résultats réels, avec leur prédiction pour une période tirée au hasard. Ces résultats confirment la précision mise en avant par les métriques obtenues. Il est intéressant de constater que le modèle prédit correctement une baisse et une hausse importante du nombre de commandes.

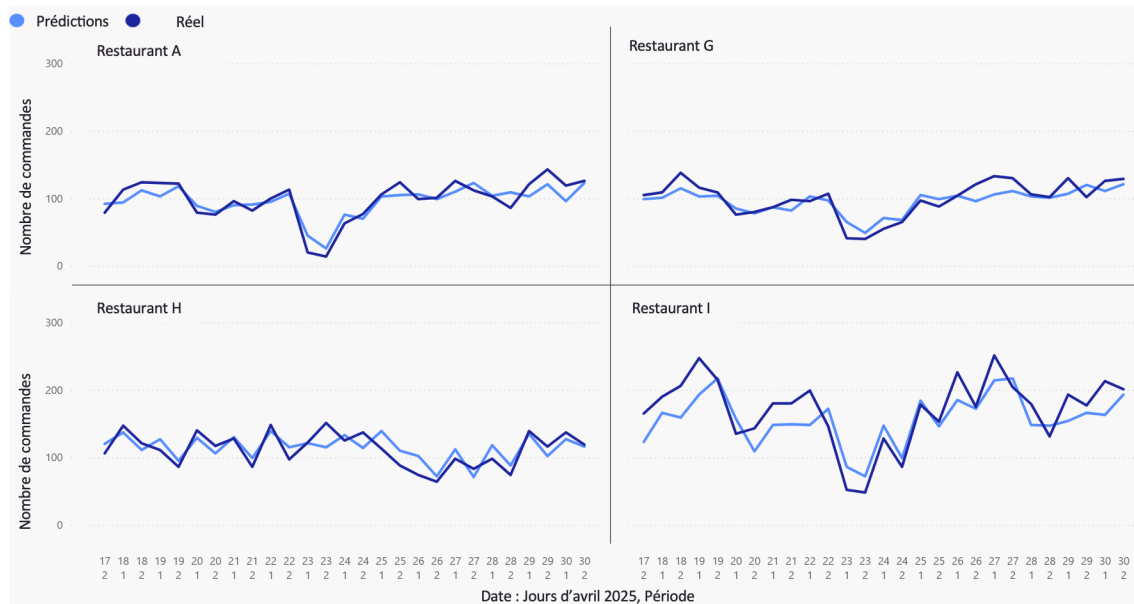


Figure 15 : Visualisation des prédictions finales

3.3. Vers une stratégie d'optimisation

Les prédictions obtenues grâce à XGBoost, montrent une précision satisfaisante dans l'anticipation du nombre de commandes par point de vente. Ces résultats permettent une visibilité anticipée de l'affluence attendue au sein des différentes zones. Par exemple, on peut analyser une carte du parc divisée en différentes zones, colorée en vert si la zone est sous le seuil de saturation et qui vire au rouge pour les zones saturées. L'outil Power Bi a permis la réalisation de la carte en annexe 5. Cependant, cette prédiction ne permet pas de régler le problème de saturation. Pour remédier à cette saturation, il est nécessaire de combiner ces prédictions avec une stratégie d'optimisation. Cela permet de procéder à des suggestions opérationnelles de réorientation optimale du flux de visiteurs. C'est dans cette logique que s'inscrit la seconde partie de ce mémoire qui a pour but la modélisation d'un modèle mathématique d'optimisation, visant à trouver les réorientations optimales au sein du parc, pour améliorer l'expérience du visiteur.

4. Optimisation

4.1. Méthodologie

4.1.1. Optimisation du nombre de visiteurs

Après avoir prédit le nombre de commandes futures dans un restaurant, pour une tranche d'une heure, il est possible, avec les seuils de saturation, d'optimiser le flux de visiteurs au sein des restaurants. Le but de cette partie est de construire un modèle mathématique qui autorise la réorientation de visiteurs d'un restaurant à un autre, avec pour but d'optimiser l'affluence au sein de chaque point de restauration. Cependant, ces réorientations doivent respecter une série de contraintes. Pour cela, le modèle est modélisé en langage mathématique et implémenté en Mosel. Les données de restauration depuis 2020 sont utilisées pour évaluer les performances des modèles sur des données historiques. Le but étant de voir quelle aurait été la situation si la solution optimale du modèle avait été utilisée pour toutes les données historiques. Cela met en avant une comparaison de la situation avant et après utilisation du modèle.

4.1.2. Présentation du modèle de base

Cette partie est dédiée à la présentation du modèle d'optimisation créé pour le cas de Pairi Daiza. Le rôle de ce modèle est de minimiser le dépassement de capacité de chaque restaurant tout en respectant des contraintes opérationnelles. Le modèle est un problème d'optimisation linéaire en nombres entiers (MILP). Un tableau récapitulatif des notations se trouve en annexe 6.

Indices et ensembles

i, j : Indices désignant les restaurants

V : Ensemble de tous les restaurants

V_i : Ensemble des restaurants voisins au restaurant i

Paramètres

C_i : Capacité maximale du restaurant i

D_i : Demande du restaurant i

A_{ij} : Possibilité de redirection du restaurant i vers le restaurant j

R_{max} : Nombre maximal de redirections activées

Variables de décision

f_{ij} : Variable indiquant le flux de visiteurs du restaurant i vers j

y_{ij} : Variable binaire indiquant si la redirection de i vers j est activée

o_i : Variable indiquant le dépassement positif de capacité du restaurant i

Fonction objective et contraintes

La fonction objective a pour but de minimiser le dépassement total de capacité o pour chacun des restaurants i :

$$\min \sum_{i \in V} o_i$$

Contraintes de redirections :

$$(C1) \quad y_{ij} \leq A_{ij} \quad \forall i, j \in V: i \neq j$$

$$(C2) \quad \sum_{j \in V} (y_{ij} + y_{ji}) \leq 1 \quad \forall i \in V$$

La première contrainte a pour objectif d'activer la possibilité d'un flux entre i et j , seulement si la matrice d'adjacence le permet. Sachant que A_{ij} est la matrice d'adjacence des sommets qui contient 1 si la redirection de i vers j est possible et 0 sinon. La seconde contrainte garantit qu'un restaurant i ne peut pas à la fois recevoir un flux et en envoyer un.

Contraintes de capacité :

$$(C3) \quad \sum_{i \in V} \sum_{j \in V, j \neq i} y_{ij} \leq R_{max}$$

$$(C4) \quad f_{ij} \leq D_i \cdot y_{ij} \quad \forall i, j \in V: i \neq j$$

La troisième contrainte autorise un nombre défini maximal de réorientation R_{max} . Le rôle de cette contrainte est de contrôler le nombre d'interventions simultanées afin d'éviter des flux de réorientations trop nombreux. La quatrième contrainte limite le flux sortant d'un restaurant i ,

afin qu'il ne dépasse pas la demande initiale de i . Cette contrainte assure également l'annulation du nombre de redirections f_{ji} si la variable y_{ij} n'est pas activée.

Contraintes de flux :

$$(C5) \quad \sum_{j \in V_i} f_{ji} \leq \max\{0, 0,95 \cdot C_i - D_i\} \quad \forall i \in V$$

$$(C6) \quad o_i \geq D_i + \sum_{j \in V_i} f_{ji} - \sum_{j \in V_i} f_{ij} - C_i \quad \forall i \in V$$

La cinquième contrainte limite le flux total de réorientation f_{ji} dans le but d'éviter un dépassement de capacités du restaurant qui reçoit un flux. Le nombre de redirections vers un restaurant j est donc au maximum égal à la différence entre 95% de la capacité du restaurant j et sa demande. Dans le cas où le restaurant j est saturé, soit que sa demande excède sa capacité, le flux arrivant sur j est nul. La sixième contrainte définit le dépassement de capacité pour chaque restaurant (o_i). Ce dépassement correspond à la somme des flux réorientés vers le restaurant i ($\sum_{j \in V_i} f_{ji}$), ajouté à la demande du restaurant i (D_i), à quoi est soustraite la capacité du restaurant i (C_i), ainsi que la somme des flux sortants du restaurant i ($\sum_{j \in V_i} f_{ij}$).

Contrainte de non-négativité et d'intégralité :

$$(C7) \quad o_i \geq 0 \quad \forall i \in V$$

$$(C8) \quad f_{ij} \geq 0 \quad \forall i, j \in V : i \neq j$$

$$(C9) \quad y_{ij} \in \{0,1\} \quad \forall i, j \in V : i \neq j$$

La septième contrainte est une contrainte de non-négativité qui assure une prise en compte dans la fonction objective, seulement des dépassements positifs de capacité, pour chacun des restaurants. Cela permet de ne pénaliser dans la fonction objective uniquement lorsqu'un restaurant dépasse sa capacité maximale. Les deux dernières contraintes du modèle s'assurent que les flux sont toujours positifs et que la variable d'activation y_{ij} soit binaire.

Ce modèle d'optimisation permet de proposer la réorientation optimale des visiteurs dans les restaurants du parc. Les résultats renvoyés par le modèle sont les flux optimaux de visiteurs entre les différents restaurants, en tenant compte des possibilités de redirections. La solution

optimale du modèle propose la meilleure combinaison permettant de minimiser les dépassements de capacité dans chacun des restaurants.

4.2. Résultats

Cette partie est dédiée à la présentation des résultats du modèle d'optimisation, qui prend en paramètre la prédiction du modèle d'apprentissage machine, dont l'objectif de minimiser le dépassement de capacité des points de restauration. Ce modèle d'optimisation a pour but de proposer les réorientations optimales de visiteurs entre les différents restaurants pour optimiser l'utilisation des ressources disponibles et améliorer l'expérience du client. La figure 16 illustre le réseau connexe utilisé dans ce modèle d'optimisation, qui permet de définir les réorientations activables.

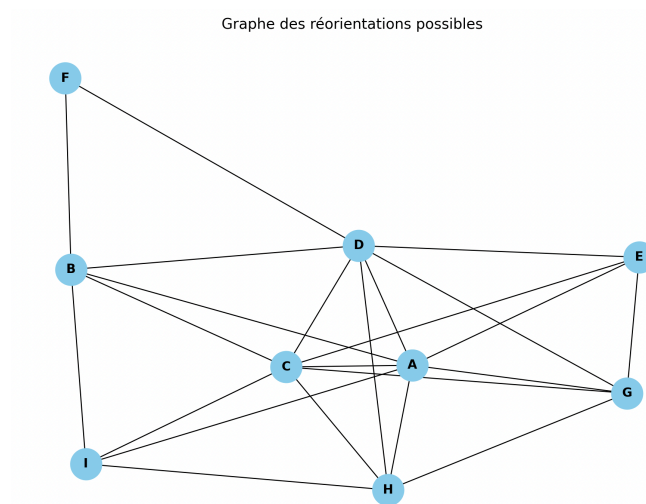


Figure 16 : Réseau des points de restauration

4.2.1. Scénarios spécifiques

Pour le scénario où tous les restaurants du parc montrent une demande supérieure à leur capacité maximale, le modèle ne propose pas de réorientation. Cela est modélisé via la cinquième contrainte, qui empêche l'activation d'un flux vers un restaurant dont la saturation est déjà dépassée. Ensuite, le cas où un seul restaurant montre un dépassement de capacité peut être étudié. Le test suivant permet de constater la capacité des redirections pour chaque restaurant en fixant la demande des autres restaurant à 80% de leur capacité maximale. Ensuite, l'algorithme d'optimisation est utilisé afin de trouver le moment où les redirections, autorisées par le modèle, ne permettent plus d'amortir le dépassement. Cela permet de définir quels sont les points de restauration où le dépassement de capacité est le plus problématique. Chaque

restaurant a sa propre disposition dans le parc, ce qui exerce une influence sur la matrice d'adjacence et donc bloque ou autorise certaines réorientations.

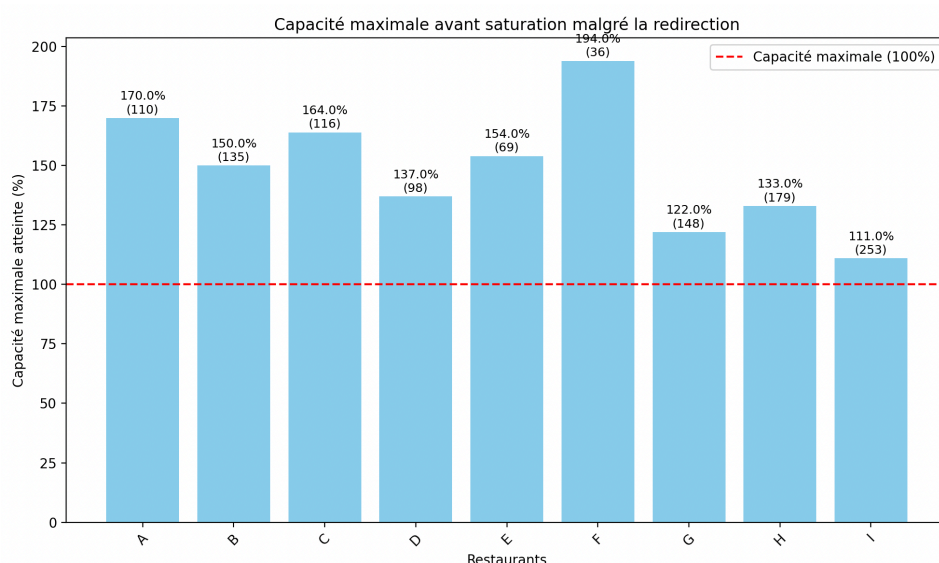


Figure 17: Capacité de dépassement maximale de chaque restaurant

La figure 17 montre, pour une demande dans chacun des restaurants de 80% de leur capacité, quel est le taux de saturation maximum si seul ce restaurant est saturé. Les points de restauration I et G sont, d'après le modèle, les lieux montrant des marges de manœuvre les plus petites. En effet, ce visuel montre qu'au-delà de 111% pour I et de 122% pour G de capacité maximale, le lieu sera inévitablement saturé. À l'inverse, les points de restauration A, C et F acceptent dans ce cas un dépassement important. En observant cela, il pourrait être intéressant de modifier le modèle afin de permettre les réorientations d'un restaurant vers plusieurs différentes zones. En effet, si le restaurant I se trouve dans une zone dans laquelle il y a un événement au sein du parc, la demande de ce point de vente risque d'augmenter drastiquement. Ce cas de redirections multiples est étudié dans la partie variation du modèle. Enfin, dans le cas où aucun restaurant ne serait saturé, le modèle ne propose pas de réorientations, car l'objectif du modèle est de minimiser le dépassement. Une variante qui a pour objectif de lisser le taux d'occupation des restaurants est étudiée dans la partie variation du modèle.

4.2.2. Effets sur les données historiques

Le modèle mathématique est testé à l'aide des données historiques des dernières années via son implémentation en Mosel. Grâce à ces tests, il est possible de quantifier l'amélioration que le modèle aurait eue sur les données historiques. Les résultats nous montrent une nette amélioration du nombre d'heures saturées, en passant de 1009 à 313, si seule une redirection

par tranche d'heure est autorisée ($K_{\max} = 1$), ce qui constitue une diminution de 69% des moments saturés. La figure 18 montre quels sont les résultats du modèle pour différents paramétrages de $K_{\max}$. Plus le nombre de réorientations est autorisé, au mieux est le nombre de périodes de dépassement. Pour un nombre de 2 flux activés, la situation s'améliore de nouveau, avec seulement 192 périodes saturées. Au-delà de 2 réorientations, la situation ne s'améliore plus que très peu, pour atteindre un minimum à 4 redirections. Ces résultats posent la question du nombre de flux à activer.

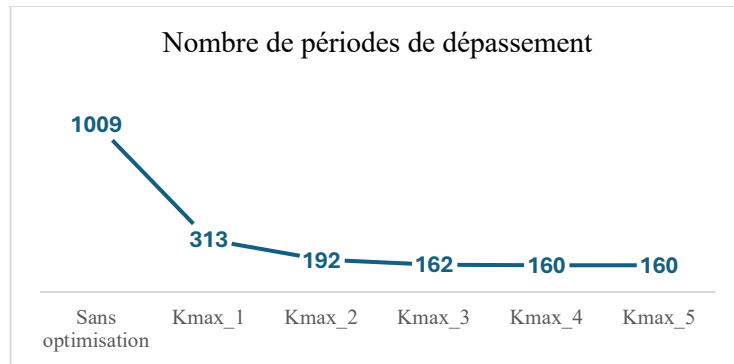


Figure 18 : Nombre de périodes de dépassement pour le modèle de base

Pour parvenir à cette amélioration, le modèle propose une série de réorientations de visiteurs. La figure 19 montre quels sont toutes les réorientations proposées par le modèle lorsque $K_{\max}$ est fixé à 1. Sur cette figure on constate que les restaurants H et I sont les points de vente qui s'échangent le plus de visiteurs.

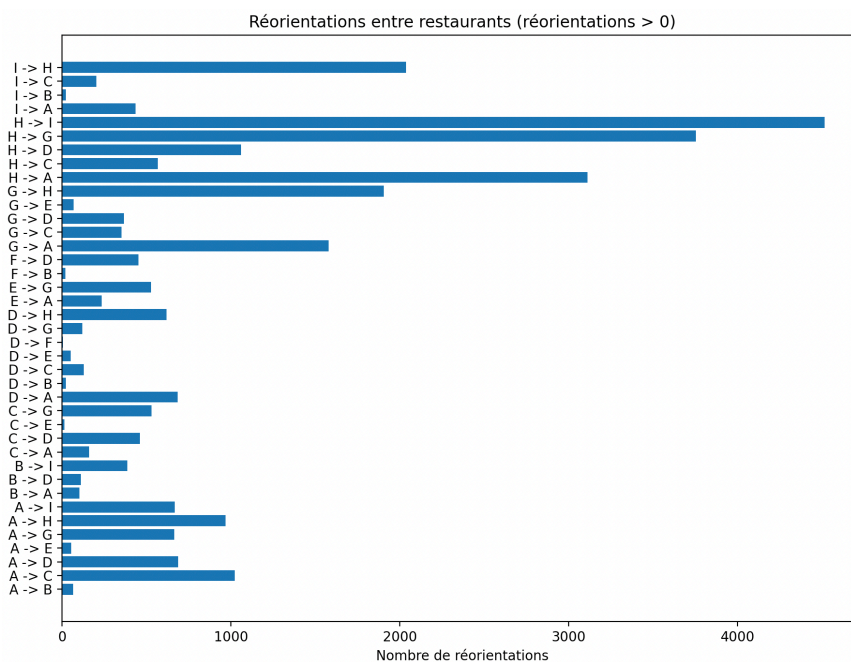


Figure 19 : Réorientations proposées par le modèle de base ($K_{\max} = 1$)

Ces réorientations permettent non seulement de réduire le nombre de jours de saturation pour chacun des restaurants, mais parfois aussi d'alléger la saturation d'un point de vente. En effet, le modèle a pour objectif de minimiser le dépassement de capacité, cela signifie que même lorsque le dépassement ne peut pas être supprimé, il est réduit. Le graphique en figure 20 montre la quantité totale de dépassements, avant et après optimisation, pour chacun des restaurants. À noter que le modèle permet, une amélioration du nombre de tranches horaires saturés, mais également, lorsqu'un restaurant est toujours saturé, une importante baisse de sa saturation. Pour le cas $K_{\max} = 1$, une analyse plus approfondie a montré que la moyenne des dépassements pour les restaurants saturés après le modèle passe de 25,4 commandes à 12 commandes. Cela illustre bien le fait que le modèle a un rôle indispensable dans la minimisation du dépassement même lorsqu'il ne peut pas être réglé par le modèle.

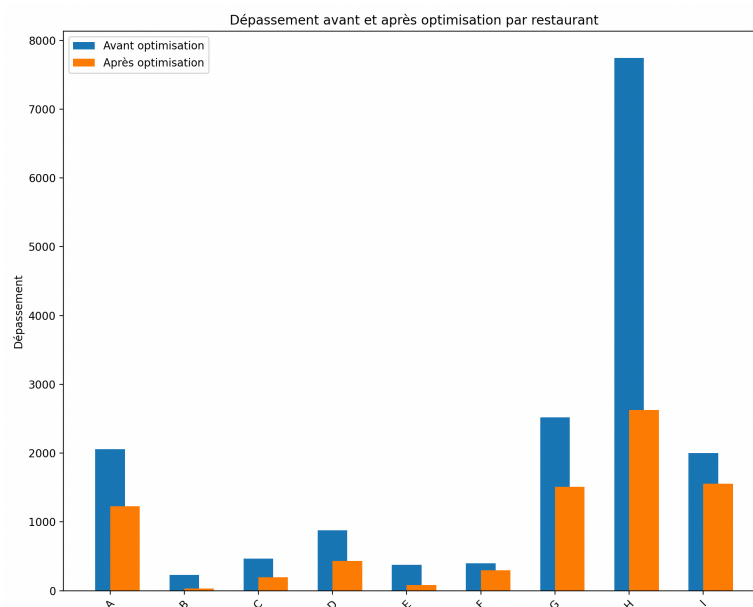


Figure 20 : Amélioration du dépassement par restaurant avec le modèle de base

Certains points de vente réagissent très bien au modèle, en diminuant drastiquement leur dépassement (H, E ou B), alors que d'autres, malgré les réorientations autorisées par le modèle, ne s'améliorent que très peu (F ou I). Ce phénomène est partiellement dû aux contraintes de redirection mises en place par le modèle. En effet, pour certains points de vente, la redirection est plus difficile que pour d'autres, car ils se situent géographiquement plus loin dans le parc. Dans le modèle, cela se traduit par la matrice d'adjacence qui active seulement quelques possibilités de redirection. De plus, le modèle n'autorise que la redirection d'un point de vente vers un seul autre. Une possibilité est de permettre à un restaurant de réorienter vers différents

endroits, afin de favoriser le dégorgement d'un lieu. Ces différentes variations sont étudiées dans le point suivant.

4.2.3. Variations du modèle

Multi-réorientations

Pour commencer, la contrainte 2 est adaptée afin de permettre à un point de vente de réorienter vers un ou plusieurs lieux. Un point de vente est donc capable de multi-réorienter vers plusieurs restaurants. Cela, tout en maintenant la contrainte qu'un restaurant ne peut pas avoir de flux sortant s'il reçoit des visiteurs réorientés. L'intégration d'une variable s_i au modèle et la division de la contrainte 2 en plusieurs équations permettent de répondre à cette variation.

$$(C2) \quad \sum_{j \in V_i} (y_{ij} + y_{ji}) \leq 1 \quad \forall i \in V$$

Adaptée en :

s_i : Variable prenant 1 si le restaurant i redirige vers un autre, 0 sinon

$NbrRéorientation$: Paramètre fixant le maximum de flux sortant d'un restaurant

$$(C2.1) \quad \sum_{j \in V_i} y_{ij} \leq NbrRéorientation \cdot s_i \quad \forall i \in V$$

$$(C2.2) \quad \sum_{j \in V_i} y_{ji} \leq 1 - s_i \quad \forall i \in V$$

Permettre à un restaurant de réorienter vers plusieurs points de vente différents facilite le désengorgement. Appliquer cela sur les données historiques permet de descendre le nombre de périodes saturées jusqu'à 133. La figure 21 montre les performances de cette variation du modèle sur les données historiques, pour les différents nombres de réorientations simultanées. Permettre au modèle de réorienter un restaurant vers plusieurs points de vente n'a qu'un faible effet positif sur les résultats.

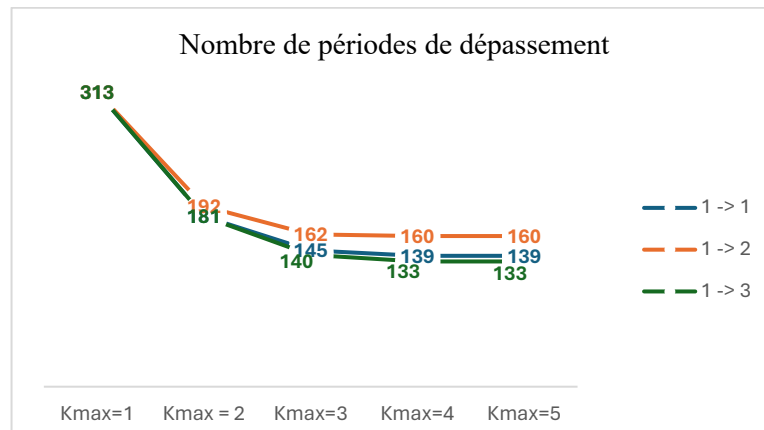


Figure 21 : Nombre de périodes de dépassement pour le modèle de multi-réorientation

L'un des facteurs les plus limitants de ce modèle est la matrice d'adjacence qui ne permet que quelques réorientations. Le point suivant est dédié à l'activation de toutes les possibilités de la matrice d'adjacence, pour obtenir un graphe complet.

Graphe complet

Au sein du parc, chaque restaurant a son emplacement spécifique. C'est pourquoi, dans le modèle de base est définie une matrice d'adjacence permettant d'activer ou pas un flux de réorientations entre points de restauration. Cette matrice n'autorise que les redirections vers un restaurant proche à moyennement loin. Le cas suivant étudie les résultats des données historiques dans le cas où l'on autoriserait aussi les réorientations plus longues.

Les résultats nous montrent une baisse importante du nombre de périodes saturées. La figure 22 nous montre une baisse de périodes saturées pour $K_{\max} = 1$ de 45 périodes dans ce cas. Ensuite, en augmentant le nombre de redirections simultanées, le nombre de périodes saturées chute jusqu'à 96, ce qui correspond à moins de 10% des moments saturés sans optimisation. Ces périodes saturées restantes correspondent très probablement à des jours où le parc est fortement fréquenté. Pour aller encore plus loin, il est possible de combiner les deux scénarios essayés, soit l'activation de toutes les arêtes du graphe, ainsi que les multi-réorientations. Ce cas extrême est testé dans le but de voir à quel point il est possible de minimiser le nombre de périodes saturés via un modèle largement moins contraignant. Sur la figure 22, on peut voir une légère amélioration par rapport au graphe complet sans multi-réorientations. Il serait possible d'ouvrir le modèle davantage en autorisant un plus grand nombre de réorientations, cependant, cette étude a pour but d'améliorer les jours saturés en restant dans des solutions

envisageables. Permettre plus de réorientations sort donc du cadre de cette étude. De plus, après d'autres analyses de la moyenne des dépassements des différents cas, aucune variation flagrante n'est à constater. Le dépassement moyen varie entre 12 et 14 commandes après utilisation du modèle, face à 25,4 commandes avant optimisation.

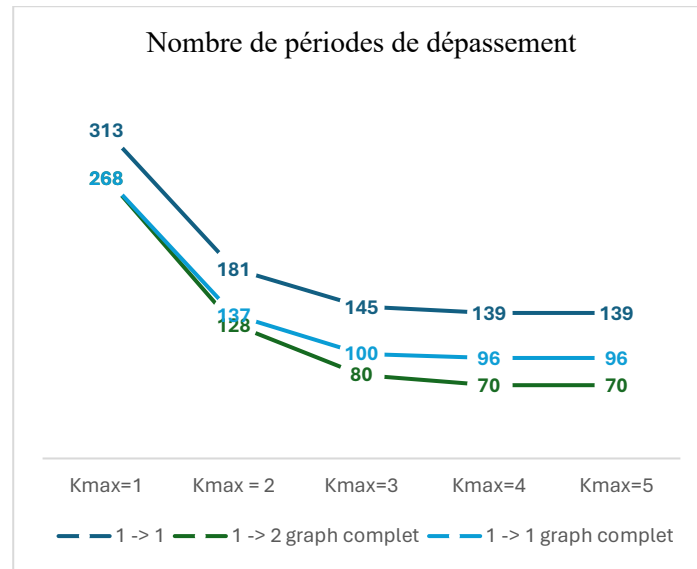


Figure 22 : Nombre de périodes de dépassement pour le modèle avec graphe complet

Lissage

Le but de cette variation est de lisser la capacité d'utilisation des points de restauration qui ne sont pas saturés. La priorité du modèle reste de minimiser les dépassements de capacités, tout en lissant la capacité d'utilisation des points de vente si aucune capacité n'est dépassée. Cela a pour objectif d'améliorer l'expérience du client en évitant les grosses disparités de fréquentation. Le modèle est donc passé en multi-objectifs avec une partie gérant les dépassements de capacité, ainsi qu'une autre s'occupant du lissage. La deuxième partie de l'objectif est pondérée de sorte à ce qu'elle ne s'active qu'en l'absence de saturation.

Nouvel objectif :

$$\min \sum_{i \in V} o_i + \text{tauxOcc}$$

Nouvelle contrainte liée à l'objectif :

$$\text{C6b)} \quad \text{tauxOcc} \geq \frac{D_i + \sum_{j \in V_i} f_{ji} - \sum_{j \in V_i} f_{ij}}{C_i} \quad \forall i \in V$$

La nouvelle contrainte C6b permet à tauxOcc de prendre une valeur supérieure ou égale au taux d'occupation des restaurants. Cette valeur est ensuite minimisée dans la fonction objective. Cela a pour impact de minimiser le taux d'occupation du restaurant le plus proche de son seuil de saturation et a donc pour conséquence le lissage des seuils de saturation. Pour l'évaluation de cette variation du modèle, seulement les périodes ne comprenant pas de dépassement sont utilisées. Les résultats de ce lissage se montrent particulièrement intéressants. La figure 23 montre le taux d'occupation maximal des points de vente, avant et après optimisation. La zone rouge montre un pic entre 80% et 100%, ce qui met en avant l'importance de lisser l'occupation des restaurants. En effet, cela signifie que lorsqu'aucun restaurant n'est saturé, il est tout de même fort probable qu'un point de vente soit occupé au-delà de 80% de sa capacité. La zone bleue montre l'impact qu'a cette variation de modèle sur les données. Avec une répartition suivant une distribution quasi-normale autour de 60%, le modèle se montre extrêmement efficace pour le lissage.

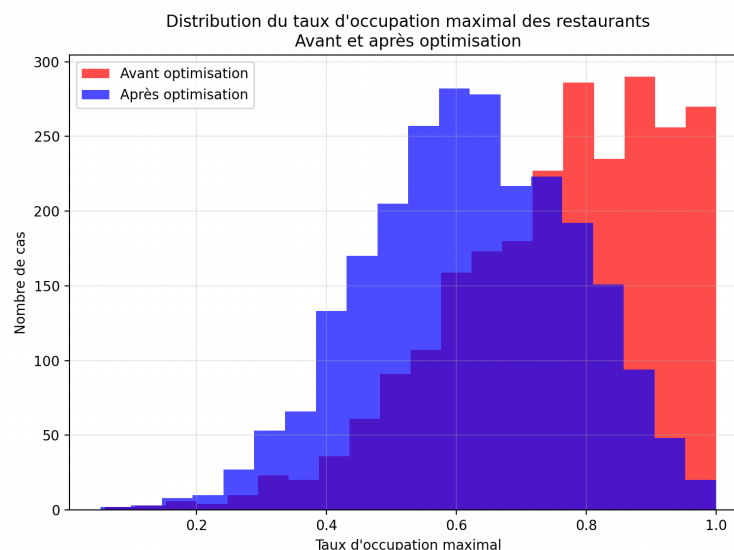


Figure 23 : Taux d'occupation maximal avec lissage

L'étude de cette variation de modèle montre l'importance de travailler sur le lissage de l'occupation même dans le cas où il n'y a pas de saturation. En effet, il est plus confortable pour les visiteurs de se rendre dans une zone occupée à 60% que dans une zone à la limite de la saturation.

5. Discussion

5.1. Interprétation des résultats

Les résultats obtenus dans les deux différentes parties de ce mémoire, démontrent une approche intéressante de la gestion des flux de visiteurs. La première partie de ce travail se concentre sur la prédiction du nombre de commandes au sein de chacun des restaurants de Pairi Daiza. Ce concept de parc à thème complique la prédiction d'affluence des visiteurs au sein du parc à cause du comportement aléatoire des déplacements (Kataoka et al., 2005). Pour faire face à ces difficultés, les données historiques des points de vente du parc ont été récupérées dans un but d'entraînement de modèles de régression. Après plusieurs tests sur différents modèles de régression, le modèle XGBoost a été sélectionné sur la base de ses excellents résultats ($R^2 = 0,916$) ainsi que de son efficacité computationnelle. En effet, ce modèle est particulièrement utilisé pour capter des relations complexes dans un contexte de prédiction de demande (Kang et al., 2022 ; Zhou et al., 2025 ; Hobor et al., 2025).

La deuxième partie de ce travail a montré l'importance d'une approche proactive dans la gestion des flux de visiteurs. Grâce à des contraintes d'optimisation linéaire basées sur la théorie des graphes, le modèle montre une diminution de 69% des périodes saturées, ainsi qu'une amélioration des 31% restants. La création de flux entre différents points névralgiques est l'une des applications de la théorie des graphes qui permet de travailler dans un graphe orienté sous contraintes de conservation des flux (Diestel, 2017, chap. 1).

Plusieurs variations du modèle ont été étudiées afin de voir quelles sont les solutions envisageables pour le parc. Les variations testées avaient pour but d'évaluer les bénéfices du modèle si un plus grand nombre de redirections était autorisé. Avec une nette amélioration lors de l'activation d'un seul flux, la question se pose du nombre de réorientations à activer simultanément. Pour répondre à cette question, il est intéressant de se concentrer sur les techniques de redirection envisageables : la signalisation dynamique au sein du parc, des offres ciblées ainsi que l'installation d'éléments de décor. (Clavé, 2023c). Le choix du nombre de réorientations à activer va donc dépendre de la capacité opérationnelle du parc à mettre en place une ou plusieurs stratégies, mais aussi de leur acceptabilité par les visiteurs.

De la même manière, cette étude propose des variations, comme l'activation de redirections d'un restaurant vers plusieurs différentes zones, ou l'activation du graphe complet malgré les

distances entre les points de restauration. Bien que la réorientation d'un restaurant vers plusieurs ne semble pas montrer de résultats indiquant une grande amélioration, l'activation du graphe complet montre un net bénéfice. Cependant, cela nécessiterait de demander au client de se déplacer et de patienter pour obtenir son repas, ce qui provoquerait de l'insatisfaction (Tang et al., 2022). Selon Jin et al. (2011), la tolérance à l'attente du visiteur dépend du lieu ainsi que du contexte dans lequel il se trouve. Dans ce cas bien précis, il pourrait être imaginable de réorienter les visiteurs via un service de petit train. Cela créerait un environnement où, comme le disent Jin et al. , la tolérance à l'attente serait fortement augmentée.

Enfin, la satisfaction du visiteur est liée à la congestion (Yuan et al., 2018 ; Jin et al., 2011). C'est pourquoi la dernière variation du modèle étudiée est le lissage du taux d'occupation des restaurants. Cette variation montre une amélioration nette des résultats après optimisation. Dans le cas où aucun restaurant n'est saturé, là où avant optimisation, le restaurant le plus saturé avait tendance à être rempli au-delà de 80%, après optimisation, les résultats montrent une saturation maximale moyenne aux alentours de 60%. Cette variation est proposée dans le cas où le parc souhaiterait procéder à l'amélioration de l'expérience du client, au-delà des périodes saturées, ce qui impacterait positivement la compétitivité du parc (Ma et al., 2013).

Dans ce contexte de redirection de visiteurs vers des zones non saturées, il est indispensable d'éviter les oscillations de foule (Brown et al., 2013; Kataoka et al., 2005). C'est la raison pour laquelle la seconde partie de ce mémoire a été réalisée, car le but est de quantifier le nombre de réorientations optimal. Sur base de la suggestion opérationnelle contenant le nombre de visiteurs à réorienter, le parc s'assure de ne pas déplacer la saturation dans un autre point de restauration, ce qui provoquerait des oscillations de foule.

5.2. Limites méthodologiques

Cette méthodologie rencontre plusieurs limites. L'une d'entre elles est la capacité du modèle de régression à prédire le nombre de visiteurs dépassant la capacité des restaurants. En effet, le modèle s'entraîne sur les données historiques des caisses de chaque point de vente. Cette capacité a donc été absorbée. Le modèle prédictif renvoie donc la capacité qui pourrait être prise en charge par chacun des restaurants. Pour pallier cela, l'utilisation de caméras de comptage est une solution, car les données montreraient donc la quantité réelle de visiteurs à proximité des restaurants. Cette technologie sera implémentée par le parc, ce qui permettra une meilleure estimation des dépassements de capacité.

De plus, les données utilisées dans le modèle de prédiction ne considèrent pas la taille des commandes. Or, une commande familiale sera inévitablement plus consistante et donc plus longue à traiter que la commande d'un individu. Les commandes ont également tendance à varier de taille en fonction de la période de l'année concernée. Tout cela provoque des variations des capacités maximales des restaurants. Pour plus de précision, il pourrait être intéressant de considérer des seuils de saturation variables.

Enfin, la combinaison de prédictions suivies d'optimisation n'est pas sans risques. En effet, les erreurs provoquées par l'algorithme prédictif sont utilisées dans le modèle d'optimisation. Cela peut influencer les résultats du modèle d'optimisation qui, par conséquent, se basera sur des erreurs. Pour limiter ces erreurs, il est fondamental de sélectionner le meilleur modèle de régression pour le projet, ainsi que de le fournir avec des données de qualité. Malgré ces limites, le projet montre un potentiel d'amélioration de la gestion des flux de visiteurs au sein du parc.

5.3. Perspectives

Bien que de nombreuses variables différentes aient déjà été prises en compte dans le modèle prédictif, d'autres types de données pourraient être intéressants à implémenter, tels que les événements, les promotions ou fermetures au sein du parc. Cela permettrait de mieux représenter la situation du parc au moment de la prédiction, afin de capter de nouvelles variations contextuelles et d'améliorer les prédictions. D'autres modèles plus complexes d'apprentissage machine pourraient également être testés, tels que des extensions de XGBoost comme proposé dans l'étude réalisée par Kang et al. (2022).

Pour aller plus loin dans ce projet, comme mentionné dans les limites de ce travail, il est intéressant d'utiliser des données provenant de caméras de comptage. Ceci afin de mieux déterminer le nombre réel de personnes se trouvant à proximité des points de restauration.

Concernant les réorientations de visiteurs, procéder à une étude comparative pour trouver quelles sont les méthodes les plus adaptées à la situation, permettrait d'aller plus loin dans cette analyse.

Enfin, bien que la prédiction d'affluence soit intéressante sur une perspective d'une heure, il pourrait être envisageable de construire un modèle capable de réaliser une prédiction avec un horizon d'une semaine. Ce qui laisserait au parc le temps de s'organiser et d'anticiper les affluences.

Conclusion

Ce mémoire avait pour objectif de répondre à deux enjeux cruciaux du parc Pairi Daiza : anticiper l'affluence de clients au sein des points de restauration de Pairi Daiza, dans le but d'optimiser la gestion du flux de visiteurs. La solution proposée se concentre sur l'amélioration de deux concepts : l'expérience du visiteur, ainsi que l'amélioration de l'efficacité opérationnelle, qui mène à une meilleure gestion des ressources du parc. Cette meilleure gestion des ressources, ainsi que l'amélioration de l'expérience du client, optimisent la rentabilité du parc.

Pour atteindre cet objectif, ce mémoire s'est divisé en deux parties complémentaires : la prédiction d'affluence de clients au sein des différents points de vente, suivie de l'optimisation proactive des flux de visiteurs, via un modèle d'optimisation mathématique. Cette combinaison de techniques a permis, en plus d'un simple modèle prédictif, de proposer une suggestion opérationnelle optimale basée sur des prédictions. L'utilisation combinée de techniques novatrices telles que l'apprentissage machine et de l'optimisation mathématique permet de proposer une solution concrète et fiable pour aider la prise de décision opérationnelle.

Dans la première partie de ce mémoire, dédiée à la création d'un modèle de régression, plusieurs modèles ont été évalués. Le plus performant pour cette étude s'est révélé être le modèle de régression XGBoost. Celui-ci a montré des résultats prometteurs, avec un R^2 de 0,916 et des temps d'entraînement et de prédiction intéressants. Ces résultats ont été validés via une validation croisée, afin de s'assurer de la bonne généralisation du modèle.

La seconde partie de ce mémoire a consisté à la conception d'un modèle d'optimisation linéaire. Un ensemble de contraintes a été modélisé afin de répondre aux contraintes opérationnelles du parc. Les résultats de cette partie nous ont montré l'efficacité du modèle sur la base des simulations effectuées sur les données historiques. En effet, le modèle de base a mis en avant une réduction des moments saturés de 69%. Afin d'avoir une analyse complète et rigoureuse, plusieurs variations du modèle ont été analysées. Parmi ces variantes, le cas d'un lissage du taux d'occupation des restaurants s'est montré prometteur.

Cette combinaison de méthodes propose deux différents livrables qui permettent l'amélioration de la prise de décision opérationnelle. Premièrement, la capacité de visualiser la saturation future des différentes zones de restauration du parc. Cela permet, avant tout, de réagir

proactivement à cette information. Deuxièmement, en se basant sur ces prédictions, le modèle d'optimisation linéaire offre une suggestion opérationnelle des réorientations de visiteurs optimales à faire au sein du parc durant la prochaine heure. Le parc peut ensuite se baser sur ces livrables pour mettre en place différentes méthodes concrètes de réorientation des visiteurs. Par exemple, la mise en place de signalétiques interactives, des réductions ciblées ou des transports prévus entre différentes zones.

Enfin, cette combinaison de méthodes comporte des risques et a ses limites. En basant le modèle d'optimisation sur des prédictions, les résultats s'exposent à de la propagation d'erreur. C'est la raison pour laquelle il est important d'avoir un modèle de prédiction robuste et précis. En perspective future de ce travail, il peut être intéressant de se baser sur des données provenant de caméras de comptage, d'étudier les meilleures stratégies de réorientation, ou encore, d'appliquer la méthodologie de ce mémoire à un horizon plus grand qu'une heure.

Pour conclure, cette méthodologie démontre une approche puissante et pertinente pour répondre aux enjeux que constituent les foules au sein des parcs à thème. Ce travail constitue la base solide d'une approche proactive et basée sur les données pour la gestion de flux de visiteurs. De plus, la méthodologie développée est applicable et généralisable à tout espace clos, comportant des points de vente.

Références

- Anton Clavé, S., Carlà-Uhink, F., & Freitag, F. (2023a). Economic Strategy : Conceptual, Customer-Based, and Environmental, Social, and Governance Strategies in the Theme Park Economy. In F. Freitag, F. Carlà-Uhink, & S. Anton Clavé (Éds.), *Key Concepts in Theme Park Studies : Understanding Tourism and Leisure Spaces* (p. 97-111). Springer International Publishing. https://doi.org/10.1007/978-3-031-11132-7_5
- Anton Clavé, S., Carlà-Uhink, F., & Freitag, F. (2023b). Space : Representing, Producing, and Experiencing Space in Theme Parks. In F. Freitag, F. Carlà-Uhink, & S. Anton Clavé (Éds.), *Key Concepts in Theme Park Studies : Understanding Tourism and Leisure Spaces* (p. 261-275). Springer International Publishing. https://doi.org/10.1007/978-3-031-11132-7_13
- Anton Clavé, S., Carlà-Uhink, F., & Freitag, F. (2023c). Visitors : The Roles of Guests as Customers, Pilgrims, Fans, Performers, and Bodies in the Theme Park. In F. Freitag, F. Carlà-Uhink, & S. Anton Clavé (Éds.), *Key Concepts in Theme Park Studies : Understanding Tourism and Leisure Spaces* (p. 323-339). Springer International Publishing. https://doi.org/10.1007/978-3-031-11132-7_16
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Introduction To Tree Classification. In *Classification and Regression Trees* (p. 2-41). Chapman and Hall/CRC.
- Brown, A., Kappes, J., & Marks, J. (2013). Mitigating Theme Park Crowding with Incentives and Information on Mobile Devices. *Journal of Travel Research*, 52(4), 426-436. <https://doi.org/10.1177/0047287512475216>
- Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>

- Diestel, R. (2017). The Basics. In R. Diestel, *Graph Theory* (Vol. 173, p. 1-34). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-53622-3_1
- Géron, A. (2019). End-to-End Machine Learning Project. In *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : Concepts, tools, and techniques to build intelligent systems* (2^e éd.). O'Reilly Media, Inc.
- Hobor, L., Brcic, M., Polutnik, L., & Kapetanovic, A. (2025). *Comparative Analysis of Modern Machine Learning Models for Retail Sales Forecasting* (No. arXiv:2506.05941). arXiv. <https://doi.org/10.48550/arXiv.2506.05941>
- Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE) : When to use them or not. *Geoscientific Model Development*, 15(14), 5481-5487. <https://doi.org/10.5194/gmd-15-5481-2022>
- Holmberg, K., & Yuan, D. (1997). A Lagrangean Heuristic Based Branch-and-Bound Method for the Capacitated Network Design Problem. In U. Zimmermann, U. Derigs, W. Gaul, R. H. Möhring, & K.-P. Schuster (Éds.), *Operations Research Proceedings 1996* (p. 78-83). Springer. https://doi.org/10.1007/978-3-642-60744-8_15
- Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning : A Brief Primer. *Behavior Therapy*, 51(5), 675-687. <https://doi.org/10.1016/j.beth.2020.05.002>
- Jin, Q., & and Pearce, P. (2011). Tourist Perception of Crowding and Management Approaches at Tourism Sites in Xi'an. *Asia Pacific Journal of Tourism Research*, 16(3), 325-338. <https://doi.org/10.1080/10941665.2011.572667>
- Kang, J., Guo ,Xingyu, Fang ,Lei, Wang ,Xiangrong, & and Fan, Z. (2022). Integration of Internet search data to predict tourism trends using spatial-temporal XGBoost composite model. *International Journal of Geographical Information Science*, 36(2), 236-252. <https://doi.org/10.1080/13658816.2021.1934476>

- Kataoka, T., Kawamura, H., Kurumatani, K., & Ohuchi, A. (2005). Distributed Visitors Coordination System in Theme Park Problem. In T. Ishida, L. Gasser, & H. Nakashima (Éds.), *Massively Multi-Agent Systems I* (p. 335-348). Springer.
https://doi.org/10.1007/11512073_25
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM : A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30.
https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
- Khelifi, M., Saidi, M., & Boudjit, S. (2017). Solving the capacitated network design problem in two steps. *Advances in Science, Technology and Engineering Systems Journal*, 291-301.
<https://doi.org/10.25046/aj020339>
- Kim, J., & Kim, S. (2024). A Study on Estimating Theme Park Attendance Using the AdaBoost Algorithm Based on Weather Information from the Korea Meteorological Administration Web. *Journal of Web Engineering*, 869-884. <https://doi.org/10.13052/jwe1540-9589.2368>
- Konak, A., & and Smith, A. E. (2004). Capacitated network design considering survivability : An evolutionary approach. *Engineering Optimization*, 36(2), 189-205.
<https://doi.org/10.1080/03052150310001633223>
- Ma, J., Gao, J., Scott, N., & Ding, P. (2013). Customer delight from theme park experiences : The Antecedents of Delight based on Cognitive Appraisal Theory. *Annals of Tourism Research*, 42, 359-381. <https://doi.org/10.1016/j.annals.2013.02.018>
- Xin, Y., & Su, X. (avec World Scientific (Firm)). (2009). *Linear regression analysis : Theory and computing*. World Scientific Pub. Co.

Yuan, Y., & Zheng, W. (2018). How to Mitigate Theme Park Crowding? A Prospective Coordination Approach. *Mathematical Problems in Engineering*, 2018(1), 3138696.

<https://doi.org/10.1155/2018/3138696>

Zhou, L., Zhu, Q., Chen, Q., Wang, P., & Huang, H. (2025). Predicting hospital outpatient volume using XGBoost : A machine learning approach. *Scientific Reports*, 15(1), 17028.

<https://doi.org/10.1038/s41598-025-01265-y>

6. Annexes

Annexe 1 : Réaction de la demandes dans les restaurants aux conditions météorologiques

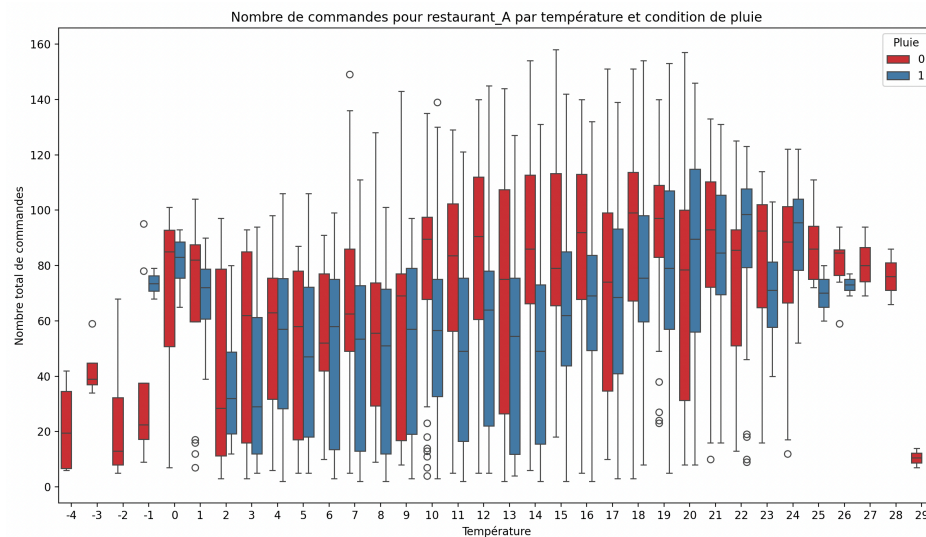


Figure 24 : Réaction du restaurant A aux conditions météorologiques

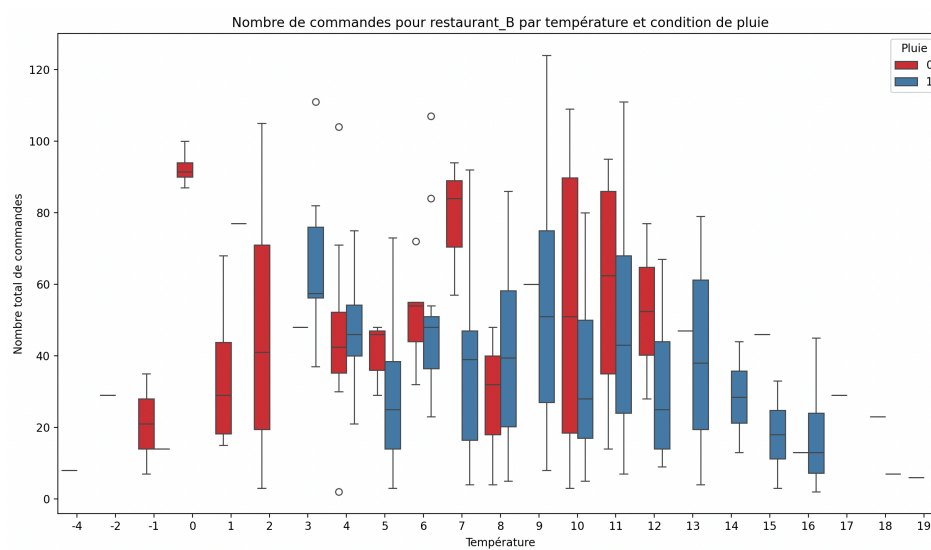


Figure 25 : Réaction du restaurant B aux conditions météorologiques

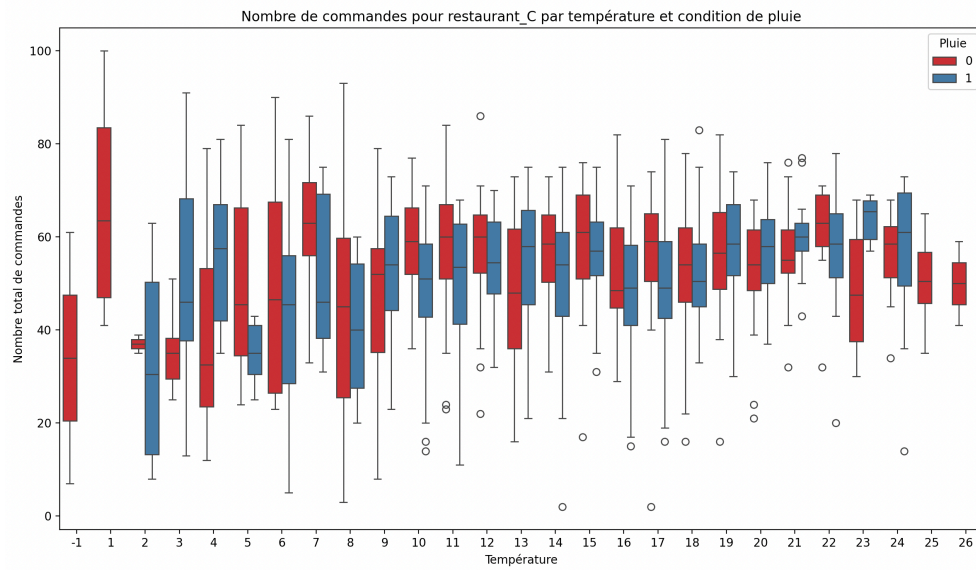


Figure 26 : Réaction du restaurant C aux conditions météorologiques

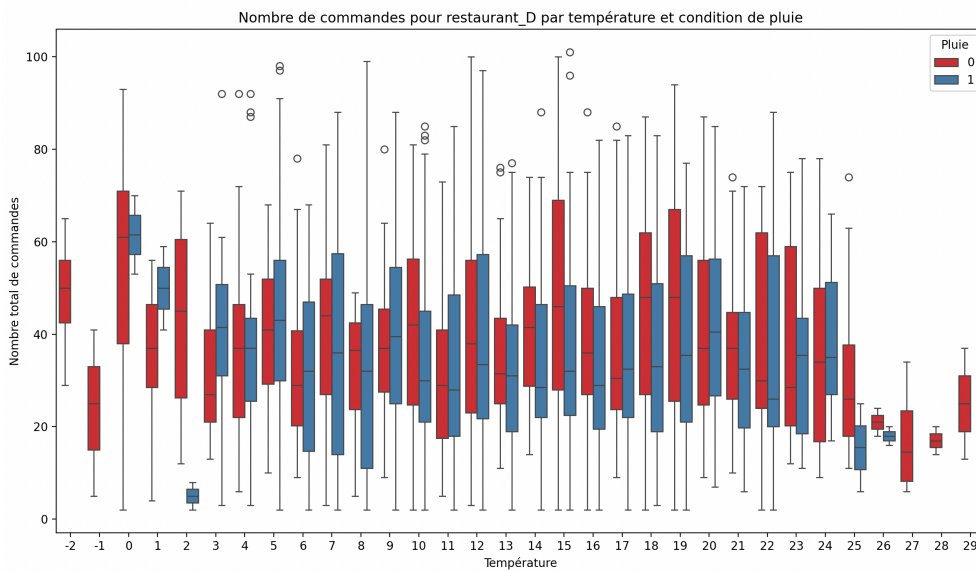


Figure 27 : Réaction du restaurant D aux conditions météorologiques

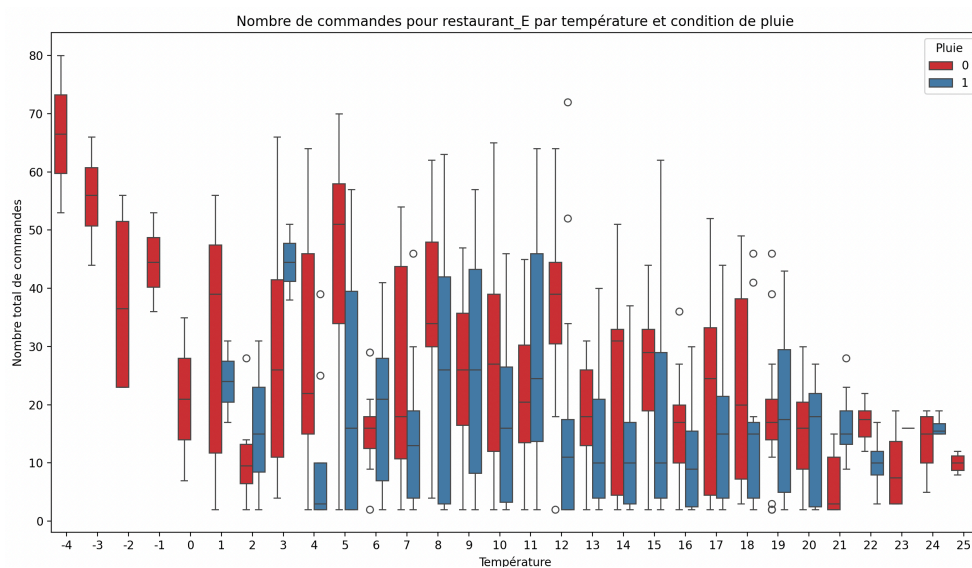


Figure 28 : Réaction du restaurant E aux conditions météorologiques

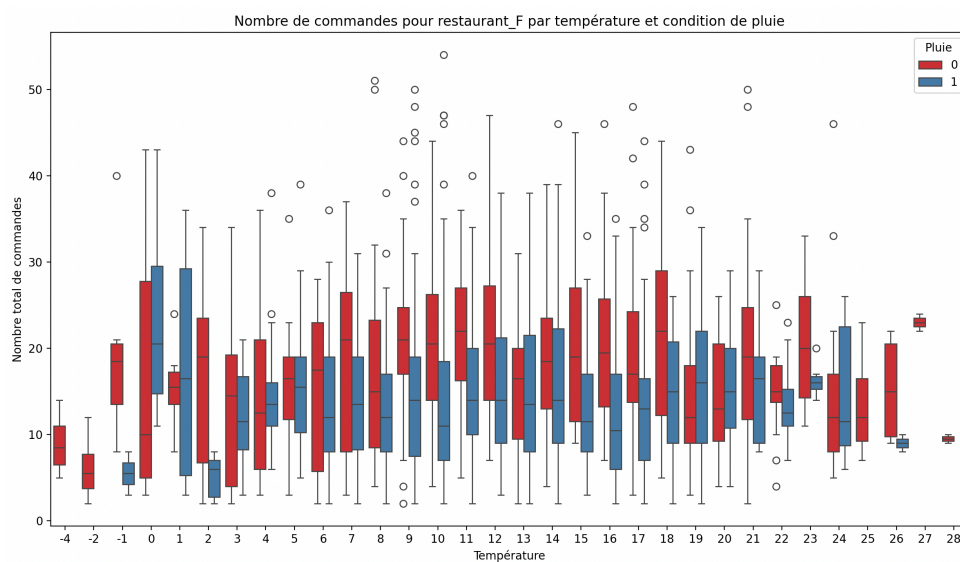


Figure 29 : Réaction du restaurant F aux conditions météorologiques

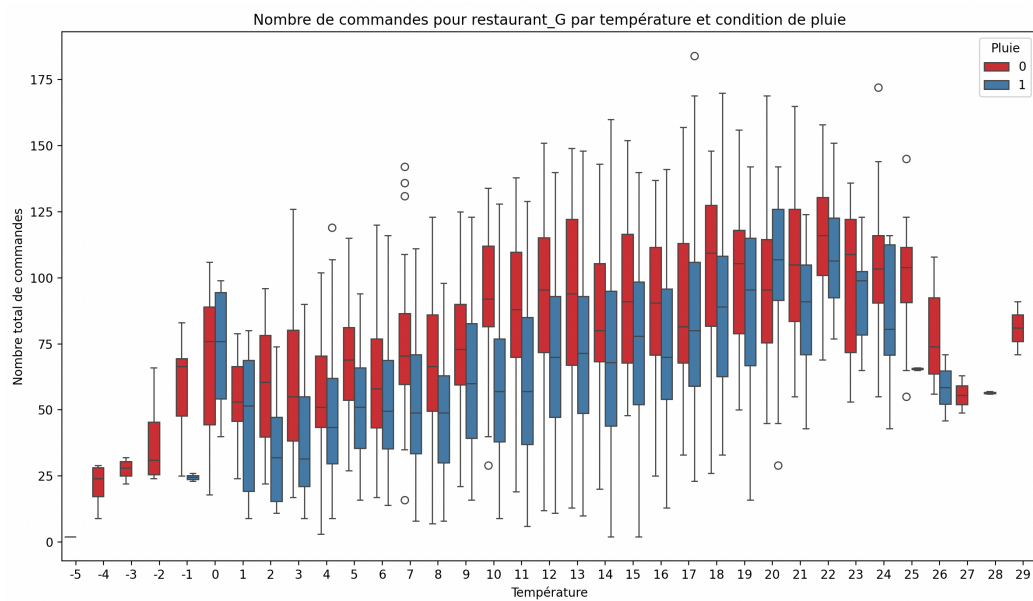


Figure 30 : Réaction du restaurant G aux conditions météorologiques

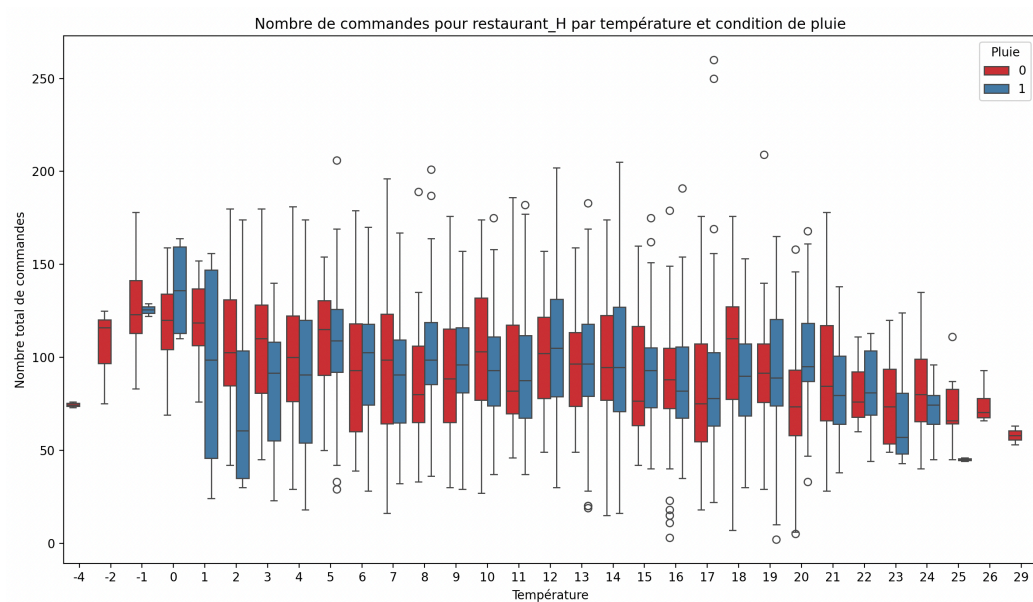


Figure 31 : Réaction du restaurant H aux conditions météorologiques

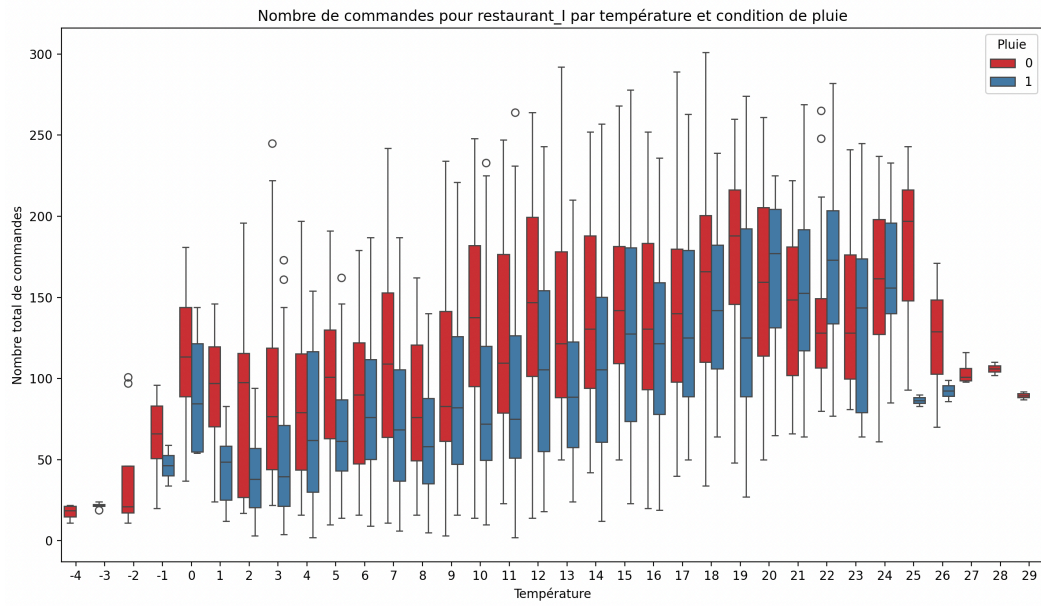


Figure 32 : Réaction du restaurant I aux conditions météorologiques

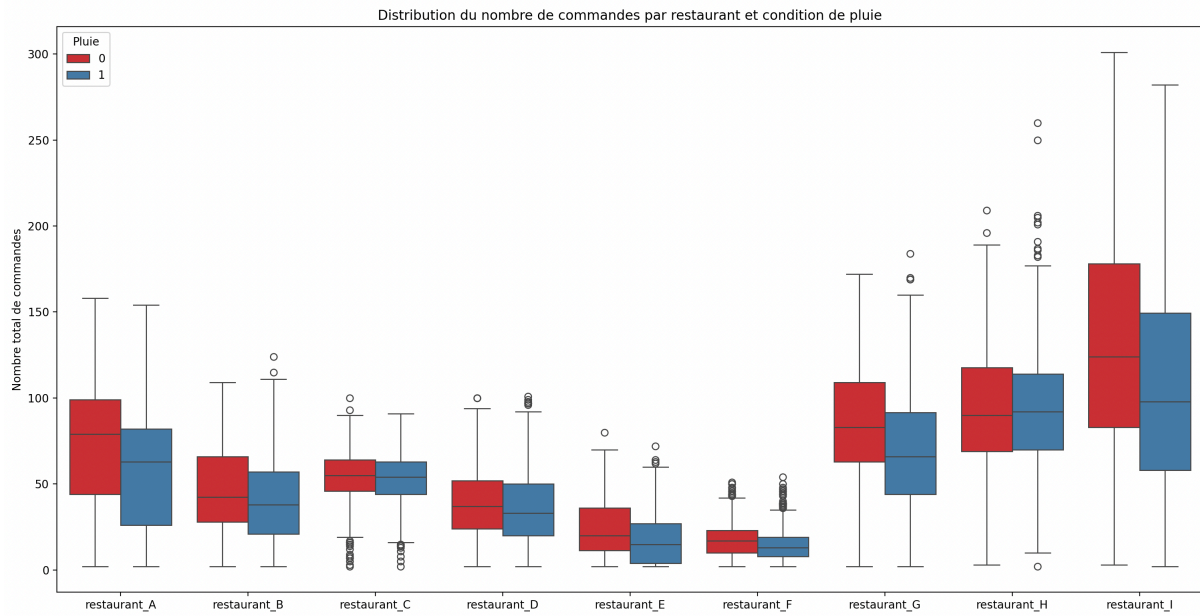


Figure 33 : Comportement de chaque restaurant à la pluie

Annexe 2 : Réaction de la demandes dans les restaurants aux conditions météorologiques

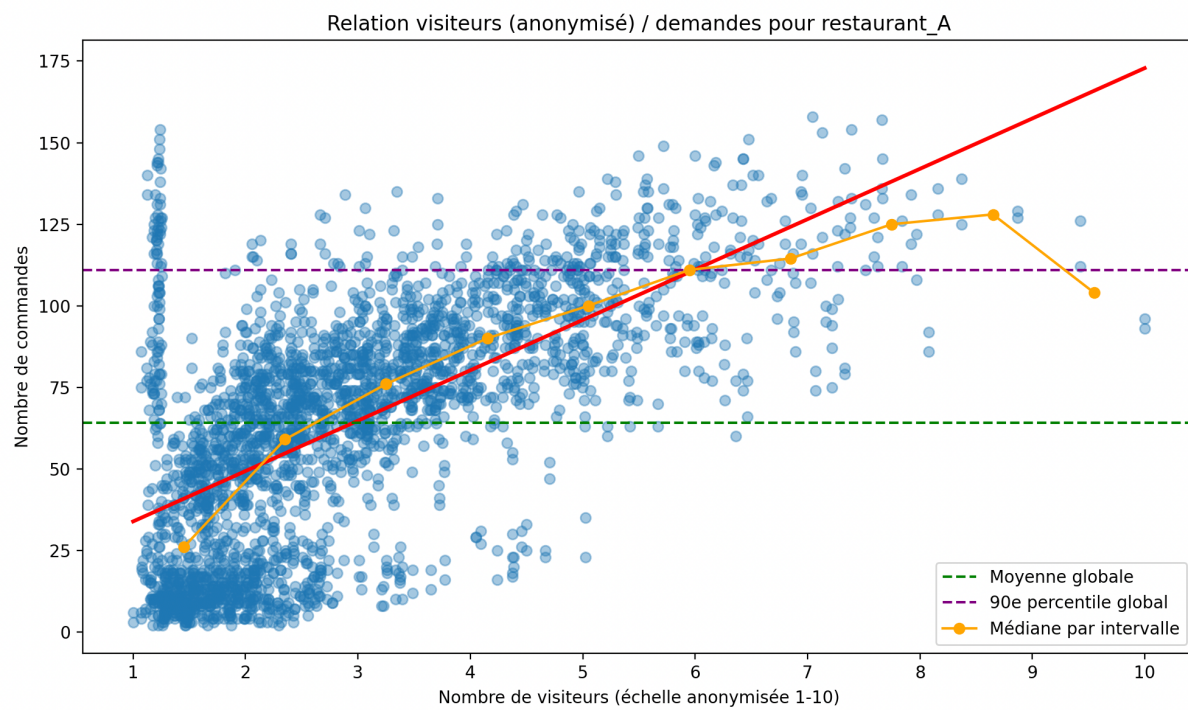


Figure 34 : Comportement spécifique du restaurant A

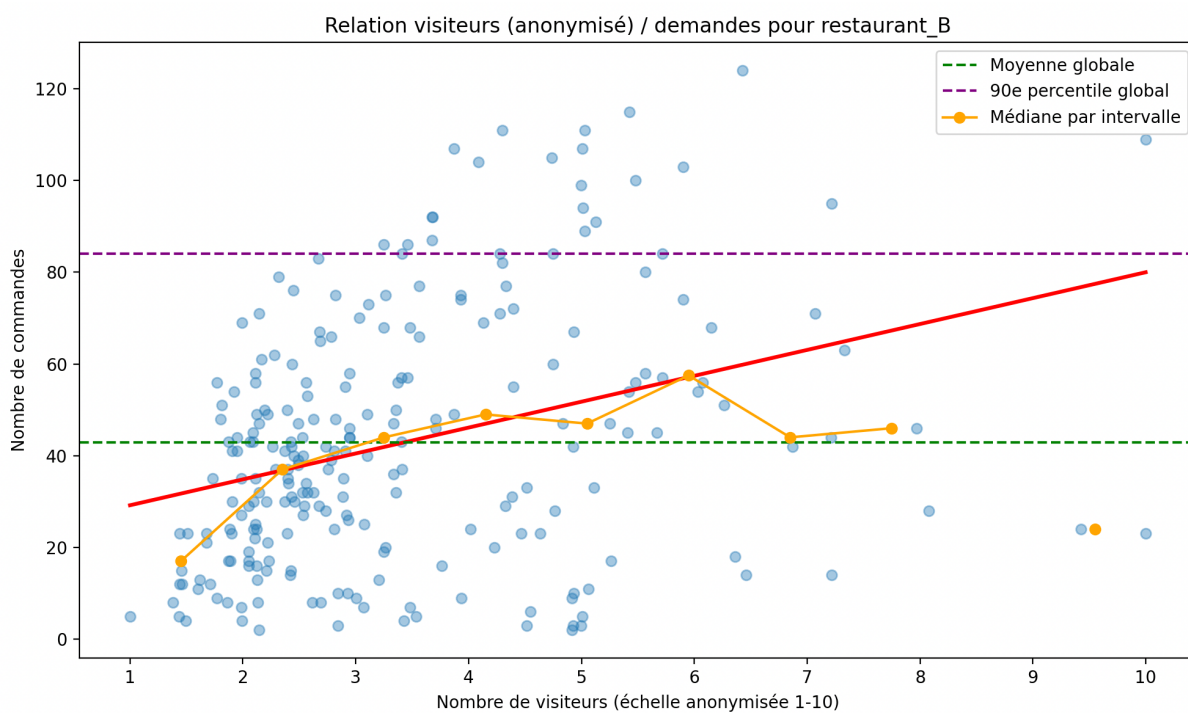


Figure 35 : Comportement spécifique du restaurant B

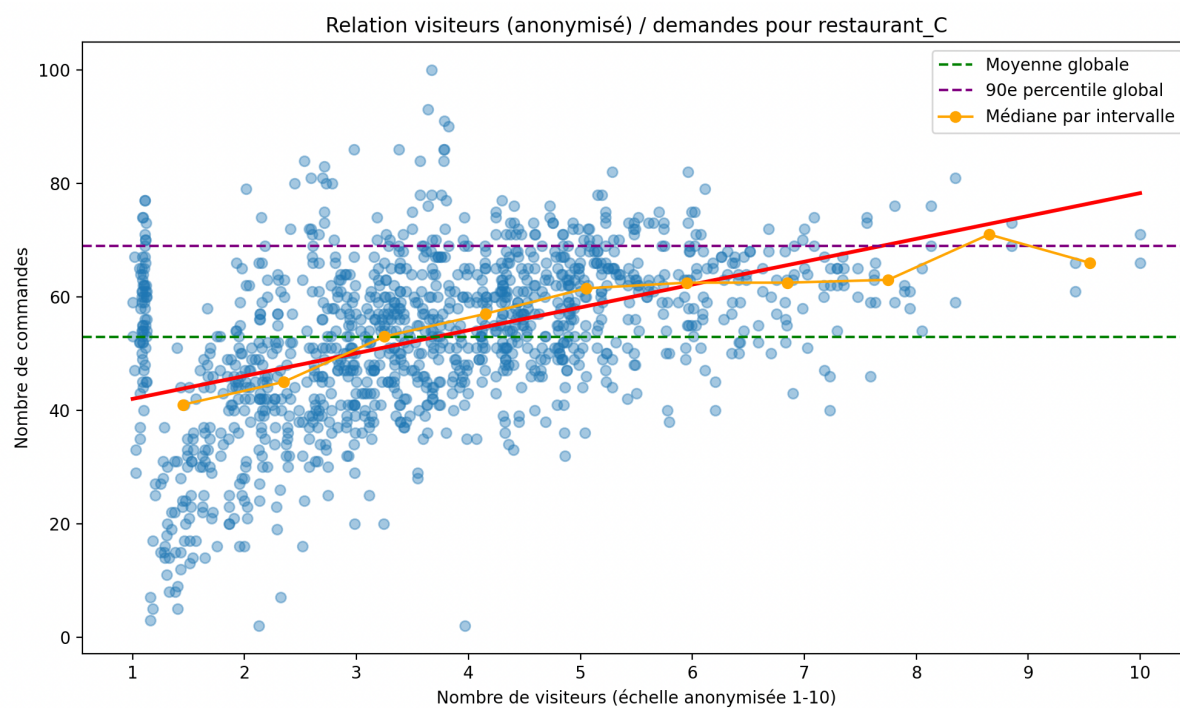


Figure 36 : Comportement spécifique du restaurant C

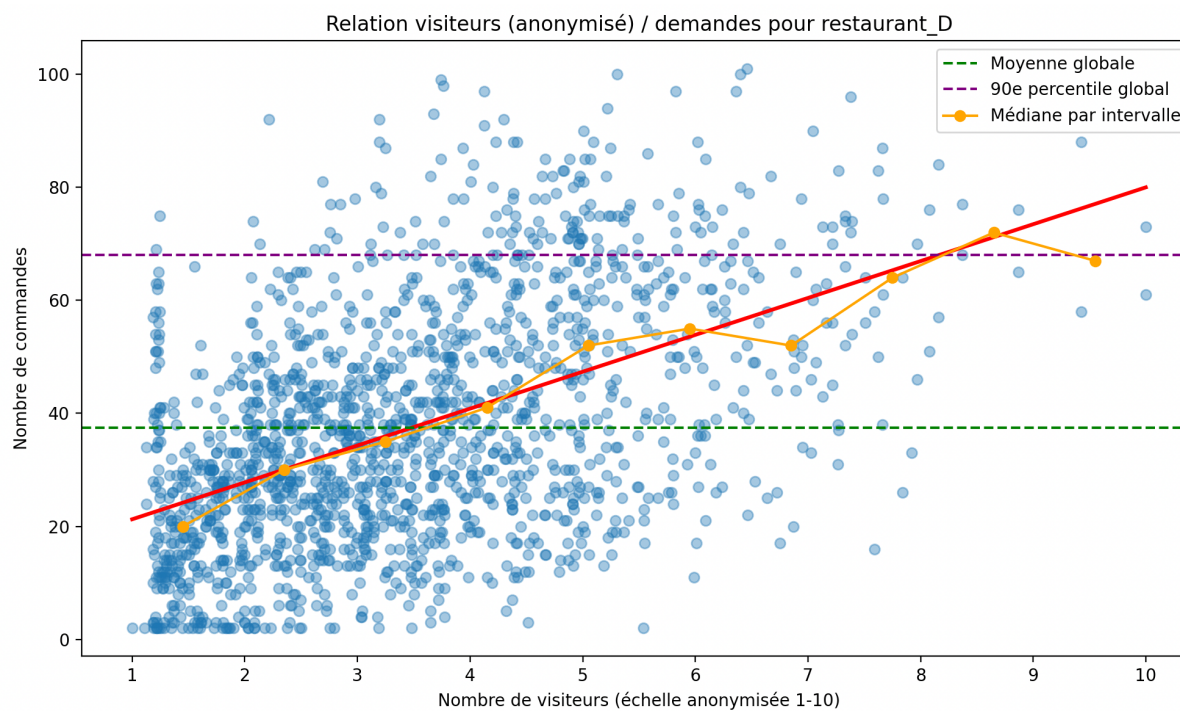


Figure 37 : Comportement spécifique du restaurant D

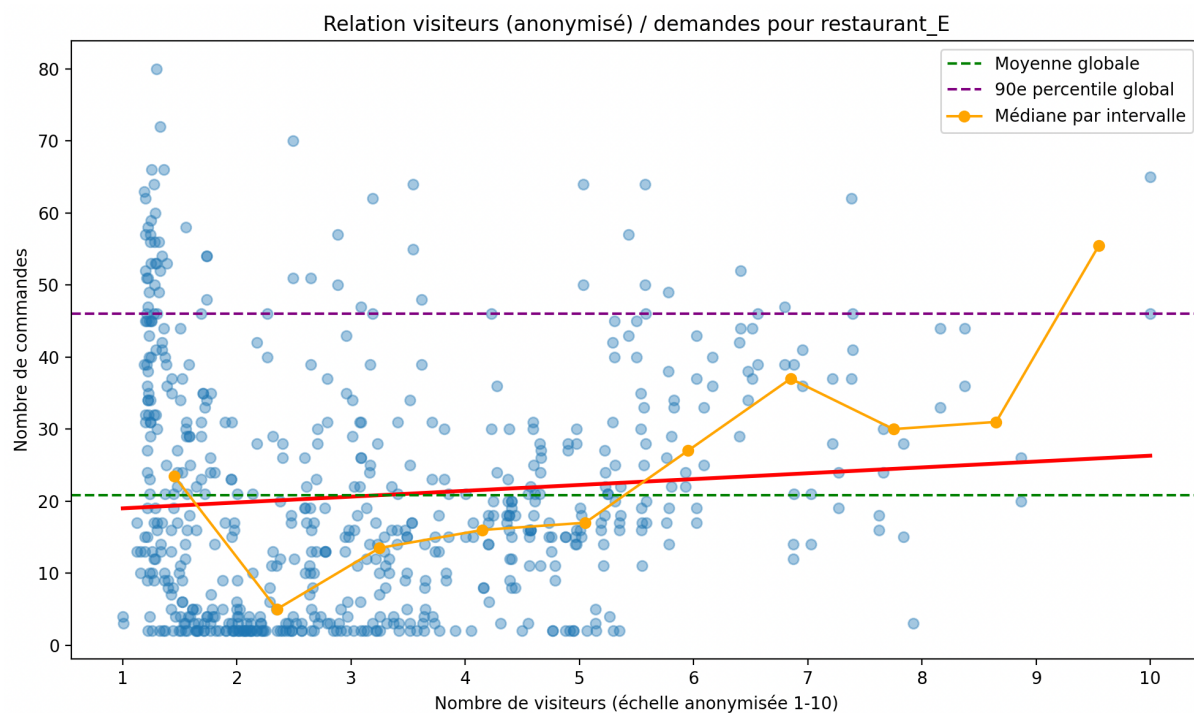


Figure 38 : Comportement spécifique du restaurant E

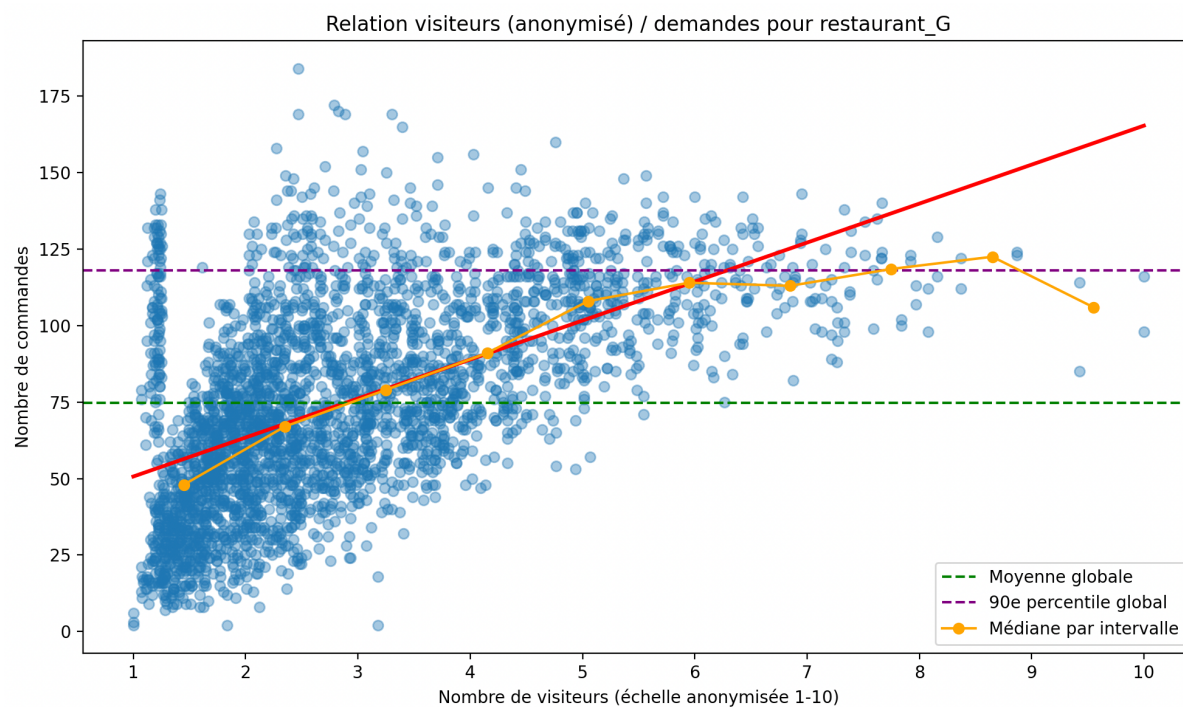


Figure 39 : Comportement spécifique du restaurant G

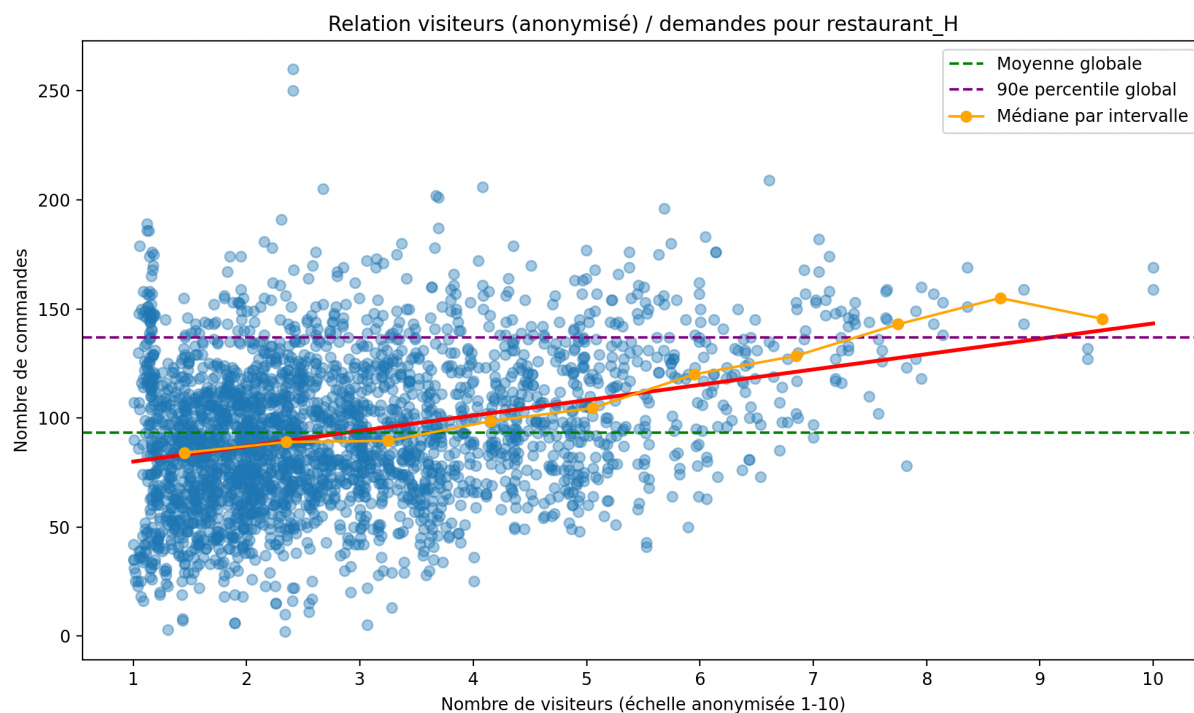


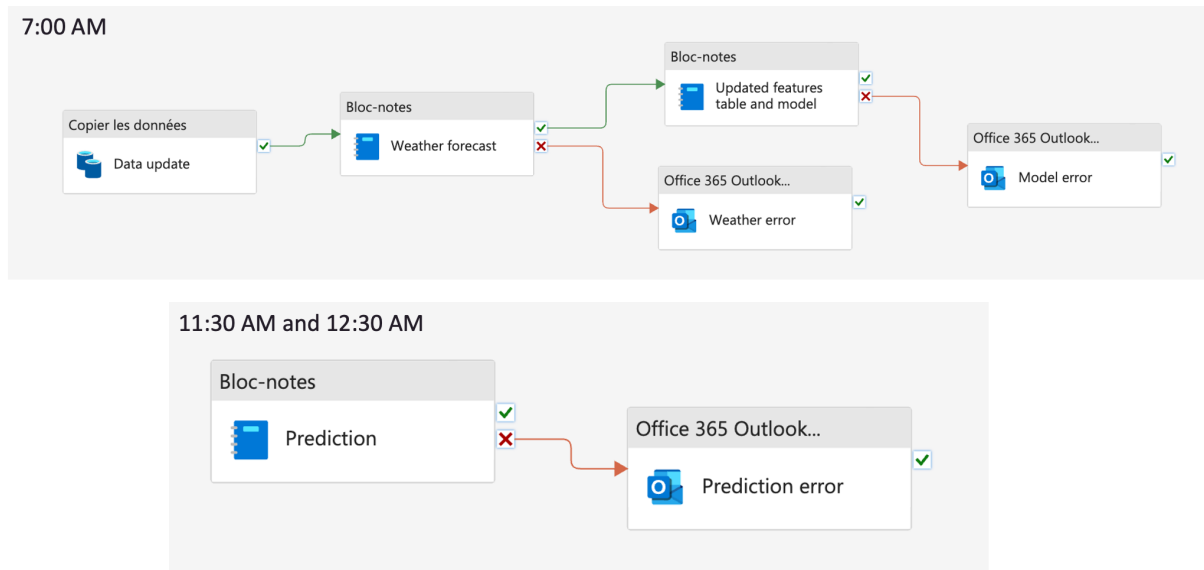
Figure 40 : Comportement spécifique du restaurant H

Annexe 3 : Tableau de description des variables pour le modèle prédictif

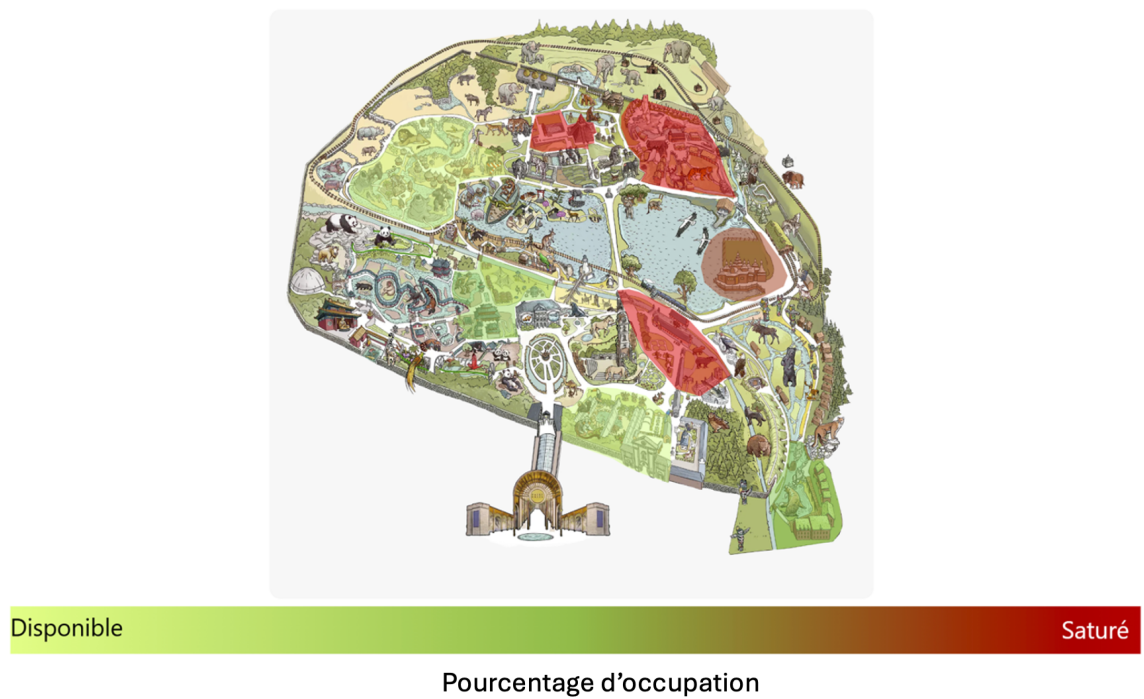
Nom de la donnée	Format de donnée	Type de variable	Explication
Année	Int	Features	Année
Mois	Int	Features	Le numéro mois (1 pour janvier, 12 pour décembre)
Jour_semaine	Int	Features	Le jour de la semaine (0 = lundi, 6 = dimanche)
Tranche_heure	Int	Features	Tranche horaire de la prochaine heure (1 pour 11h-12h, 2 pour 12h-13h et 3 pour 13h-14h).
Visiteurs	Int	Features	Nombre de visiteurs prévus pour le jour actuel.

Nom de la donnée	Format de donnée	Type de variable	Explication
Wallonia	Bool	Features	Indicateur binaire (0 ou 1) indiquant si le jour est un jour férié en Wallonie.
Pluie	Bool	Features	1 si les précipitations du jour sont supérieures à 0mm, 1 sinon
Temp_categorie	Int	Features	Température moyenne du jour arrondi à l'unité à la localisation du parc
VenteTranche1	Int	Features	Nombre de tickets vendus pour le restaurant et la tranche horaire précédente (une heure avant l'heure actuelle).
VenteTranche2	Int	Features	Nombre de tickets vendus pour le restaurant et la tranche horaire d'il y a deux heures.
Colonnes restaurant_*	Bool (One-Hot)	Features	Colonnes qui représentent chaque restaurant sous forme de variables "one-hot". Colonnes créées pour chaque restaurant, avec des valeurs binaires (0 ou 1).
Commandes	Int	Target	Nombre d'achat dans ce restaurant pour la période suivante

Annexe 4 : Flux d'optimisation Microsoft Fabric



Annexe 5 : Carte Power BI du niveau de saturation des zones



Annexe 6 : Tableau des notations

Notation	Description
i, j	Indices désignant les restaurants
V	Ensemble de tous les restaurants
V_i	Ensemble des restaurants voisins du restaurant i
C_i	Capacité maximale du restaurant i
D_i	Demande prédite du restaurant i
A_{ij}	Possibilité de redirection du restaurant i vers le restaurant j
K_{max}	Nombre maximal de redirection activées
f_{ij}	Variable indiquant le flux de visiteurs du restaurant i vers le restaurant j
y_{ij}	Variable binaire indiquant si la redirection de i vers j est activée
o_i	Variable indiquant le dépassement positif de capacité du restaurant i

