CIRRELT

CIRRELT-2024-23

Estimation des émissions de CO_2 des véhicules commerciaux du Québec à partir de données de vérification mécanique

Arnaud Teinturier
Amaury Philippe
Martin Trépanier

Juillet 2024

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

Estimation des émissions de CO_2 des véhicules commerciaux du Québec à partir de données de vérification mécanique

Arnaud Teinturier[1,*], Amaury Philippe[2], Martin Trépanier[2]

1. École Polytechnique, Paris
2. Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport (CIRRELT), Polytechnique Montréal et Chaire en innovation du transport

Résumé. L'objectif de cette étude est de déterminer, à partir de données kilométriques relevées lors de vérifications mécaniques les kilométrages annuels de chaque camion circulant au Québec. Cependant, la présence d'anomalies dans les données collectées nécessite un prétraitement pour garantir la fiabilité des résultats. De plus, environ 25% des camions ne disposent pas de données kilométriques, ce qui rend l'estimation de leur kilométrage annuel plus complexe. Pour pallier ce problème, nous déterminons une estimation des kilométrages annuels par interpolation par spline monotone pour les camions ayant des données exploitables. Ensuite, nous utilisons des méthodes de Machine Learning pour estimer les distances parcourues annuellement par les camions sans données kilométriques, en nous basant sur leurs caractéristiques techniques, géographiques et personnelles. Enfin, en agrégeant l'ensemble de ces informations, nous pouvons estimer les émissions de CO_2 par différents modèles des camions du Québec et analyser les résultats obtenus.

Mots-clés : Transport, estimation, kilométrage, émission CO_2

Remerciements : Les auteurs désirent remercier la Société de l'assurance-automobile du Québec pour la fourniture des données utilisées dans cette étude. Les auteurs remercient également le Ministère de l'Économie, de l'Innovation et de l'Énergie du Québec, qui finance la Chaire en transformation du transport.

---

*Auteur correspondant : arnaud.teinturier@polytechnique.edu

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2024

Table des matières

Glossaire

Acronymes

NIV Numéro d'identification du véhicule

CO_2 Dioxyde de Carbone

MSE Mean Square Error

R^2 Coefficient de régression

PTAC Poids Total Autorisé en Charge

CP3 Code Postal à 3 caractères

MRC Municipalité Régionale de Comté

GES Gaz à Effet de Serre

SAAQ Société de l'assurance automobile du Québec

CA Camion

Type de carburants

Diesel D

Essence E

GPL P

GNL N

Éthanol T

Électrique L

Hybride H

Hybride rechargeable W

Hydrogène C

Non-propulsé S

Autre A

Classes de véhicules

BCA Camion ou tracteur routier de plus de 3 000 kg pour le transport de biens.

COT Dépanneuse, véhicule de service, corbillard, ambulance, véhicule de conduite, ou véhicule avec plaque amovible.

HCA Camion ou tracteur routier de plus de 3 000 kg pour le transport de biens (circulation restreinte).

RCA Camion ou tracteur routier de plus de 3 000 kg pour le transport de biens (hors réseau).

1 Introduction

Dans le contexte des accords de Paris et de l'engagement du Canada à réduire ses émissions de gaz à effet de serre, le gouvernement québecois dispose de nombreuses données liées au transport et souhaite pouvoir quantifier davantage les émissions spécifiques par catégorie de véhicules. En effet, en 2021, le secteur du transport représente 42,6% des émissions de CO_2 au Québec [1] et correspond au premier poste d'émissions dans la région. Une analyse plus fine de la répartition de ces émissions au sein des transports s'avère intéressante.

Cette étude porte sur la flotte de véhicules commerciaux et se concentre en particulier sur les camions circulant au Québec. En effet, les données des entreprises ne sont pas publiques et coûtent très chères, ce qui les rend très peu accessible. Développer une estimation des kilométrages des véhicules par une approche statistique et d'apprentissage s'avère ainsi intéressant. Cette approche vient suivre une première étude menée par la Chaire en tranformation du transport sur les données du gouvernement concernant les voitures des particuliers. Les données équivalentes liées aux véhicules commerciaux (essentiellement des camions) n'ont jamais été étudiées et cette étude permettera ainsi de proposer une nouvelle approche pour estimer les émissions de CO_2 du transport routier, que l'on pourra comparer aux estimations gouvernementales.

En somme, le travail se découpe en plusieurs axes. Après avoir exploré, analysé et prétraité les données disponibles, nous développons un modèle permettant d'estimer les distances parcourues par les camions. Nous pouvons y comparer différentes approches. Pour les véhicules pour lesquels nous n'avons pas pu estimer les kilométrages, nous utilisons des méthodes d'apprentissage. Une fois les kilométrages obtenus, nous calculons les émissions de CO_2 associées pour chaque camion et nous discutons des résultats.

2 Analyse des données

2.1 Sources et type de données

La source principale de données est la Société de l'assurance automobile du Québec (SAAQ), qui nous a fourni un ensemble d'informations sur la flotte de véhicules circulant au Québec entre 2011 et 2022.

Nous disposons de trois types de données différentes :

- Base des véhicules autorisés à circuler en date du 31 décembre de chaque année
- Base des vérifications mécaniques concernant les véhicules en circulation à l'année courante depuis leur entrée dans le parc automobile québécois
- Base des transactions concernant les véhicules en circulation à l'année courante depuis leur entrée dans le parc automobile québécois.

Ces trois tables sont représentées sur la figure 17 en annexe. Le numéro d'identification du véhicule (NIV) correspond à l'index unique d'un véhicule et relie ainsi nos tables entre elles. Nous disposons d'une version de chaque table pour chaque année entre 2011 et 2022.

Base de données	Nombre d'attributs	Taille totale
Véhicules en circulation	19	8,8 Go
Vérifications mécaniques	4	1 Go
Transactions	8	12,2 Go

TABLE 1 – Informations sur les bases de données

Les tables de circulation contiennent des caractéristiques techniques, géographiques et personnelles concernant les véhicules. Dans les tables de transactions et de vérifications mécaniques, on retrouve des relevés kilométriques du compteur des véhicules à différents instants : lors d'un achat/vente du véhicule ou lors d'un contrôle technique. Ces informations sont plus ou moins fiables et les valeurs sont souvent mal renseignées. Ceci nécessite donc un prétraitement des données en amont du calcul estimant les distances parcourues chaque année par les véhicules.

En résumé, pour chaque véhicule, nous disposons donc dans le meilleur des cas des informations suivantes :

- Informations techniques, géographiques et personnelles
- Données kilométriques issues des vérifications mécaniques et des transactions (figure 1 où l'on peut déjà remarqué la présence d'anomalies)

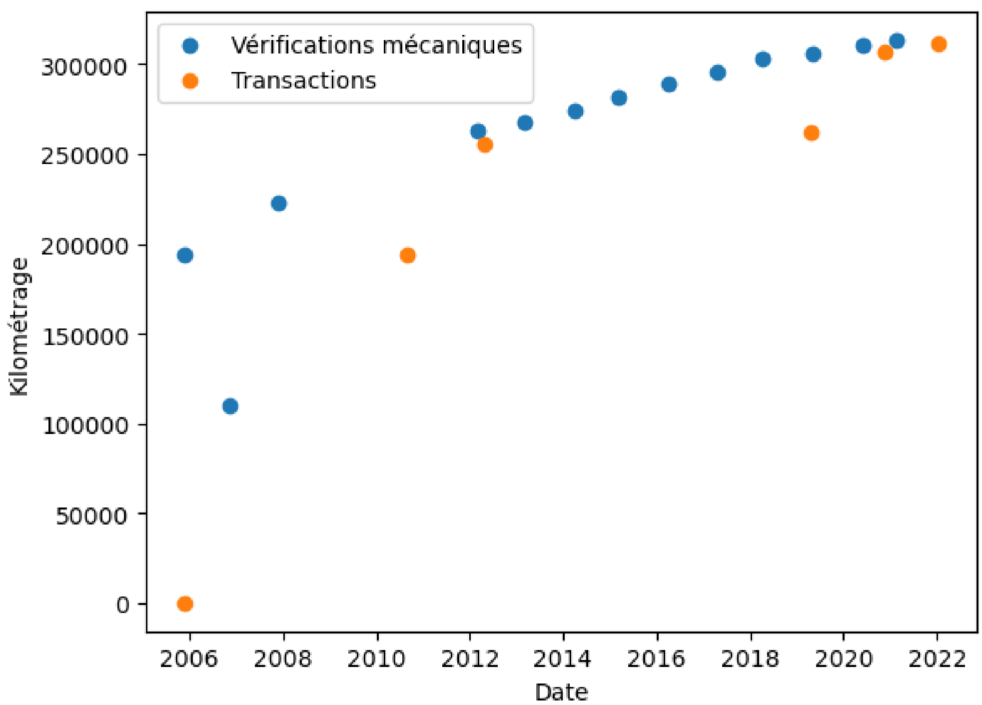

FIGURE 1 – Distribution des données issues de transactions et de vérifications mécaniques pour un véhicule

Au total, les données représentent 22Go de mémoire, ce qui est massif. Cela nécessite de réfléchir à un processus visant à réduire la quantité de données sans perdre d'information.

2.2 Exploration et prétraitement

Tout d'abord, l'importation des données a été effectué via le module *Pandas* de Python. En effet, les dataframes *Pandas* se prêtent bien à l'analyse de données, notamment lorsque celles-ci sont sous format csv.

Lorsqu'on observe le format des données, on se rend compte qu'elles comportent beaucoup de redondances. En effet, en ce qui concerne les transactions et les vérifications mécaniques, si un véhicule est présent dans les fichiers de circulation à plusieurs années distinctes, ses informations apparaîtront dans chacune des tables de transactions et de vérifications mécaniques des différentes années. De même, au niveau des données de circulation, beaucoup de véhicules ne changent pas d'informations d'une année à l'autre et toutes les données (hormis l'année de circulation) sont redondantes.

Ainsi, afin de réduire la taille des données à analyser, nous avons effectué différents prétraitements des données :

- Pour les données de transactions et de vérifications mécaniques :
  1. Fusion des données entre 2011 et 2022
  2. Suppression des duplicatas
- Pour les données gouvernementales :
  1. Fusion des données entre 2011 et 2022
  2. Récupération de la présence des véhicules et création d'une table de présence encodée en one-hot
  3. Récupération des données des véhicules pour lesquels il y a des changements d'une année à l'autre et création d'une table avec ces données variables
  4. Création d'une table avec les données les plus récentes pour chaque véhicule et suppression des duplicatas

Ce prétraitement des données a permis de réduire la taille des données par 2, passant de 22Go à 10Go. Par la suite, on transforme l'ensemble des dates en format *np.datetime64*, ce qui facilitera l'analyse à l'issue. On résume ce prétraitement dans la figure 16 en annexe.

Comme on s'intéresse au camion et pour diminuer encore la taille des fichiers avec lesquels on va travailler, on réalise le même processus en filtrant au départ les véhicules de type 'CA' (camions). Cela permettra d'améliorer grandement la vitesse d'exécution des différents algorithmes.

2.3 Statistiques descriptives pour les camions

L'effectif total étudié est de **328 777 camions**. A l'aide des différentes tables produites, nous pouvons analyser différentes caractéristiques des camions et leur distribution.

A partir des figures 2,3,4,5, on peut noter que chaque année, le nombre de nouveaux camions en circulation s'élève à 4500 et l'effectif totale de la flotte suit donc une loi quasi-linéaire. On observe également que la majorité des camions sont de types 'BCA', c'est-à-dire des véhicules lourds de plus de 3 tonnes conçus pour le transport de marchandise. Les véhicules roulent pour la plupart au diesiel, puis à l'essence. La majorité des camions sont récents et datent d'après l'an 2000 (85%).

Sur la figure 6, on a représenté la distribution des masses nettes des camions. La masse moyenne est de 6895 kg et on remarque qu'une grande partie des camions ont des masses aux alentours de 3500 kg ou 8000 kg. La masse est une variable importante car elle joue une rôle décisif dans les émissions de gaz à effet de serre. D'un point de vue géographique, la plupart des véhicules sont immatriculés dans la région de Montréal et de Montérégie comme le montre la figure 18 en annexe.

Enfin, la répartition des camions entre personnes morales (entreprises) et personnes physiques (particuliers) montre une nette prédominance des entreprises. En effet, 93,1 % des camions sont enregistrés au nom de personnes morales contre seulement 6,9 % par des particuliers. Ces chiffres montrent que la majorité des camions sont utilisés à des fins commerciales.

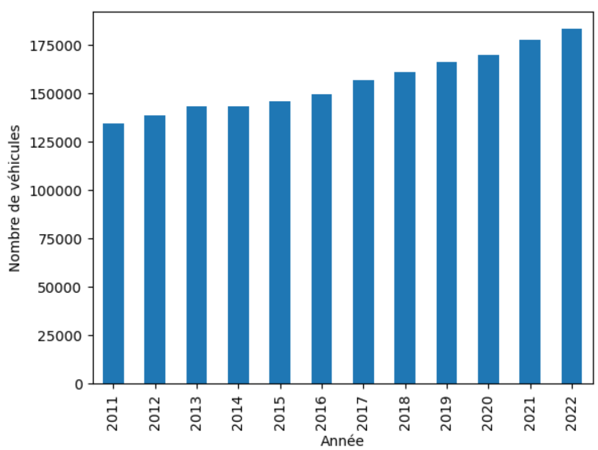

FIGURE 2 – Evolution de la flotte de camions

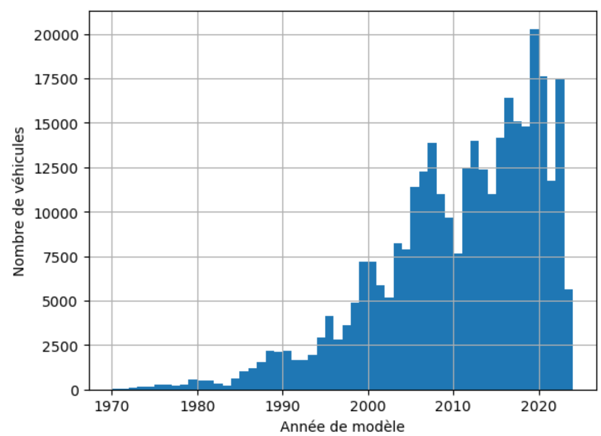

FIGURE 3 – Répartition par année de modèle

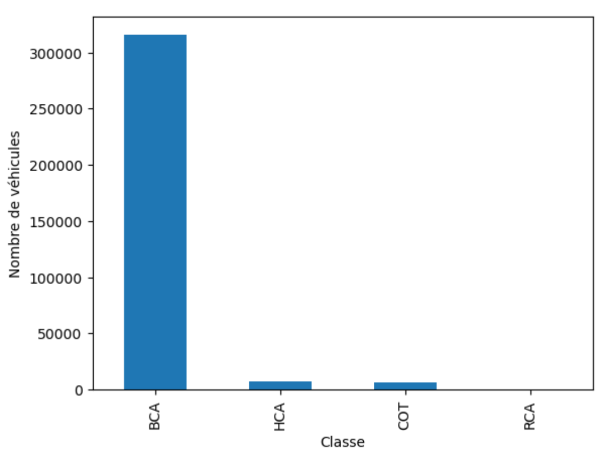

FIGURE 4 – Répartition par classe

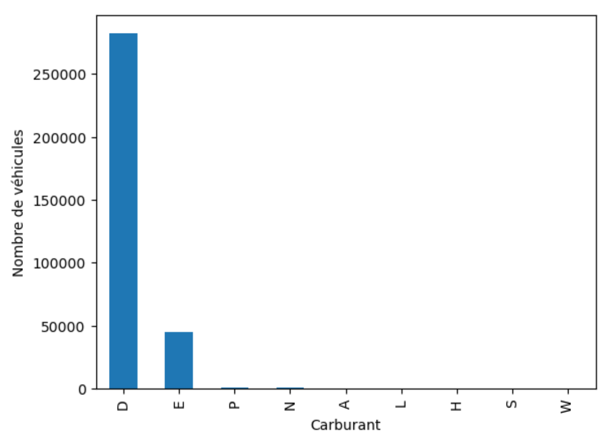

FIGURE 5 – Répartition par type de carburant

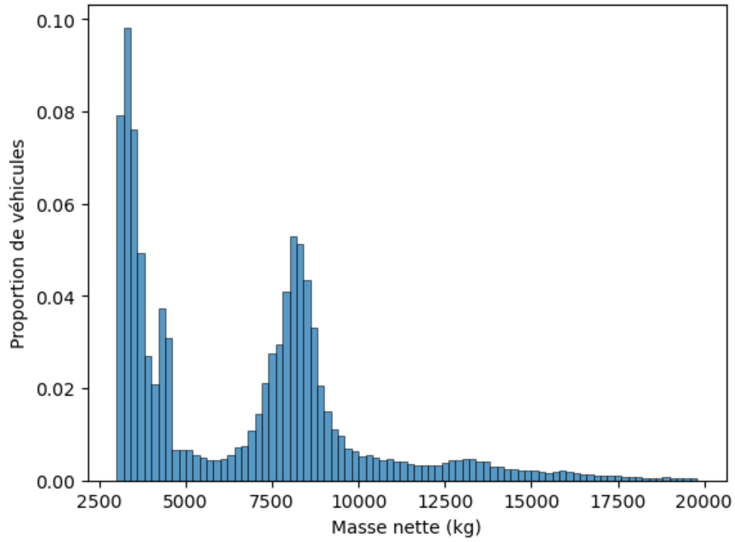

FIGURE 6 – Distribution des masses nettes des véhicules

2.4 Comparaison entre données transactionnelles et vérifications mécaniques

L'objectif étant d'estimer les distances annuelles à partir des données kilométriques, nous allons comparer les données transactionnelles avec celles issues de vérifications mécaniques. En particulier, nous allons étudié la quantité disponible ainsi que la qualité des données. A priori, les valeurs renseignées lors d'un contrôle technique sont censées être plus fiables que celles non nécessairement renseignées lors d'une transaction, par exemple entre particuliers.

Sur l'ensemble des 328 777 camions, **74,6%** possèdent au moins une vérification mécanique contre seulement **59,7%** possédant au moins une transaction. D'après la figure 7, parmi les relevés kilométriques disponibles, les vérifications mécaniques ne contiennent aucune valeur nulle ou non renseignée, tandis qu'au niveau des transactions, près de la moitié des données ne sont pas renseignées (valeur 0 par défaut). En moyenne, en se référant aux figures 19,20, un véhicule dispose de **8,8** données de vérifications et **2,7** données de transactions.

En termes de qualité, nous remarquons la présence d'anomalies dans les données. Ces anomalies sont de plusieurs types :

- Décroissance au cours du temps des kilométrages
- Ecart aberrant entre deux données
- Kilométrage nul (non renseigné)

La première étape de filtrage correspond à la suppression des points nuls pour lesquels l'information kilométrique aurait dû être renseignée mais ne l'a pas été.

On note n le nombre de véhicules distincts. Pour un véhicule i dans $\{1, ..., n\}$, on peut modéliser les données disponibles pour un type d'événements (transactions ou vérifications mécaniques) sous la forme d'un ensemble de points $(y_{i,j})$, où j varie entre 1 et T_i (nombre de

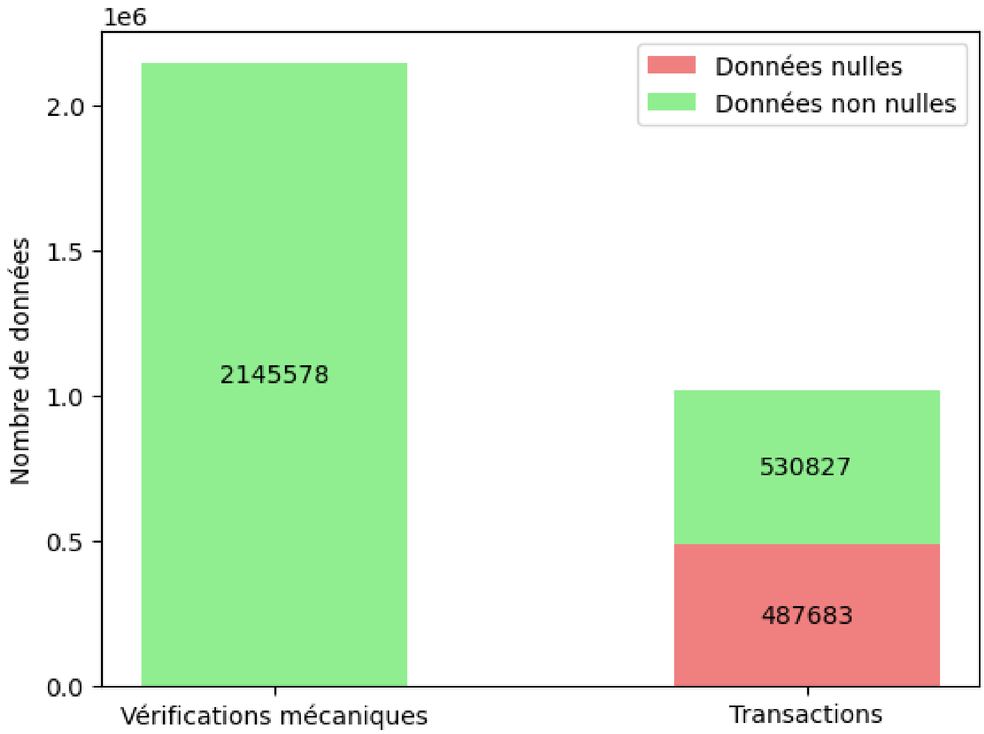

FIGURE 7 – Répartition des données kilométriques par type de données

transactions pour le véhicule i) ou V_i (nombre de vérifications mécaniques pour le véhicule i). Chaque point $y_{i,j}$ est caractérisé par une date, un kilométrage et un événement.

Pour étudier les points de chute, on suppose les $y_{i,j}$ triés par ordre chronologique et on distingue les données transactionnelles des données de vérifications mécaniques. On effectue ensuite pour chaque ensemble $y_{i,j}$ un décompte du nombre de points de chute, c'est-à-dire :

$$\begin{cases} \mathrm{N_{tr}} = \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{T_i} \mathbf{1}_{\{y_{i,j+1} < y_{i,j}\}}, \\ \mathrm{N_{ver}} = \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{V_i} \mathbf{1}_{\{y_{i,j+1} < y_{i,j}\}}, \end{cases}$$

Dans le tableau 2, nous avons répertorié les résultats du comptage par type d'évènements. A nouveau, nous constatons que beaucoup de véhicules possèdent une seule voire aucune donnée de transactions (60,9%), ce qui restreint l'échantillon pour lequel il serait possible d'estimer les kilométrages annuels. De plus, en se basant sur les vérifications mécaniques, nous atteignons 45,3% de véhicules qui ont au moins deux données et des points strictement croissants dans le temps, tandis que cela ne concerne que 33,4% de véhicules via les transactions.

Utiliser les vérifications mécaniques pour estimer les kilométrages annuels semble donc être plus fiable pour les estimations. Pour le calcul, il sera intéressant de se baser sur les données sans point de chute comportant au moins deux points, nous qualifierons ces données

Nombre de points de chute	Vérifications mécaniques		Transactions	
	Pourcentage	Effectif	Pourcentage	Effectif
0	45.3	149044	33.4	109838
1	12.6	41272	5.1	16732
2	4.5	14700	0.5	1724
3+	1.9	6199	0.1	175
0 ou 1 donnée	35.7	117562	60.9	200308

TABLE 2 – Répartition de la présence de points de chute dans les données par véhicule

de **cohérentes**. Ces données cohérentes représentent 45,3% de véhicules et constituent un bon set d'entrainement pour déterminer les estimations manquantes à l'aide d'algorithmes d'apprentissage.

Pour la suite, nous considérons uniquement les données kilométriques cohérentes issues des vérifications mécaniques. Nous désignerons un véhicule par l'indice i, les relevés kilométriques par $(y_{i,j})_{i=1..n,j=1..V_i}$ et les dates correspondantes par $(t_{i,j})_{i=1..n,j=1..V_i}$. En réalisant des premières estimations par régression linéaire basique, on se rend rapidement compte que certains points $(y_{i,j})$ créent des aberrations dans le résultat, puisque les écarts relatifs entre deux points successifs sont parfois impossibles et sont dûs à des erreurs humaines lors de la saisie.

Pour filtrer les points aberrants, il faut donc calculer l'équivalent d'une vitesse (km/jour) entre deux points succesifs. On peut ensuite utiliser une méthode de détection des valeurs extrêmes, basée sur le 99e centile des rapports kilomètres par jour entre deux points successifs. Pour calculer ce rapport, on peut procéder comme suit :

1. Trier les données $(y_{i,j})$ par ordre croissant de la date de vérification mécanique, pour chaque véhicule i.
2. Calculer le rapport en Km/jour entre deux points : $\delta_{i,j} = \frac{y_{i,j+1} - y_{i,j}}{t_{i,j+1} - i_{,j}}$ pour $j \in [1, V_i - 1]$

Une fois ceci réalisé, nous pouvons déterminer le 99e centile de l'ensemble des rapports km/jour, qui vaut **650 km/jour**. Ensuite, nous procédons de manière récursive pour éliminer les points aberrants :

1. Supprimer tout point $(y_{i,j})$ vérifiant $\delta_{i,j} > q_{99}$
2. Calculer les nouveaux rapports km/jour entre les points
3. Réitérer jusqu'à ce qu'aucun point ne soit supprimé

La méthode itérative est utile pour s'assurer que tous les points aberrants sont traités, et pas seulement le premier. En effet, si plusieurs points sont anormalement situés par rapport à d'autres, la suppression du premier point peut laisser un nouvel écart potentiellement aberrant avec le deuxième point. En répétant le processus de détection et de suppression des points aberrants jusqu'à ce qu'aucun point ne soit supprimé, on peut s'assurer que tous les écarts aberrants ont été traités.

Après filtrage, on obtient des données nettoyées pour l'estimation kilométrique pour 146 853 véhicules, ce qui signifie que le traitement récursif n'élimine que 2 191 véhicules pour lesquels il reste donc moins de deux points de données (environ 1,4% des véhicules). Un filtrage équivalent sur les données de transactions fait chuter le nombre de véhicules à 101 903, soit environ 7,2% camions en moins. Ceci souligne encore le manque de qualité des données transactionnelles.

3 Estimation kilométrique

Dans cette section, nous allons expliquer les méthodes de calcul permettant d'estimer les kilométrages annuels parcourus par chaque véhicule entre 2011 et 2022. Pour cela, nous nous appuierons sur l'ensemble des données kilométriques cohérentes. En effet, sans appliquer de prétraitement et de filtre sur les points aberrants, on obtient des valeurs kilométriques aberrantes pouvant estimer plusieurs millions de kilomètres parcourus en une année par exemple.

L'estimation sera réalisée en deux temps : estimer le kilométrage à l'odomètre de chaque véhicule au 31 décembre de l'année courante, puis calculer la distance annuelle parcourue.

3.1 Différents modèles

Le comportement d'un conducteur étant plus ou moins similaire d'une année à l'autre, une méthode de régression linéaire semble a priori cohérente. Quatre méthodes différentes sont proposées pour estimer les kilométrages.

Pour un véhicule i, on note m_i son nombre de données et elles seront représentées par :

$$y_i = \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,m_i} \end{pmatrix} \quad \text{et} \quad t_i = \begin{pmatrix} t_{i,1} \\ \vdots \\ t_{i,m_i} \end{pmatrix}$$

3.1.1 Modèle 1 : Régression linéaire classique

Dans ce modèle, nous réalisons une régression linéaire pour chaque ensemble de points kilométriques pour chacun des véhicules.

En utilisant l'erreur quadratique moyenne, le risque à minimiser s'écrit sous la forme : $\hat{R}_i(a,b) = \frac{1}{m_i}\sum_{j=1}^{m_i}(y_{i,j} - a \cdot t_{i,j} - b)^2$.

On calcule le gradient de $\hat{R}_i$:

$$\nabla \hat{R}_i(a,b) = \begin{pmatrix} \frac{\partial \hat{R}_i}{\partial a} \\ \frac{\partial \hat{R}_i}{\partial b} \end{pmatrix} = \begin{pmatrix} -\frac{2}{m_i}\sum_{j=1}^{m_i} t_{i,j}(y_{i,j} - at_{i,j} - b) \\ -\frac{2}{m_i}\sum_{j=1}^{m_i}(y_{i,j} - at_{i,j} - b) \end{pmatrix}$$

On résoud le système suivant :

$$
\begin{aligned}
\nabla \hat{R}_i(a,b) = 0 &\Leftrightarrow \begin{cases} \sum_{j=1}^{m_i} t_{i,j}(y_{i,j} - at_{i,j} - b) = 0 \\ \sum_{j=1}^{m_i} (y_{i,j} - at_{i,j} - b) = 0 \end{cases} \\
&\Leftrightarrow \begin{cases} \sum_{j=1}^{m_i} t_{i,j}y_{i,j} - a\sum_{j=1}^{m_i} t_{i,j}^2 - b\sum_{j=1}^{m_i} t_{i,j} = 0 \\ \sum_{j=1}^{m_i} y_{i,j} - a\sum_{j=1}^{m_i} t_{i,j} - bm_i = 0 \end{cases} \\
&\Leftrightarrow \begin{cases} a\sum_{j=1}^{m_i} t_{i,j}^2 + b\sum_{j=1}^{m_i} t_{i,j} = \sum_{j=1}^{m_i} t_{i,j}y_{i,j} \\ a\sum_{j=1}^{m_i} t_{i,j} + bm_i = \sum_{j=1}^{m_i} y_{i,j} \end{cases} \\
&\Leftrightarrow \begin{cases} a = \frac{m_i\sum_{j=1}^{m_i} t_{i,j}y_{i,j} - \sum_{j=1}^{m_i} t_{i,j}\sum_{j=1}^{m_i} y_{i,j}}{m_i\sum_{j=1}^{m_i} t_{i,j}^2 - \left(\sum_{j=1}^{m_i} t_{i,j}\right)^2} \\ b = \frac{\sum_{j=1}^{m_i} y_{i,j} - a\sum_{j=1}^{m_i} t_{i,j}}{m_i} \end{cases} \\
&\Leftrightarrow \begin{cases} a = \frac{\mathrm{Cov}(y_i,t_i)}{\mathbb{V}(t_i)} \\ b = \bar{y_i} - a\bar{t_i} \end{cases}
\end{aligned}
$$

où $\bar{y_i}$ et $\bar{t_i}$ sont les moyennes respectives des vecteurs y_i et t_i.

Ainsi, le kilométrage y à un instant t donné est calculé comme suit :

$$y = \frac{\mathrm{Cov}(y_i,t_i)}{\mathbb{V}(t_i)} \cdot t + \left(\bar{y_i} - \frac{\mathrm{Cov}(y_i,t_i)}{\mathbb{V}(t_i)} \cdot \bar{t_i}\right)$$

3.1.2 Modèle 2 : Interpolation linéaire avec extrapolation sur les extrémités

Dans ce modèle, nous réalisons une interpolation linéaire pour chaque ensemble de points kilométriques pour chacun des véhicules. Pour l'extrapolation, seules les données des deux points d'extrémité sont utilisées. En se basant uniquement sur ces points, le modèle peut prendre en compte une tendance spécifique, à savoir que plus un véhicule a de kilomètres, moins il en parcourt. Les segments sont définis par deux instants consécutifs $(t_{i,j}, t_{i,j+1})$ avec $j = 1, \ldots, m_i - 1$.

Le modèle fonctionne de la manière suivante pour déterminer le kilométrage y à un instant t donné :

1. Si $t \leq t_{i,1}$ (extrapolation à gauche) :

$$a_{i,1,2} = \frac{y_{i,2} - y_{i,1}}{t_{i,2} - t_{i,1}}, \quad b_{i,1,2} = y_{i,1} - a_{i,1,2} \cdot t_{i,1}$$

$$y = a_{i,1,2} \cdot t + b_{i,1,2}$$

2. Si $t \in [t_{i,j}, t_{i,j+1}]$ (interpolation par morceaux) :

$$a_{i,j,j+1} = \frac{y_{i,j+1} - y_{i,j}}{t_{i,j+1} - t_{i,j}}, \quad b_{i,j,j+1} = y_{i,j} - a_{i,j,j+1} \cdot t_{i,j}$$

$$y = a_{i,j,j+1} \cdot t + b_{i,j,j+1}$$

3. Si $t \geq t_{i,m_i}$ (extrapolation à droite) :

$$a_{i,m_i-1,m_i} = \frac{y_{i,m_i} - y_{i,m_i-1}}{t_{i,m_i} - t_{i,m_i-1}}, \quad b_{i,m_i-1,m_i} = y_{i,m_i-1} - a_{i,m_i-1,m_i} \cdot t_{i,m_i-1}$$

$$y = a_{i,m_i-1,m_i} \cdot t + b_{i,m_i-1,m_i}$$

Ainsi, le kilométrage y à un instant t donné est calculé en fonction de sa position par rapport aux instants t_i connus, soit par extrapolation à gauche, par interpolation entre deux points, soit par extrapolation à droite.

3.1.3 Modèle 3 : Interpolation linéaire avec extrapolation sur l'ensemble des points

Dans ce modèle, nous combinons la régression linéaire classique et l'interpolation linéaire. Nous réalisons une interpolation linéaire pour chaque ensemble de points kilométriques pour chacun des véhicules, tout en utilisant les formules de la régression linéaire classique pour l'extrapolation globale (c'est-à-dire pour les $t \notin [t_{i,1}, t_{i,m_i}]$).

Le modèle fonctionne de la manière suivante pour déterminer le kilométrage y à un instant t donné :

1. Si $t \notin [t_{i,1}, t_{i,m_i}]$ (extrapolation globale) :

$$a = \frac{\text{Cov}(y_i, t_i)}{\mathbb{V}(t_i)}, \quad b = \bar{y_i} - a \cdot \bar{t_i}$$

$$y = a \cdot t + b$$

2. Si $t \in [t_{i,j}, t_{i,j+1}]$ (interpolation par morceaux) :

$$a_{i,j,j+1} = \frac{y_{i,j+1} - y_{i,j}}{t_{i,j+1} - t_{i,j}}, \quad b_{i,j,j+1} = y_{i,j} - a_{i,j,j+1} \cdot t_{i,j}$$

$$y = a_{i,j,j+1} \cdot t + b_{i,j,j+1}$$

Ainsi, le kilométrage y à un instant t donné est calculé en fonction de sa position par rapport aux instants t_i connus, soit par extrapolation globale si $t \notin [t_{i,1}, t_{i,m_i}]$, soit par interpolation entre deux points si $t \in [t_{i,j}, t_{i,j+1}]$.

3.1.4 Modèle 4 : Interpolation par spline monotone avec extrapolation linéaire

Dans ce modèle, nous réalisons une interpolation par spline monotone [1] pour chaque ensemble de points kilométriques pour chacun des véhicules, tout en utilisant une extrpolation linéaire sur les extrémités. Cette méthode permet de garantir que le polynôme interpolé reste monotone car les kilométrages doivent être croissants dans le temps.

Le modèle fonctionne de la manière suivante pour déterminer le kilométrage y à un instant t donné :

1. Si $t \leq t_{i,1}$ (extrapolation à gauche) :

$$a_{i,1,2} = \frac{y_{i,2} - y_{i,1}}{t_{i,2} - t_{i,1}}, \quad b_{i,1,2} = y_{i,1} - a_{i,1,2} \cdot t_{i,1}$$

1. Utilisation de PCHIP : Piecewise Cubic Hermite Interpolating Polynomial sont des polynômes interpolateurs de degré 3

$$y = a_{i,1,2} \cdot t + b_{i,1,2}$$

2. Si $t \in [t_{i,1}, t_{i,m_i}]$ (interpolation par spline monotone) :

$$y = \text{PCHIP}(t_i, y_i)(t)$$

3. Si $t \geq t_{i,m_i}$ (extrapolation à droite) :

$$a_{i,m_i-1,m_i} = \frac{y_{i,m_i} - y_{i,m_i-1}}{t_{i,m_i} - t_{i,m_i-1}}, \quad b_{i,m_i-1,m_i} = y_{i,m_i-1} - a_{i,m_i-1,m_i} \cdot t_{i,m_i-1}$$

$$y = a_{i,m_i-1,m_i} \cdot t + b_{i,m_i-1,m_i}$$

Ainsi, le kilométrage y à un instant t donné est calculé en fonction de sa position par rapport aux instants t_i connus, soit par interpolation par spline monotone si $t \in [t_{i,1}, t_{i,m_i}]$, soit par extrapolation sur les deux points des extrémités.

3.2 Choix du modèle d'estimation

Sur la figure 8, on a représenté un exemple d'application des différents modèles d'estimation. Dans le tableau 13 en annexe, on a récapitulé les expressions de l'estimation kilométrique y à un instant t pour les différents modèles.

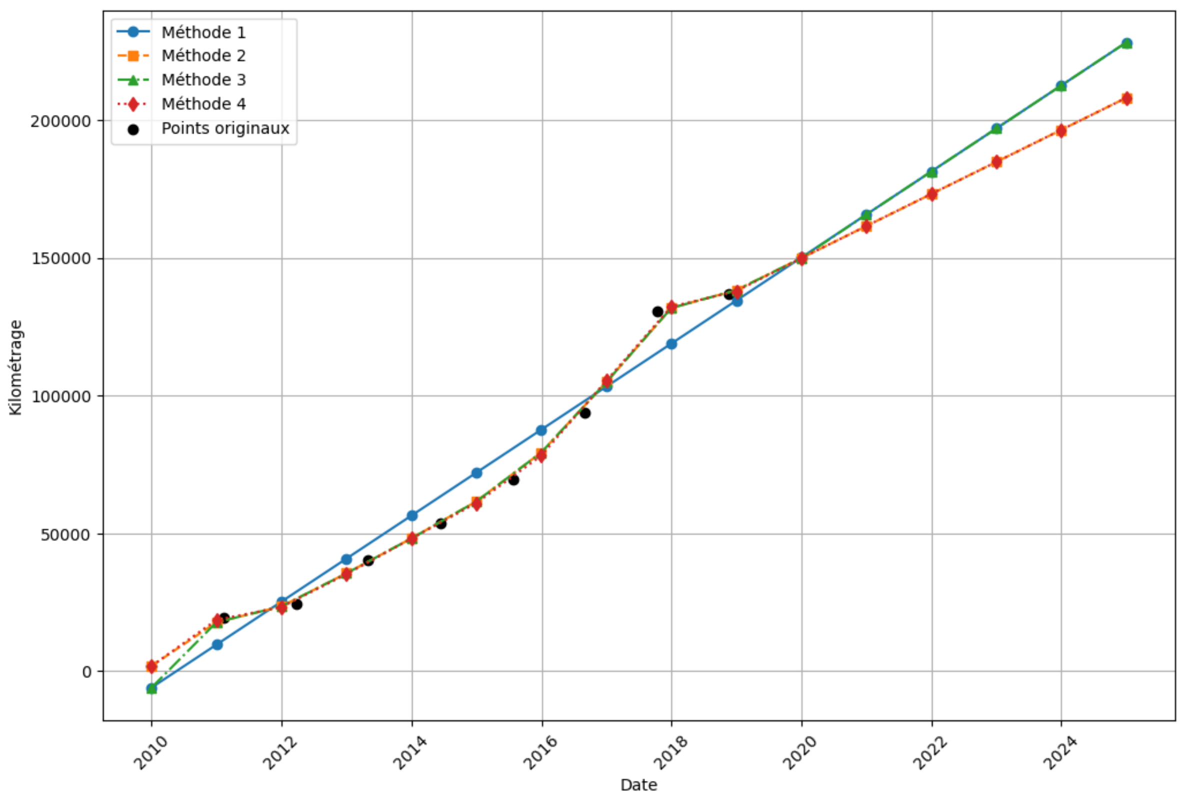

FIGURE 8 – Application des modèles d'estimation kilométrique pour une même série de points

Pour évaluer la précision des modèles, nous utilisons une procédure de validation croisée spécifique :

1. **Sélection aléatoire d'un point de données** : Pour chaque véhicule i, nous retirons un point de données $(t_{i,k}, y_{i,k})$ de manière aléatoire parmi les m_i points disponibles. Ce point est utilisé comme référence pour évaluer la précision des modèles.
2. **Estimation par les modèles** : Les modèles utilisent les $m_i - 1$ points de données restants pour chaque véhicule i pour estimer le point retiré.
3. **Prédiction et calcul de l'erreur** : Chaque modèle prédit le kilométrage $\hat{y}_{i,k}$ pour le point retiré $t_{i,k}$ et l'erreur est ensuite calculée à partir d'une fonction de perte donnée.

On réitère l'opération K fois et on moyenne l'erreur obtenue pour chaque véhicule. La moyenne empirique de l'erreur pour un modèle h donné est alors :

$$\hat{L}_i(h) = \frac{1}{K} \sum_{j=1}^{K} \ell(y_{i,k_j}, h(t_{i,k_j}))$$

où k_j est l'indice du point retiré à l'itération j.

Pour la fonction de perte, on prendra : $\ell(y, \hat{y}) = y - \hat{y}$. Pour comparer les performances des différents modèles, nous considérons les erreurs moyennes sur tous les véhicules. Si nous avons n véhicules, l'erreur moyenne pour un modèle donné est :

$$\begin{aligned} \hat{L}(h) &= \frac{1}{n} \sum_{i=1}^{n} \hat{L}_i(h) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{K} \sum_{j=1}^{K} \ell(y_{i,k_j}, h(t_{i,k_j})) \end{aligned}$$

Les erreurs $\hat{L}_i(h)$ sont indépendantes, d'espérance et de variance finies. En prenant l'hypothèse que ces erreurs sont identiquement distribuées, nous pouvons appliquer le théorème central limite (TCL) pour estimer un intervalle de confiance pour l'erreur moyenne $\hat{L}(h)$. Par le TCL, on a que :

$$\sqrt{n}\left(\hat{L}(h) - \mu\right) \xrightarrow{n} \mathcal{N}(0, \sigma^2)$$

où $\mu = \mathbb{E}[\hat{L}(h)]$ et $\sigma^2 = \mathbb{V}[\hat{L}(h)]$

Ainsi, $\hat{L}(h)$ suit approximativement une loi normale $\mathcal{N}(\mu, \sigma^2/n)$. Un intervalle de confiance à 95% pour l'erreur moyenne est alors donné par :

$$IC_{95\%}(\hat{L}(h)) = \left[\hat{L}(h) - 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \hat{L}(h) + 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

où $\hat{\sigma}$ est l'écart-type empirique des erreurs calculé comme suit :

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{K} \sum_{j=1}^{K} \ell(y_{i,k_j}, h(t_{i,k_j})) - \hat{L}(h) \right)^2}$$

Les résultats obtenus sont indiqués dans le tableau 3. Ces résultats montrent que les modèles 2 et 4 semblent fournir des estimations plus précises des points kilométriques, car leur erreur absolue moyenne est plus faible et leurs intervalles de confiance sont plus étroits, indiquant une plus grande précision.

	Modèle 1	**Modèle 2**	**Modèle 3**	**Modèle 4**
Count	62865			
Mean	-1760 ±150	440 ±75	-4710 ±130	-390 ±74
Std	18870	9590	16220	9570

TABLE 3 – Erreur moyenne absolue en km par modèle (K=5)

Les quatre modèles fournissent des résultats assez proches puisque la méthode de calcul diffère très peu d'un modèle à l'autre. Cependant, les performances des modèles d'interpolation sont meilleurs pour prédire le bon kilométrage. L'erreur moyenne absolue ainsi que son écart-type est plus faible et souligne donc une meilleure précision. La méthode d'interpolation par spline monotone permet de capturer davantage de contexte dans les données puisqu'elle conserve la monotonie des séquences, réduisant ainsi les erreurs d'approximation et elle tient compte également de la tendance sur plusieurs points, améliorant ainsi la fidélité des prédictions. Par ailleurs, en traçant les moyennes kilométriques annuelles calculées par les différentes méthodes, on remarque que le modèle 4 affiche une légère baisse des distances parcourues en 2020, ce qui semble être cohérent avec la pandémie de la Covid-19. C'est donc cette méthode que nous allons utiliser.

3.3 Résultats d'estimation

A partir des données de vérifications mécaniques cohérentes, nous appliquons la méthode d'interpolation par spline monotone pour obtenir une estimation kilométrique pour 146 586, soit 44,6% des véhicules. Ensuite, on procède à une différence entre chaque année pour obtenir la distance parcourue pendant une année.

On note $d_{i,j}$ la distance annuelle du camion i à l'année $j \in [2011; 2022]$. En considérant que toutes les distances annuelles sont indépendantes et suivent la même loi, nous pouvons appliquer la loi forte des grands nombres pour la moyenne empirique des distances parcourues à l'année j. Soit $\hat{d}_{j,n}$ cette moyenne pour n camions :

$$\hat{d}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} d_{i,j}$$

Par la loi forte des grands nombres, $\hat{d}_n$ converge presque sûrement vers la distance moyenne annuelle :

$$\hat{d}_{j,n} \xrightarrow{p.s.} \mu_j \quad \text{lorsque} \quad n \to \infty$$

.

Sur la figure 9, on observe une augmentation globale des distances parcourues par les camions avec une moyenne de 44900km par an. On remarque une faible baisse à partir de 2019 (années de pandémie). Il est possible que la moyenne de l'année 2019 soit légèrement sous-estimée, puisque le modèle tend à lisser les irrégularités. De la même manière, la proportion plus faible de véhicules circulant en 2022 répertoriés dans la base de données influencent également à la baisse l'estimation pour cette année-ci.

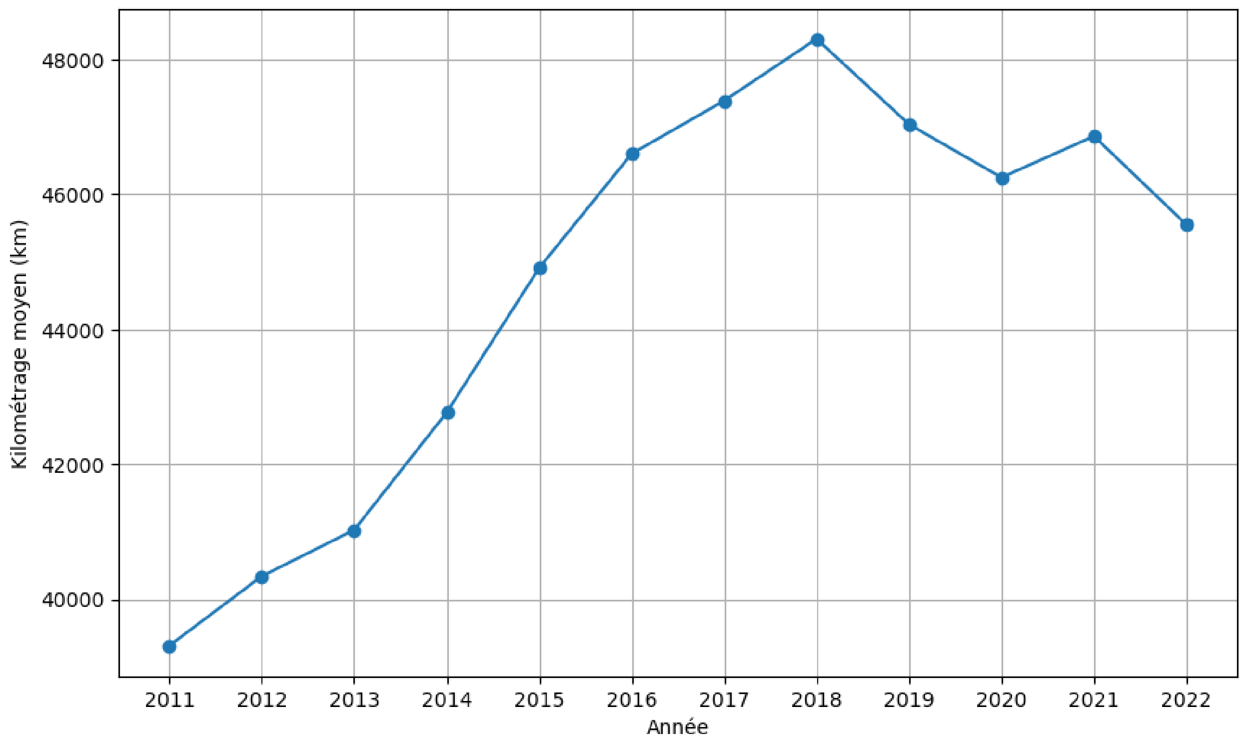

FIGURE 9 – Evolution de la distance moyenne annuelle parcourue

En distinguant par classe de véhicules, on observe dans le tableau 4 que les véhicules lourds de transport de marchandise (BCA) parcourent davantage de distance que les autres. Ces moyennes seront intéressantes à comparer aux résultats définitifs après complétion par Machine Learning afin de vérifier le maintien de ces tendances.

Classe	Distance (en km)
BCA	45700
COT	26600
HCA	20000
RCA	7500

TABLE 4 – Distance annuelle moyenne par classe de véhicules

4 Prédiction pour les estimations manquantes

Une fois les distances annuelles parcourues par 146 586 véhicules estimées, il reste encore 182 191 sans estimation. Comme nous disposons de données géographiques, personnelles et techniques sur les camions, il est intéressant d'étudier la potentialité d'algorithme d'apprentissage pour déterminer les valeurs restantes.

4.1 Analyse et choix des features

Tout d'abord, nous avons décidé de modifier légèrement le format des données. En effet, au départ, pour chaque véhicule, nous disposions d'une colonne par année avec sa distance parcourue. Nous pivotons les données afin cette fois-ci d'avoir une ligne par véhicule et par année, le nouvel index devient alors NIV_ANNEE. Cette transformation permet de faciliter l'analyse du lien entre l'année et d'autres variable. De plus, cela permettra au modèle d'apprentissage de capturer l'évolution temporelle des kilomètres.

Pour commencer, nous avons analysé les corrélations linéaires qui peuvent exister entre les différentes features. Ces corrélations sont représentées sur la figure 22 en annexe. On remarque que le kilométrage annuel est corrélé au type de camion, c'est-à-dire à son nombre d'essieux et sa cylindrée. De façon prévisible, l'année du modèle a un impact sur les kilométrages annuels.

Nous avons également exploré la distribution du kilométrage par région administrative, classe de véhicule et type de carburant (figures 23,24,25 en annexe). Ces graphiques nous indiquent que la région à laquelle le camion est rattachée joue sur les kilométrages parcourus. Conjointement, l'indication du MRC et du CP3 permettant d'affiner la localisation géographique sera également intéressante à prendre en compte. La classe du véhicule a de façon évidente un impact sur son utilisation donc sur les distances parcourues. Les camions commerciaux ou professionnels roulent davantage que les véhicules privés ou hors réseaux. Le type de carburant influence également l'autonomie du véhicule, ce qui agit aussi les kilomètres parcourus annuellement.

Par ailleurs, le profil du propriétaire du véhicule correspond à une feature importante. En moyenne, un véhicule d'entreprise parcourt 44300km par an, tandis qu'un véhicule de particulier en parcourt 24500km.

Nous choisissons ainsi les features suivantes pour notre régression :

- ANNEE
- MASSE_NETTE
- NB_CYL
- CYL_VEH
- NB_ESIEU_MAX
- NIV
- CP3
- CLAS
- MARQ_VEH
- MODEL_VEH
- ANNEE_MOD
- TYP_CARBU
- TYP_DOSS_PERS
- REG_ADM
- MRC

Pour un bon fonctionnement de la régression, un travail sur les valeurs manquantes est nécessaire (confère tableau 5). Dans le tableau, nous avons exhibé les données d'entrainement de l'ensemble des données ; ce qui permet de voir que les valeurs manquantes sont homogènement réparties sur l'intégralité des données. La méthode de complétion des valeurs manquantes n'ajoutera donc pas de biais entre données d'entrainement et données à prédire. Afin de pallier les manques, nous avons mis en œuvre une méthode d'imputation basée sur le regroupement. Cette approche consiste à regrouper les véhicules selon des critères spécifiques (CP3 ou modèle), permettant ainsi de récupérer les données géographiques ou techniques du groupe pour ensuite les attribuer aux valeurs manquantes. La procédure s'effectue en plusieurs étapes :

1. Imputation des CP3, marque et modèle en se basant sur les valeurs les plus fréquentes (très peu de véhicules concernés, moins de 0.1%, cela n'affectera pas significativement l'analyse)
2. Imputation des caractéristiques techniques (masse, nombre de cylindres, cylindrée, nombre maximum d'essieux et année de modèle) par regroupement de modèles.
3. Imputation des données géographiques (région administrative et municipalité) par regroupement de CP3.

A l'issue de ce processus, on remarque qu'il reste tout de même beaucoup de données géographiques manquantes. En regardant les données, nous avons réalisé qu'il s'agit de véhicules immatriculés hors Québec. Par conséquent, nous avons décidé d'ajouter une nouvelle feature : $Province$, créée à partir de la première lettre du $CP3$ et qui distinguera ainsi les véhicules par un nouvel échelon géographique. Ne pouvant donc pas remonter aux données de région administrative ou de MRC pour ces véhicules, on fixera les valeurs manquantes par $Autre$. Par ailleurs, les quelques dernières valeurs numériques manquantes sont complétées par la médiane de la colonne correspondante.

Nous avons ensuite encodé les variables catégorielles en utilisant du One-hot encoding. Nous avons également crée une autre feature correspondant à l'âge du véhicule à l'année courant.

Entrée	**Initial**		**Complétion par regroupement**	
	Train	**Tout**	**Train**	**Tout**
ANNEE, NIV, CLAS, TYP_CARBU, TYP_DOSS_PERS	0		0	
CP3	11	33	0	0
MARQ_VEH	177	726	0	0
MODEL_VEH	1400	4079	0	0
ANNEE_MOD	1	2	0	0
MASSE_NETTE	119	270	1	13
NB_CYL	1912	3414	1	13
CYL_VEH	6208	15922	264	996
NB_ESIEU_MAX	5483	13114	29	224
REG_ADM	3010	6262	2952	6140
MRC	3010	6262	2952	6140

TABLE 5 – Nombre de valeurs manquantes par entrée avant et après complétion

4.2 Comparaison d'algorithmes de régression

Nous avons testé plusieurs algorithmes (tableau 6) de régression afin de trouver le modèle le plus performant pour la prédiction du kilométrage. Certains modèles nécessitant une normalisation des données, nous avons utilisé le *StandardScaler* du module *Scikit-learn* pour les variables continues, qui centre et réduit chaque ensemble de variables.

Pour chaque modèle, nous avons évalué la performance en utilisant les métriques R^2 (score de régression) et RMSE (Root Mean Squared Error). De plus, nous calculons également le temps d'exécution de chaque algorithme. Les résultats obtenus sont présentés dans le tableau 6. Les scores sont calculés sur un échantillon test correspondant à 20% des données d'entrainement, ce qui permet de vérifier si le modèle présente du surapprentissage. On remarque que les régressions basées les *DecisionTree* et *RandomForest* réalisent des meilleurs scores que les autres. Le résultat semble logique car ces algorithmes sont capables de capturer des relations complexes et non linéaires dans les données, ce qui leur permet de mieux modéliser les interactions et les dépendances entre les variables. En particulier, le *RandomForestRegressor* combine plusieurs arbres de décision pour réduire la variance et éviter le surapprentissage, ce qui améliore la robustesse et la précision des prédictions. Naivement, on pouvait prédire cette performance puisqu'avec les données disponibles, leur hétérogénéité et leur diversité, une classification des caractéristiques des camions permet de capturer davantage l'impact de celles-ci sur les kilométrages annuels.

Modèle	$\mathbf{R^2}$	**RMSE**	**Temps d'exécution** (en secondes)
Linear Regression	0.561	34100	49
Hist Gradient Boosting	0.670	29600	88
LightGBM	0.670	29600	5.5
CatBoost	0.700	28200	19
Decision Tree	0.736	26500	130
RandomForest	0.844	20300	2600

TABLE 6 – Résultats des différents modèles de régression

4.3 Prédiction

Après avoir comparé les différents modèles, nous avons décidé d'affiner les hyperparamètres du modèle *RandomForestRegressor*. L'article [2] explique que le réglage des hyperparamètres peut améliorer les performances du modèle, même si les gains peuvent être toutefois modérés.

Afin d'optimiser les hyperparamètres de notre modèle *RandomForestRegressor*, nous avons décidé d'implémenter une optimisation bayésienne, proposée dans l'article [2]. Cette méthode itérative consiste à rechercher les meilleurs paramètres en maximisant une fonction cible. Notre objectif consistera à maximiser le score R^2. Nous ajoutons un terme de pénalisation afin de limiter l'overfitting (écart absolue entre le score R^2 calculé sur les données d'entrainement et les données tests). Cette fonction est disponible en annexe 8.3. Contrairement aux méthodes de *GridSearch* ou *RandomSearch*, l'optimisation bayésienne est plus efficace et permet d'explorer l'espace des hyperparamètres de manière ciblée. L'analyse de l'évolution des résultats permet également de conclure sur l'impact d'un hyperparamètre. Cette méthode est cependant plus coûteuse parfois en termes de calcul et dépend fortement du paramétrage initial. Ainsi, il est aussi intéressant de combiner différentes approches, notamment pour choisir les paramètres discrets.

Après optimisation, on choisit les hyperparamètres suivants :

- Nombre d'estimateurs (*n_ estimators*) : 300

- Nombre minimal d'échantillons pour diviser un nœud (*min_samples_split*) : 2
- Profondeur maximale de l'arbre (*max_depth*) : None
- Nombre minimal d'échantillons par feuille (*min_samples_leaf*) : 1
- Nombre maximal de caractéristiques à considérer pour diviser un nœud (*max_features*) : sqrt

Nous appliquons donc le modèle RandomForest avec les hyperparamètres choisis et nous obtenons les résultats répertoriés dans le tableau 7.

	R^2	RMSE
Avant fine-tuning	0.844	20300
Après fine-tuning	0.845	19900

TABLE 7 – Effet du fine-tuning

L'effet de l'optimisation des hyperparamètres ne semble pas significative, mais elle permet tout de même de réduire l'écart quadratique moyen d'environ 400km. Son effet sera certainement amplifié lors de l'application de l'apprentissage à l'ensemble des données.

4.4 Résultats de prédiction

Le processus de traitement a donc été appliqué de manière uniforme à l'ensemble du jeu de données. Nous avons ensuite entrainé le modèle de forêt aléatoire fine-tuné sur l'intégralité des données d'entrainement et nous avons prédit les kilométrages pour les 176 074 camions. Nous obtenons une moyenne annuelle de 42750km (contre 44900km avant ML).

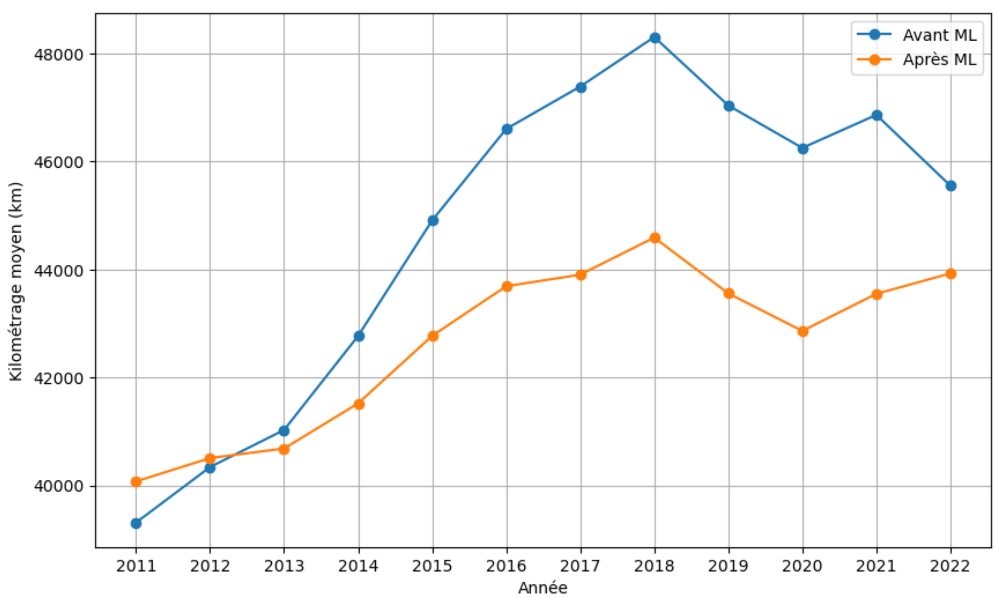

FIGURE 10 – Evolution de la distance moyenne annuelle parcourue avant et après complétion par Machine Learning

Classe	Avant ML	Après ML
BCA	45700	43800
COT	26600	24900
HCA	20000	20800
RCA	7500	7500

TABLE 8 – Distance annuelle moyenne par classe de véhicules

A partir de la figure 10 et du tableau 8, on observe que les prédictions réalisées par l'algorithme d'apprentissage diminuent légèrement les valeurs moyennes des distances annuelles parcourues. Cependant, la tendance générale de la courbe ainsi que la répartition des kilométrages en fonction des classes de véhicules restent semblables. Cela semble donc cohérent. Le *RandomForestRegressor* a donc certainement capturer le lien important entre la classe d'un véhicule et son utilisation. On peut donc conclure que les prédictions réalisées n'ont pas sensiblement modifié les ordres de grandeurs. Cela peut être du à un overfitting sur les données d'entrainement. Par ailleurs, on observe que la moyenne annuelle repart à la hausse après 2020. Cela s'explique par le fait que beaucoup de véhicules circulant en 2022 n'avaient pas été estimées avant le ML et les résultats sont donc d'autant plus cohérents en aggrégeant le ML aux premières estimations.

5 Émissions de CO_2

5.1 Méthodes de conversion

Pour chacun des 328 777 véhicules, nous avons à disposition un kilométrage parcouru chaque année de circulation, auquel on peut ajouter une information sur la masse du véhicule ainsi que le type de carburant utilisé. Le but de cette partie est d'expliquer les différentes méthodes permettant de réaliser l'estimation des émissions de CO_2 équivalent par véhicule chaque année à partir de ces données. Chaque méthode permet une approximation plus ou moins fine des émissions de gaz à effet de serre.

Une première méthode naïve consiste à prendre en compte uniquement la distance parcourue dans l'année. On calcule ainsi les émissions selon la formule suivante :

$$E_i = d_i \cdot F_e$$

où :

- ⋆ E_i désigne les émissions de CO_2 du véhicule i
- ⋆ d_i désigne la distance parcourue par le véhicule i
- ⋆ $F_e = 903\ gCO_2 - eq/km$ correspond au facteur d'émissions moyennes d'un camion en France [3]

Une deuxième méthode consiste à considérer la masse nette du véhicule comme donnée d'entrée supplémentaire pour estimer les émissions de CO_2. Cette méthode permet d'approcher les émissions réelles car la consommation est directement reliée à la masse. La formule utilisée est la suivante :

$$E_i = d_i \cdot m_i \cdot F_{tkm}$$

où :

- ⋆ m_i désigne la masse nette du véhicule i
- ⋆ $F_{tkm} = 191.7\ gCO_2 - eq/tkm$ correspond au facteur d'émission moyenne par tonne-kilomètre [2] d'un camion au Canada [4]

Une troisième méthode plus précise consiste à considérer également le type de carburant. En effet, même si la majorité des camions roulent au diesel et à l'essence, il est intéressant de développer un calcul plus poussé permettant de différencier les émissions des camions électriques, hydrogènes, hybrides, etc. Le calcul s'effectue de la façon suivante :

Pour l'estimation, on s'appuiera sur les données de carburant et d'efficacité énergétique répertoriées dans les tableaux 9,10.

PTAC	**Efficacité énergétique (gep/tkm)**
3,5 à 16t	75,7
16 à 32t	55,1
>32t	41,5
Tous les camions	45

TABLE 9 – Efficacité énergétique [5]

Type de carburant	**Facteur d'émission CO_2**	**Densité énergétique**
Diesel	3,00 kg/L	38,68 MJ/L
Essence	2,37 kg/L	31,54 MJ/L
Méthanol	1,52 kg/L	15,74 MJ/L
Ethanol	0,082 kg/L	21,00 MJ/L
Gaz naturel	1,18 kg/L	22,50 MJ/L
GPL	1,54 kg/L	25,13 MJ/L
Electrique	34,5 g/kWh	3,6 MJ/kWh
H2	75,0 g/kg	119,74 MJ/kg
Hybride	Mix 20% électrique et 80% essence	/

TABLE 10 – Données de carburants [6], [7]

En sachant que 1 $gep = 0,042\ MJ$, on peut donc calculer les émissions CO_2 à partir de la formule suivante :

$$E_i = \frac{d_i \cdot m_i \cdot Ef_i \cdot Fe_i}{\rho_i} \tag{1}$$

où :

2. La tonne-kilomètre (tkm) est une unité de mesure couramment utilisée en transport de marchandise et sert à mesurer l'impact environnemental.

- ⋆ Ef_i correspond à l'efficacité énergétique du véhicule i
- ⋆ Fe_i est le facteur d'émission du carburant
- ⋆ ρ_i est la densité énergétique du carburant

Dans nos données, certains véhicules ont pour type de carburant : 'Autre'. Nous utilisons à nouveau un processus permettant de récupérer la donnée pour des camions de même marque et de même modèle. Cela permet ainsi d'avoir le bon carburant pour chaque camion. Pour les véhicules dont le carburant est de type 'non propulsé', on considère qu'il n'y a pas d'émissions de CO_2. De plus, n'ayant pas accès au PTAC, nous utiliserons donc la valeur moyenne pour tous les camions proposée dans le tableau 9.

5.2 Choix de la méthode de calcul

De manière naïve, la méthode 3 semble être la meilleure car elle offre la plus grande précision en prenant en compte la distance parcourue, la masse du véhicule, ainsi que le type de carburant et l'efficacité énergétique. Les méthodes 1 et 2 sont plus simples et dépendent grandement des constantes choisies, qui sont souvent des moyennes globales d'émissions. De surcroit, le fait de s'appuyer sur les types de carburant ainsi que sur l'efficacité énergétique dans la méthode 3 paraît plus intéressant dans les années récentes et futures, puisque le nombre de véhicules électriques, hybrides ou à hydrogène va augmenter. En termes de complexité de calcul, les trois méthodes sont assez équivalentes.

Les trois méthodes fournissent les résultats répertoriés dans le tableau 11.

Méthode	**Emissions annuelles moyennes ($MtCO_2eq$)**	**Emission annuelle moyenne par camion (tCO_2eq)**
1	6.0	42.5
2	10.0	68.6
3	7.7	52.3

TABLE 11 – Résultats généraux des différentes méthodes de calcul des émissions de CO_2

Pour comparer les résultats, nous avons récupéré les estimations d'émissions des véhicules lourds effectuées par le gouvernement québecois [1].

Pour donner davantage de contexte, le gouvernement québecois s'est appuyé sur les données calculées à l'échelle canadienne. Comme expliqué dans le rapport d'inventaire national [8], la méthode utilise principalement le modèle MOVES3, développé par l'Agence de protection de l'environnement des États-Unis. Les émissions sont calculées en plusieurs étapes :

- **Données sur les parcs de véhicules** : Les véhicules sont classés par type de carburant, type de carrosserie, et poids nominal brut. Les données sont issues des bases de données d'immatriculation de Statistique Canada et des consultants spécialisés.
- **Pénétration technologique** : On estime le nombre de véhicules équipés de technologies antipollution, en tenant compte des normes réglementaires de différents niveaux (Niveau 1 à 3).

— **Taux d'accumulation de kilométrage** : Ces taux mesurent le kilométrage annuel moyen des véhicules, basés sur les lectures d'odomètres en Ontario et en Colombie-Britannique, appliqués à l'ensemble du pays.

— **Taux de consommation de carburant** : MOVES3 incorpore des taux de consommation sous forme de taux d'énergie (kJ/s) variables en fonction de nombreux paramètres (classe de véhicule, vitesse, etc.). Ces taux sont convertis en volume de carburant en utilisant des facteurs de conversion énergétiques.

Les données extraites pour le Québec sont donc issues d'un modèle globale basé sur des données moyennes de deux autres régions. Leur fiabilité n'est donc pas parfaite. La comparaison des modèles aux données gouvernementales permet donc de tirer des conclusions relatives. Il est cependant intéressant de comparer l'approche micrométrique réalisée dans cette étude par rapport à celle macrométrique du gouvernement.

En termes d'émissions annuelles moyennes, le rapport gouvernemental fournit une moyenne de 7.4 $MtCO_2eq$, ce qui s'approche de celle du modèle 3. Sur la figure 11, on a tracé l'évolution des émissions annuelles totales. On remarque tout d'abord que les ordres de grandeurs estimés sont proches. On observe également que le modèle 3 semble le plus proche des données gouvernementales. Cela semble logique puisque la méthodologie est assez proche. Les trois méthodes fournissent toutefois une croissance plus ou moins linéaire des émissions tandis que les données gouvernementales semblent fluctuées davantage.

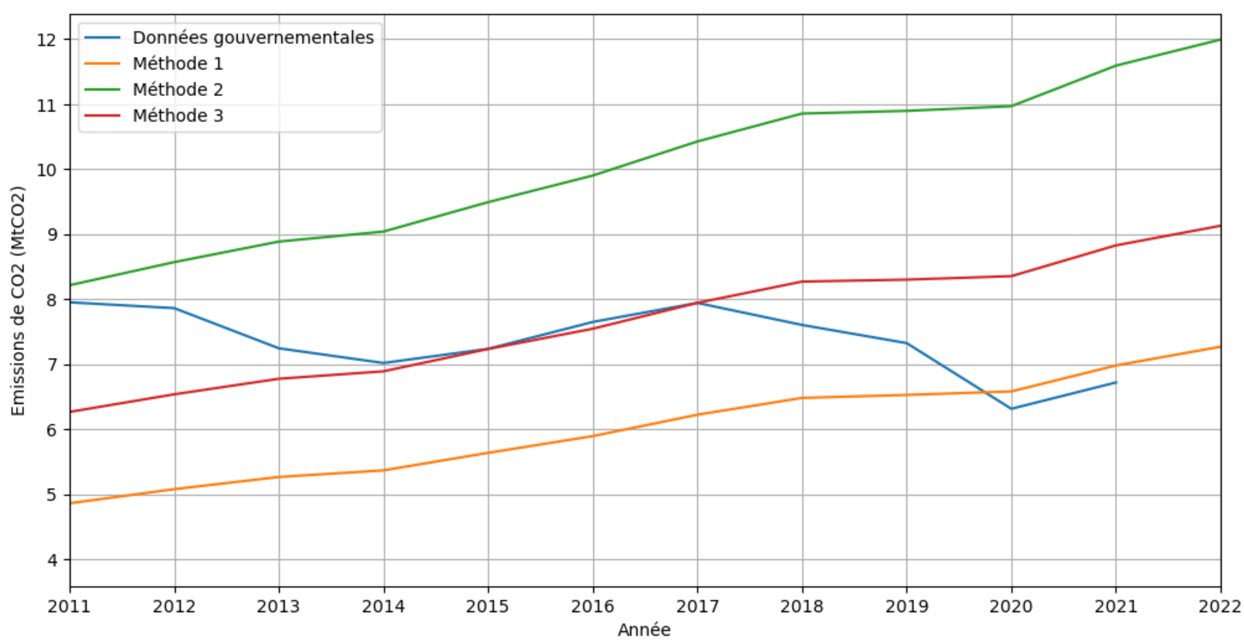

FIGURE 11 – Emissions totales annuelles des véhicules lourds au Québec

Ainsi, la méthode 3 semble aboutir à une meilleure précision, puisqu'elle capture à la fois les paramètres majeurs des émisisons (distance, masse) mais elle permet aussi une distinction plus fine vis-à-vis des véhicules peu émetteurs de CO_2, dont l'effectif a augmenté dans les dernières années (restant toujours en proportion assez faible). Pour l'analyse finale, nous garderons les émissions de CO_2 réalisées par la méthode 3.

6 Résultats finaux

6.1 Analyse

6.1.1 Bilan général

Notre étude a permis d'obtenir une estimation des kilomètres annuels ainsi que des émissions de CO_2 associées pour l'intégralité des 328 777 camions. Pour 44,6% d'entre eux, qui disposaient de données en quantité et qualité suffisantes, nous nous sommes basés sur une approche statistique. Ensuite, nous avons complété les estimations par une approche d'apprentissage pour les véhicules restants.

De manière globale, en considérant que tous les camions sont indépendants et que leur distance annuelle est identiquement distribuée, on obtient une distance annuelle moyenne de **42750 km**. Sur la figure 12, on a tracé l'évolution du kilométrage annuel. Depuis 2011, on assiste à une intensification progressive du transport routier de marchandises au Québec. En 2020, l'effet de la pandémie est visible sur la courbe et on remarque que l'activité routière est repartie à la hausse à l'issue.

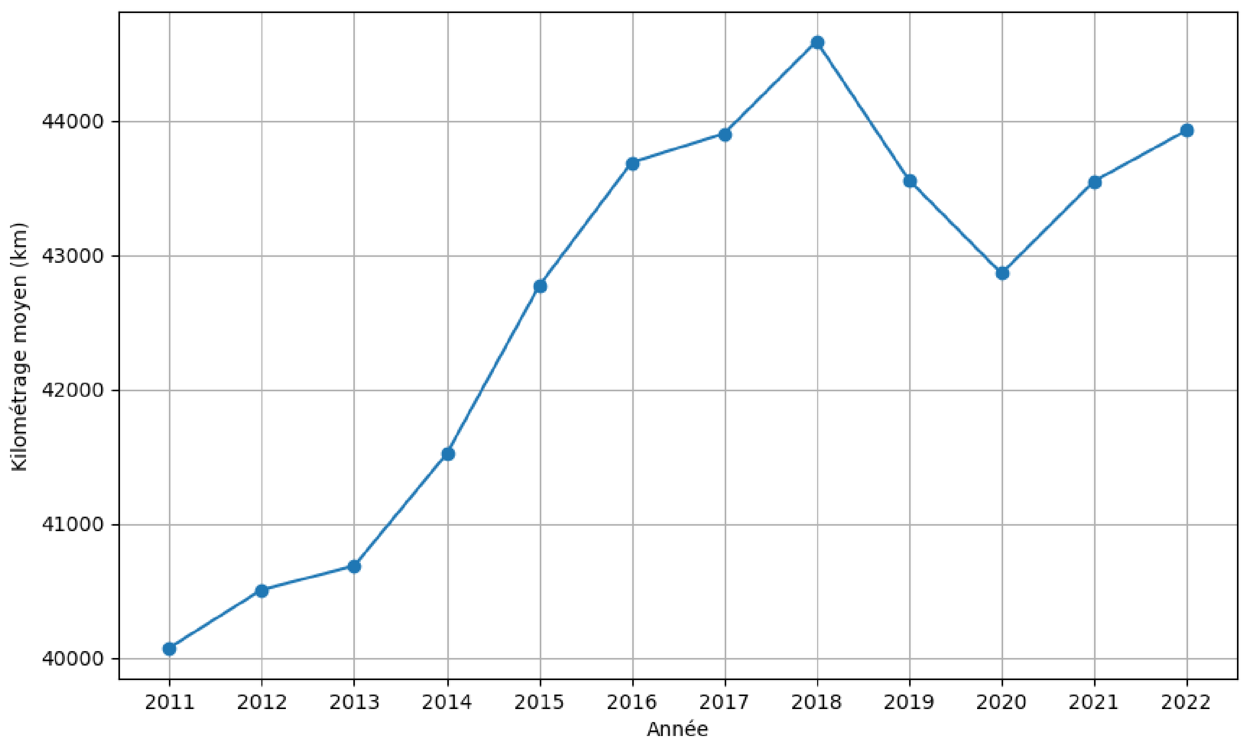

FIGURE 12 – Kilométrage annuel moyen des camions analysés

Afin d'affiner les valeurs moyennes obtenues, on les distingue dans le tableau 12 en fonction de la classe du véhicules. La majorité du transport routier correspond aux camions lourds de type 'BCA'.

Classe	Distance (km)	Proportion
BCA	43800	96%
COT	24900	2.1%
HCA	20800	1.8%
RCA	7500	0.1%

TABLE 12 – Distance annuelle moyenne et proportion par classe de véhicules

En ce qui concerne les émissions de gaz à effet de serre, on obtient une moyenne annuelle par véhicule de **52.3 tCO_2eq** et pour l'ensemble de la flotte : **7.7 MtCO_2eq** par année. Ces ordres de grandeur sont cohérents avec les estimations réalisées par le gouvernement tout en estimant davantage d'émissions que ce dernier.

6.1.2 Analyse détaillée

A l'aide des données géographiques d'immatriculation des véhicules, on obtient les figures 26 et 27 en annexe. En moyenne, on observe que les véhicules qui roulent le plus sont immatriculés entre Montréal et Québec, qui correspondent aux zones de forte activité économique. D'un point de vue des émissions, la plupart des camions étant immatriculés dans la région de Montréal et de Montérégie (localisation des transporteurs), ces régions affichent de grandes valeurs d'émissions de CO_2. Comme nous ne disposons pas de données sur les trajets des camions, la répartition géographique des émissions est certainement différente mais se concentre essentiellement autour du Saint-Laurent, entre Montréal et Québec.

Pour la suite, les graphiques représentent une répartition moyenne pour une année de la contribution aux émissions selon divers paramètres des camions.

D'un point de vue de la masse des véhicules, on a représenté sur la figure 13 les émissions de CO_2 en fonction de la catégorie de masse des véhicules. Les véhicules entre 8 et 12 tonnes sont les plus polluants, suivis par ceux de 5 à 8 tonnes. Les véhicules de plus de 20 tonnes, bien qu'ayant une capacité de charge plus élevée, sont moins représentés dans le parc automobile, ce qui explique leur contribution moindre aux émissions globales. Ces observations soulignent l'importance de cibler les véhicules de poids moyen pour les stratégies de réduction des émissions. Par exemple, l'électrification de la flotte de camions devraient se concentrer sur cette catégorie de poids.

A l'aide de la figure 28 en annexe, on observe que la majorité des émissions sont dues aux véhicules datant d'après 2000 (plus de 60% pour les modèles après 2010). Ce résultat est logique puisque ce sont les camions les plus récents qui parcourent le plus de distance et donc qui émettent le plus. Toutefois, on remarque tout de même que les vieux modèles de camions (dont les facteurs d'émissions sont plus élevés) représentent 10% des émissions de CO_2.

Enfin, l'analyse de la figure 29 de l'annexe montre la distribution des émissions de CO_2 par région administrative. En corrélation avec la carte 26, les résultats indiquent que les régions de Montréal et de la Montérégie sont les plus contributrices en termes d'émissions de CO_2, représentant une part significative du total des émissions. Cette tendance peut être expliquée par l'implantation de la plupart des entreprises de transport dans ces régions, entraînant un

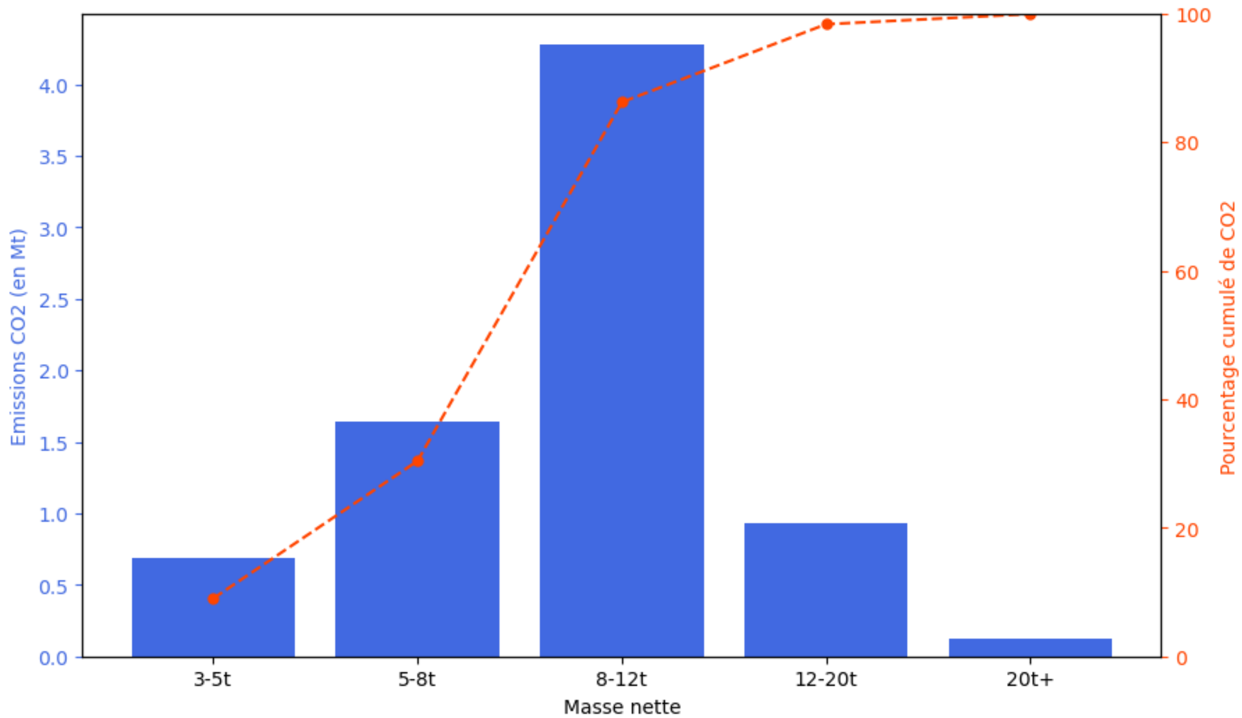

FIGURE 13 – Répartition des émissions de CO_2 par catégorie de masse des véhicules

nombre élevé de véhicules en circulation immatriculés dans ces zones. D'autres régions, telles que l'Abitibi-Témiscamingue et la Côte-Nord, affichent des émissions nettement inférieures, probablement en raison de la moindre densité de population et du trafic routier réduit.

6.1.3 Evolution temporelle

Entre 2011 et 2022, on observe une augmentation quasi-linéaire des émissions de CO_2 du transport routier. Sur la figure 14, on remarque une lègère atténuation de cette augmentation pendant l'année de pandémie, mais reprend la même tendance à l'issue. Cette augmentation est de l'ordre de **260 ktCO_2eq** par année.

La figure 15 nous montre la proportion relative des différentes catégories de masse de camions au cours du temps. On observe à nouveau l'augmentation de la part des camions de 8 à 12t, dont le nombre et donc les émissions ont augmenté ces dernières années.

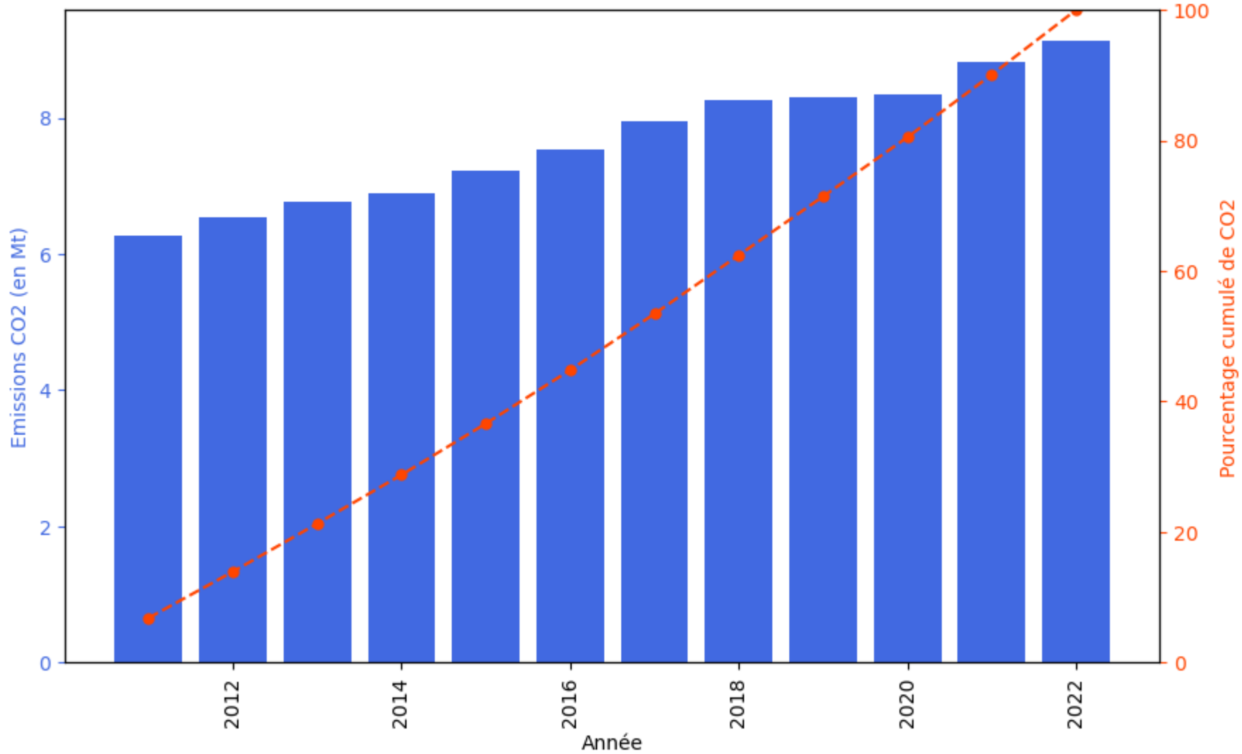

FIGURE 14 – Evolution des émissions de CO_2 au cours du temps de l'ensemble de la flotte

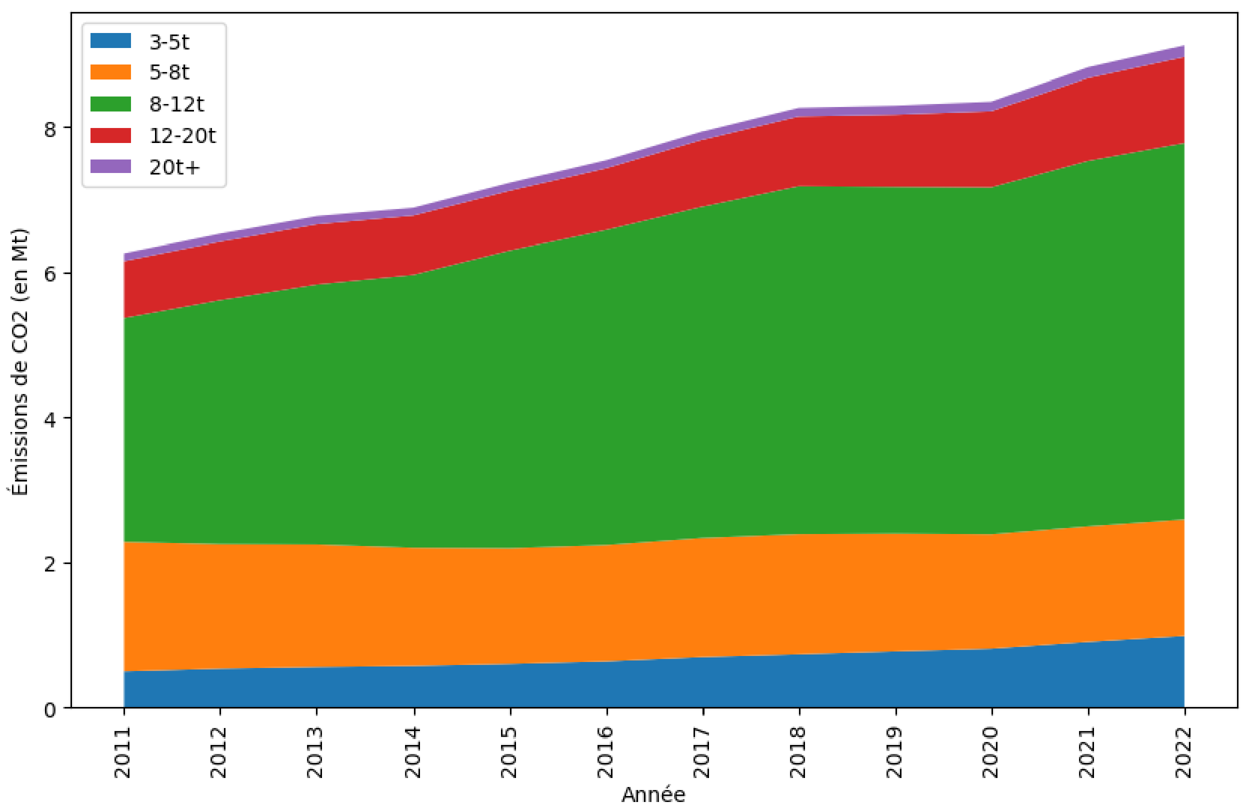

FIGURE 15 – Evolution des émissions de CO_2 par catégorie de masse

6.2 Limitations

Tout d'abord, cette étude s'est basée sur les données issues de la Société d'assurance automobile du Québec. Ces données n'avaient auparavant été que très peu voire par pas analysées en ce qui concerne les camions. Les données comportaient beaucoup de valeurs manquantes, des erreurs de saisie humaines, des incohérences internes ou des NIV non uniques. Le travail réalisé a donc essayé de tenir compte de ces incohérences afin de réduire au maximum les erreurs associées dans les estimations. Cependant, il est à noter que la qualité première des données influence nécessairement la précision des estimations obtenues.

Si les estimations kilométriques basées sur l'interpolation permettent d'obtenir des intervalles de confiance relativement faibles, les prédictions réalisées par le *RandomForestRegressor* laissent une marge d'erreur significative au vue de la moyenne kilométrique annuelle. Les estimations individuelles pour chaque camion peuvent donc comporter des erreurs importantes, mais le grand nombre de camions dans l'échantillon assure que les moyennes annuelles obtenues sont représentatives et précises.

En outre, la complétion du CP3, des marques et des modèles de véhicules peut introduire un certain effet sur les résultats, puisque cette complétion a été basée sur la valeur la plus fréquente. Cependant, la très faible proportion de valeurs manquantes dans ces trois catégories de données permet de limiter l'effet produit. Par ailleurs, la faible diversité de données au sein de l'ensemble d'entrainement peut provoquer un overfitting non négligeable. Cette faible diversité se retrouve dans l'ensemble de prédiction, dont les principaux paramètres ressemblent à ceux du jeu d'entrainement. Cela réduit ainsi l'effet du surapprentissage vis-à-vis de la qualité des prédictions.

Concernant la modélisation des émissions de gaz à effet de serre, il est important de souligner que les modèles choisies sont assez simples et comprennent des hypothèses génériques sur les véhicules. En effet, ne disposant pas de l'information sur la charge utile des véhicules, les calculs d'émissions se basent sur des véhicules à vide. Les émissions sont donc sous-estimées, mais permettent d'avoir une estimation de leur borne inférieure.

6.3 Perspectives

Tout d'abord, il serait intéressant d'appliquer la méthodologie développée dans cette étude à d'autres catégories de véhicules, comme les autobus, les voitures des particuliers ou autres. Toutefois, la qualité des données kilométriques joue un rôle majeur dans la méthodologie utilisée et il est fort possible que le manque de données de vérifications mécaniques (jugées plus fiables) altèrent la précision des estimations. En effet, les contrôles techniques sont moins fréquents pour les véhicules particuliers que pour les camions et cela réduit grandement la quantité de données disponibles.

Concernant le travail effectué sur les camions, nous n'avons considéré que les données de vérifications mécaniques dans le calcul final. L'utilisation conjointe des données de transactions ainsi que des vérifications pourraient permettre d'agrandir l'échantillon de camions à données cohérentes et ainsi élargir l'ensemble d'entrainement pour le Machine Learning.

Comme expliqué précédemment, le modèle d'estimation des émissions de CO_2 pourrait être amélioré. Pour cela, il faudrait collecter des données sur les charges utiles des véhicules pour affiner les estimations et obtenir des résultats plus réalistes. Une approche consisterait à utiliser le guide gouvernemental de charge [9], ce qui nécessiterait d'obtenir davantage d'informations techniques sur les camions. Par ailleurs, le modèle d'estimation des émissions ne tient pas compte d'une éventuelle amélioration technologique sur l'efficacité énergétique des camions. Cela pourrait également affiner les résultats. Enfin, quantifier d'autres types d'émissions, telles que le méthane (CH4) et le monoxyde de carbone (CO), pourrait également enrichir l'analyse et fournir une vision plus large de l'impact environnemental des camions.

7 Conclusion

L'étude réalisée a permis de développer une méthode d'estimation des émissions de CO_2 des véhicules commerciaux au Québec à partir de données kilométriques. L'analyse approfondie des données issues des vérifications mécaniques et des transactions liées aux véhicules a mis en évidence la présence d'un certain nombre d'anomalies nécessitant un prétraitement rigoureux. Les véhicules ont donc été scindés en deux groupes : ceux disposant de données cohérentes (croissance des kilométrages dans le temps avec au minimum deux points) et les autres. Pour le premier groupe, différentes méthodes de calcul ont été testées et il a été décidé d'utiliser une interpolation par spline monotone afin d'estimer les kilométrages annuels des véhicules. Pour le deuxième groupe, une approche d'apprentissage a été mise en oeuvre. Après analyse des corrélations et complétion des valeurs manquantes, l'algorithme *RandomForestRegressor* a démontré les meilleurs résultats et a permis de prédire les kilométrages restants à déterminer. Enfin, les émissions de CO_2 ont été estimées à partir des distances annuelles parcourues ainsi que d'autres données spécifiques à chaque véhicule.

Les résultats aboutissent à une distance annuelle moyenne de 42750km et montrent une augmentation globale des émissions de CO_2, avec une moyenne annuelle de 52,3 tonnes par camion et un total de 7,7 Mt pour l'ensemble de la flotte. Ces estimations dépassent légèrement celles faites par le gouvernement et montrent une augmentation quasi-linéaire des émissions. Plusieurs pistes d'amélioration peuvent être envisagées telles que la prise en compte de la charge utile des camions dans le calcul des émissions ou l'intégration des données transactionnelles.

Remerciements

Les auteurs désirent remercier la Société de l'assurance-automobile du Québec pour la fourniture des données utilisées dans cette étude. Les auteurs remercient également le Ministère de l'Économie, de l'Innovation et de l'Énergie du Québec, qui finance la Chaire en transformation du transport.

8 Annexe

8.1 Données

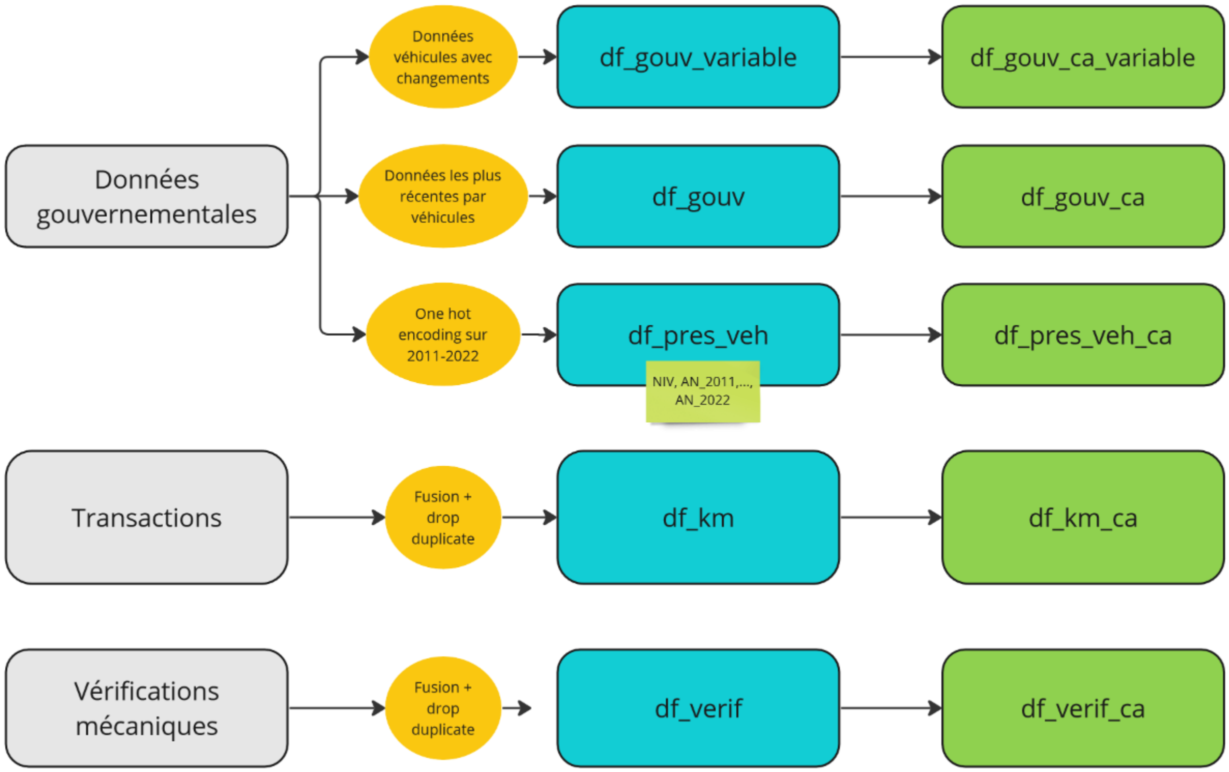

FIGURE 16 – Séquence de pré-traitement

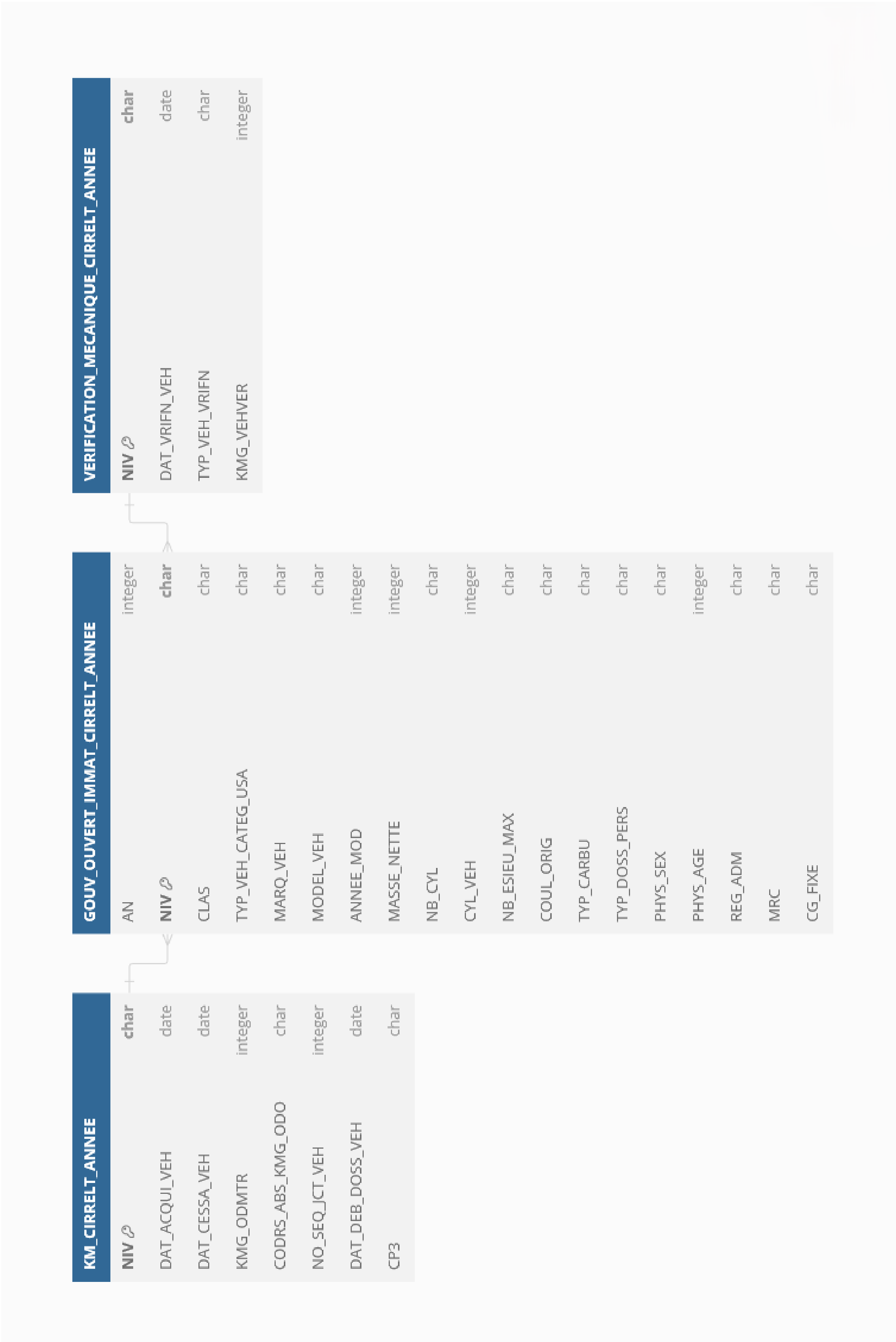

FIGURE 17 – Représentation des liens entre les bases de données

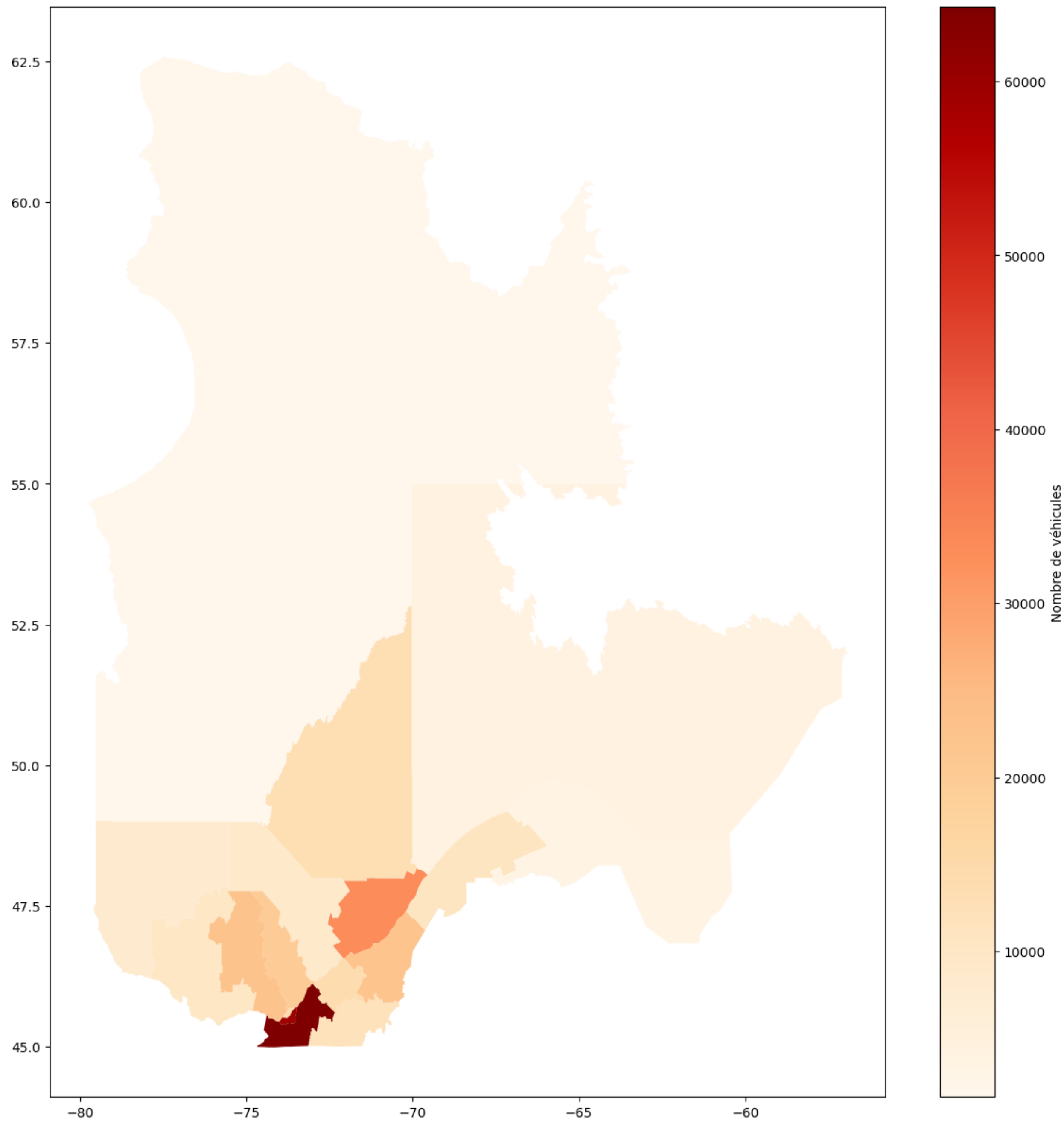

FIGURE 18 – Répartition des véhicules dans les différentes régions administratives du Québec

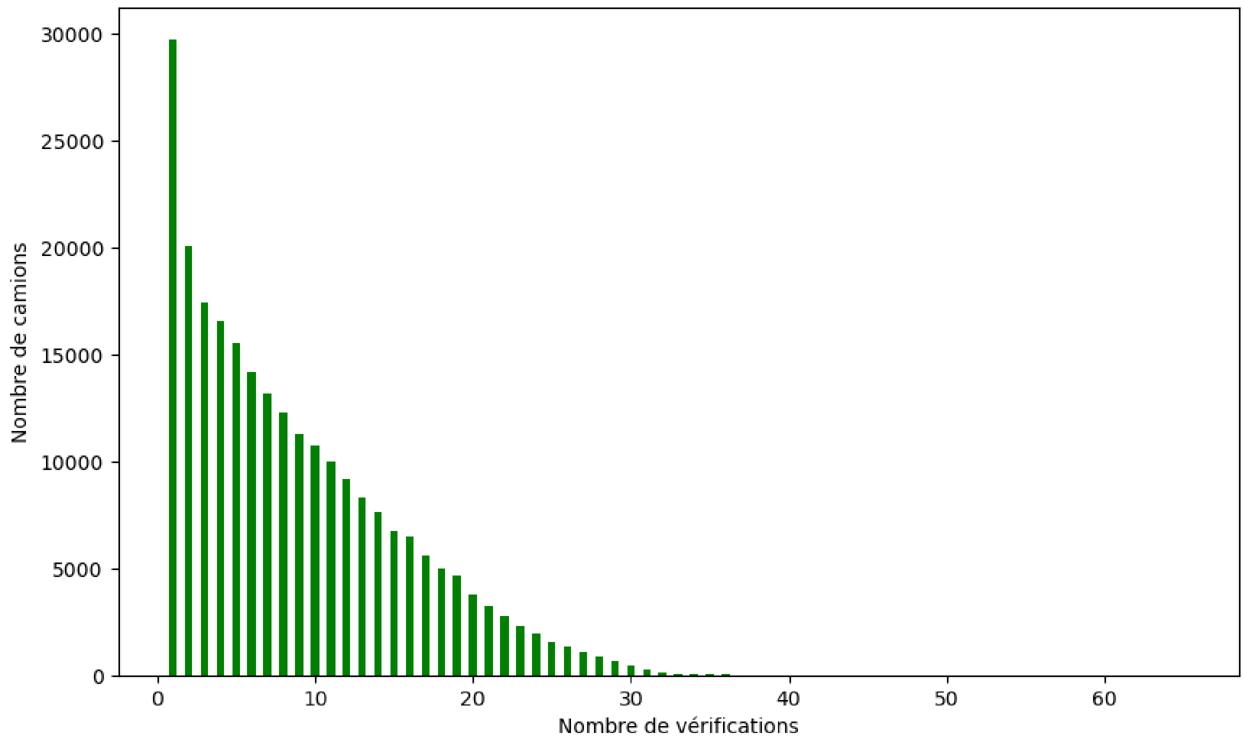

FIGURE 19 – Distribution du nombre de points kilométriques par camion (vérifications)

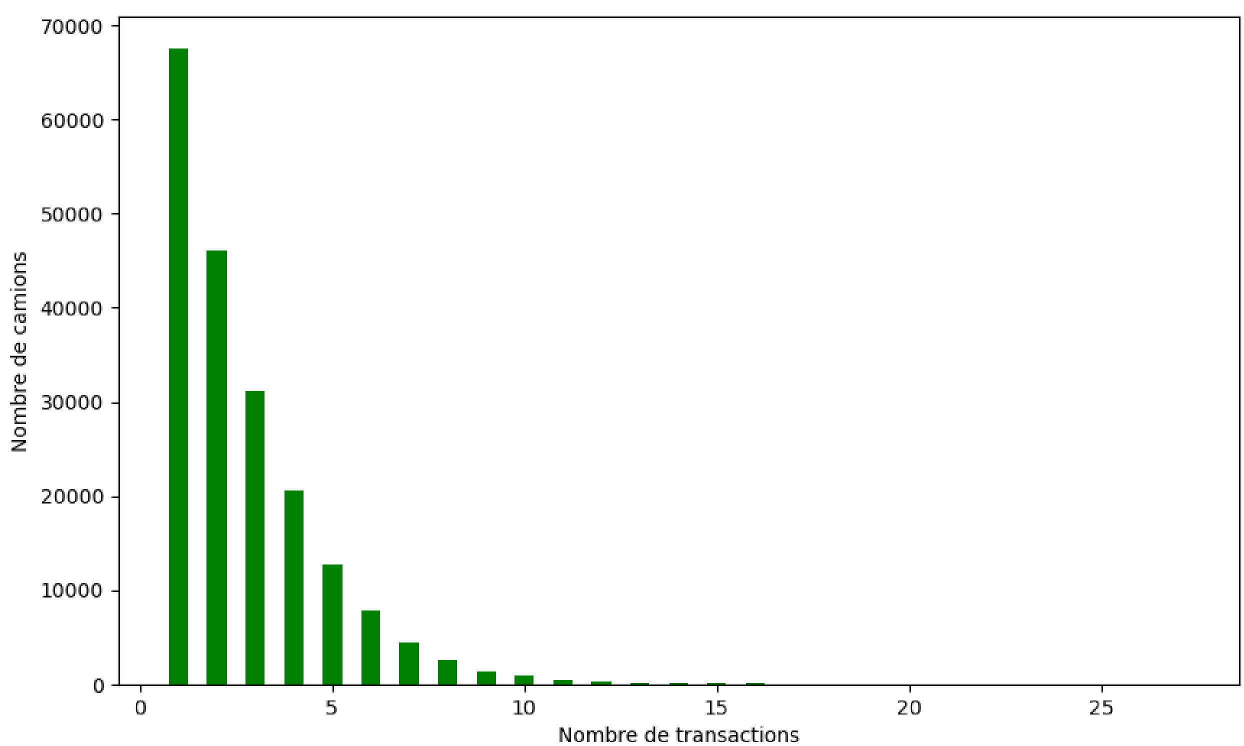

FIGURE 20 – Distribution du nombre de points kilométriques par camion (transactions)

8.2 Modèles d'estimation

Modèle	Cas	Expression de y
1	Tout t	$y = \frac{\text{Cov}(y_i,t_i)}{\mathbb{V}(t_i)} \cdot t + \left(\bar{y_i} - \frac{\text{Cov}(y_i,t_i)}{\mathbb{V}(t_i)} \cdot \bar{t_i}\right)$
2	$t \leq t_{i,1}$	$y = \frac{y_{i,2}-y_{i,1}}{t_{i,2}-t_{i,1}} \cdot t + \left(y_{i,1} - \frac{y_{i,2}-y_{i,1}}{t_{i,2}-t_{i,1}} \cdot t_{i,1}\right)$
	$t \in [t_{i,j}, t_{i,j+1}]$	$y = \frac{y_{i,j+1}-y_{i,j}}{t_{i,j+1}-t_{i,j}} \cdot t + \left(y_{i,j} - \frac{y_{i,j+1}-y_{i,j}}{t_{i,j+1}-t_{i,j}} \cdot t_{i,j}\right)$
	$t \geq t_{i,m_i}$	$y = \frac{y_{i,m_i}-y_{i,m_i-1}}{t_{i,m_i}-t_{i,m_i-1}} \cdot t + \left(y_{i,m_i-1} - \frac{y_{i,m_i}-y_{i,m_i-1}}{t_{i,m_i}-t_{i,m_i-1}} \cdot t_{i,m_i-1}\right)$
3	$t \notin [t_{i,1}, t_{i,m_i}]$	$y = \frac{\text{Cov}(y_i,t_i)}{\mathbb{V}(t_i)} \cdot t + \left(\bar{y_i} - \frac{\text{Cov}(y_i,t_i)}{\mathbb{V}(t_i)} \cdot \bar{t_i}\right)$
	$t \in [t_{i,j}, t_{i,j+1}]$	$y = \frac{y_{i,j+1}-y_{i,j}}{t_{i,j+1}-t_{i,j}} \cdot t + \left(y_{i,j} - \frac{y_{i,j+1}-y_{i,j}}{t_{i,j+1}-t_{i,j}} \cdot t_{i,j}\right)$
4	$t \leq t_{i,1}$	$y = \frac{y_{i,2}-y_{i,1}}{t_{i,2}-t_{i,1}} \cdot t + \left(y_{i,1} - \frac{y_{i,2}-y_{i,1}}{t_{i,2}-t_{i,1}} \cdot t_{i,1}\right)$
	$t \in [t_{i,1}, t_{i,m_i}]$	$y = \text{PCHIP}(t, \{t_i\}, \{y_i\})$
	$t \geq t_{i,m_i}$	$y = \frac{y_{i,m_i}-y_{i,m_i-1}}{t_{i,m_i}-t_{i,m_i-1}} \cdot t + \left(y_{i,m_i-1} - \frac{y_{i,m_i}-y_{i,m_i-1}}{t_{i,m_i}-t_{i,m_i-1}} \cdot t_{i,m_i-1}\right)$

TABLE 13 – Résumé des quatre modèles de régression pour estimer le kilométrage y en fonction du temps t.

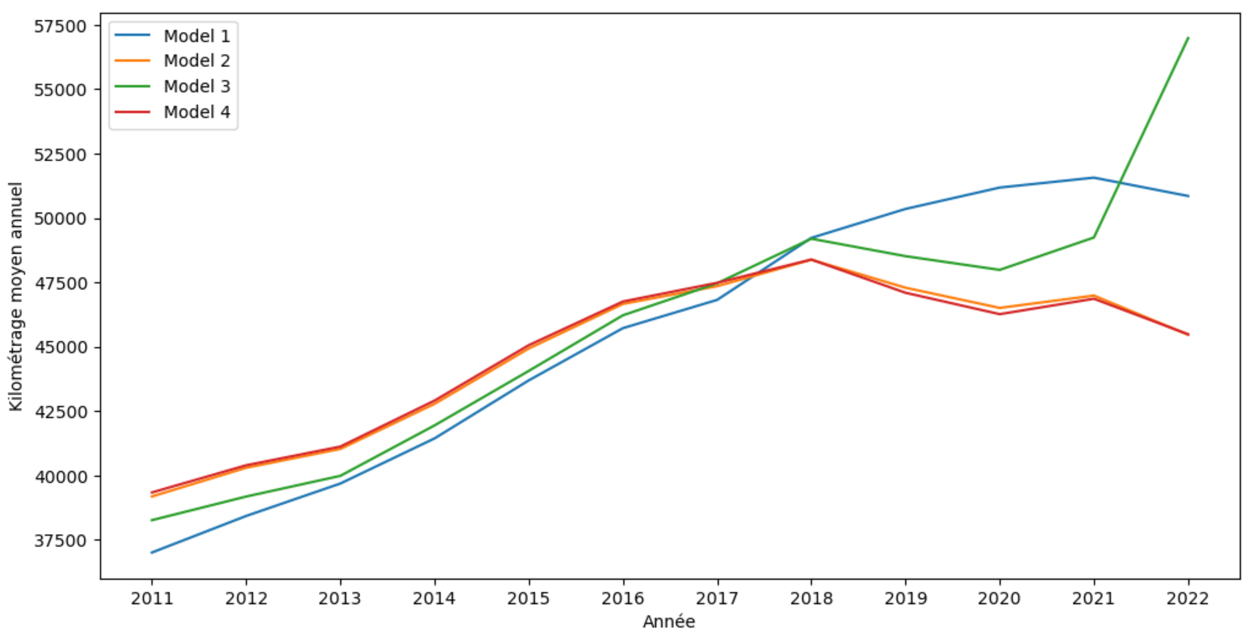

FIGURE 21 – Evolution des kilométrages annuels moyens par méthode d'estimation

8.3 Machine Learning

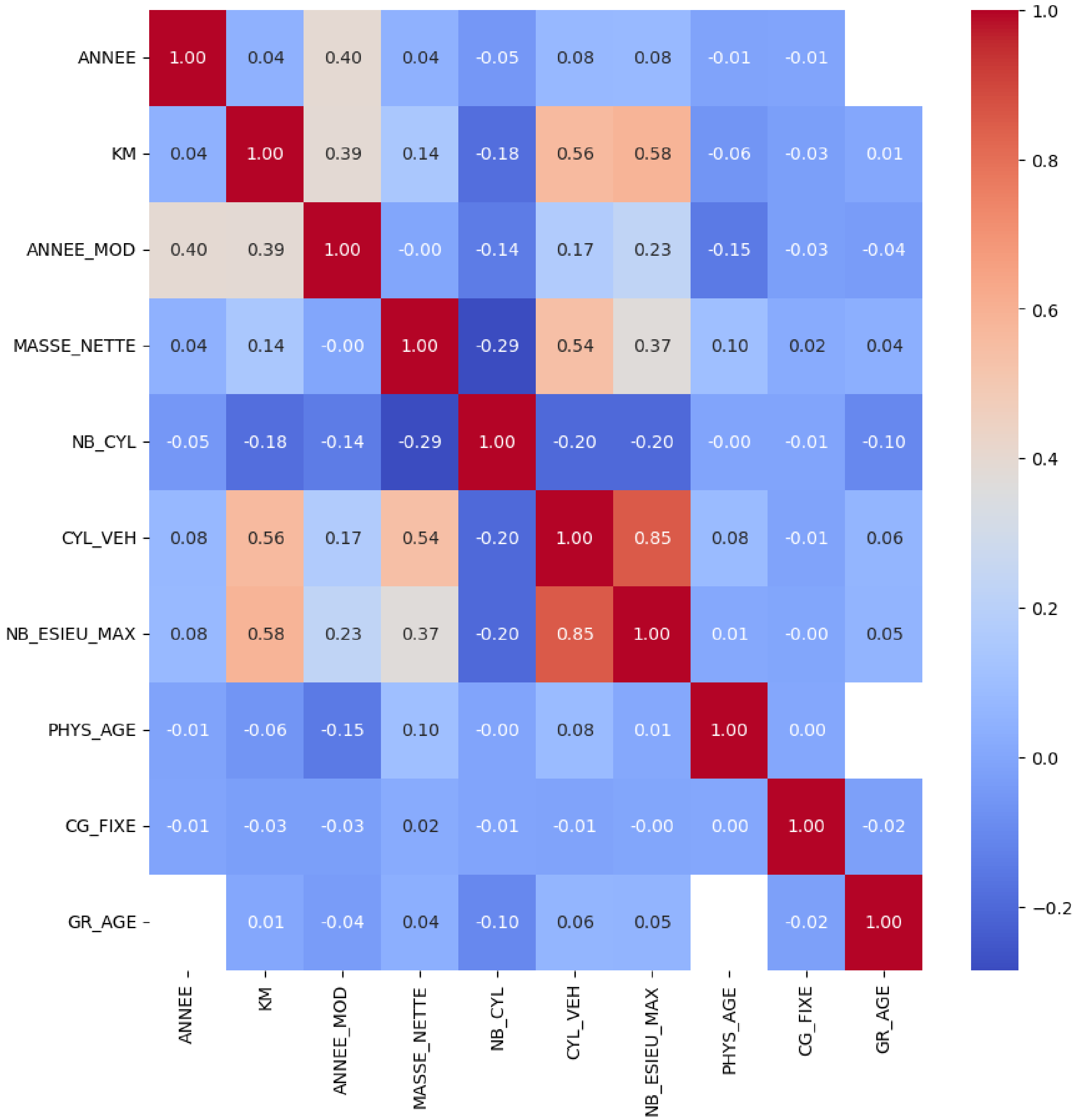

FIGURE 22 – Matrice de corrélation

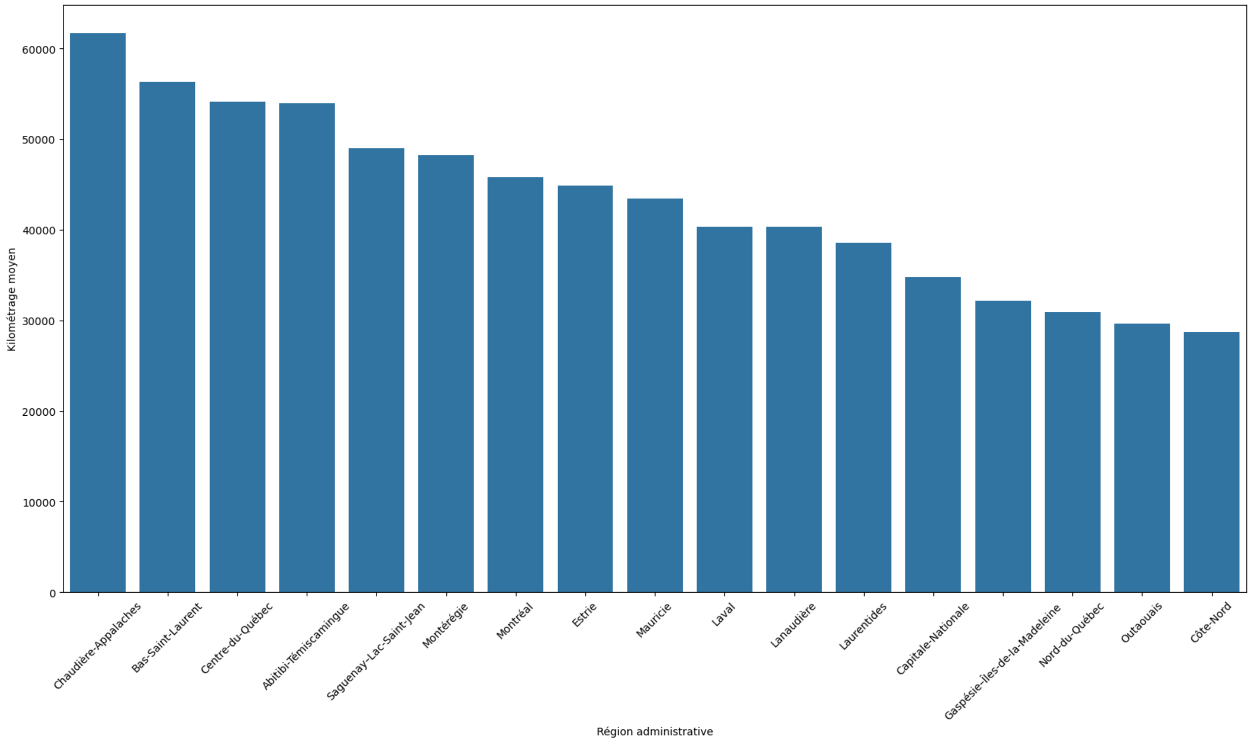

Figure 23 – Distance annuelle moyenne par région administrative

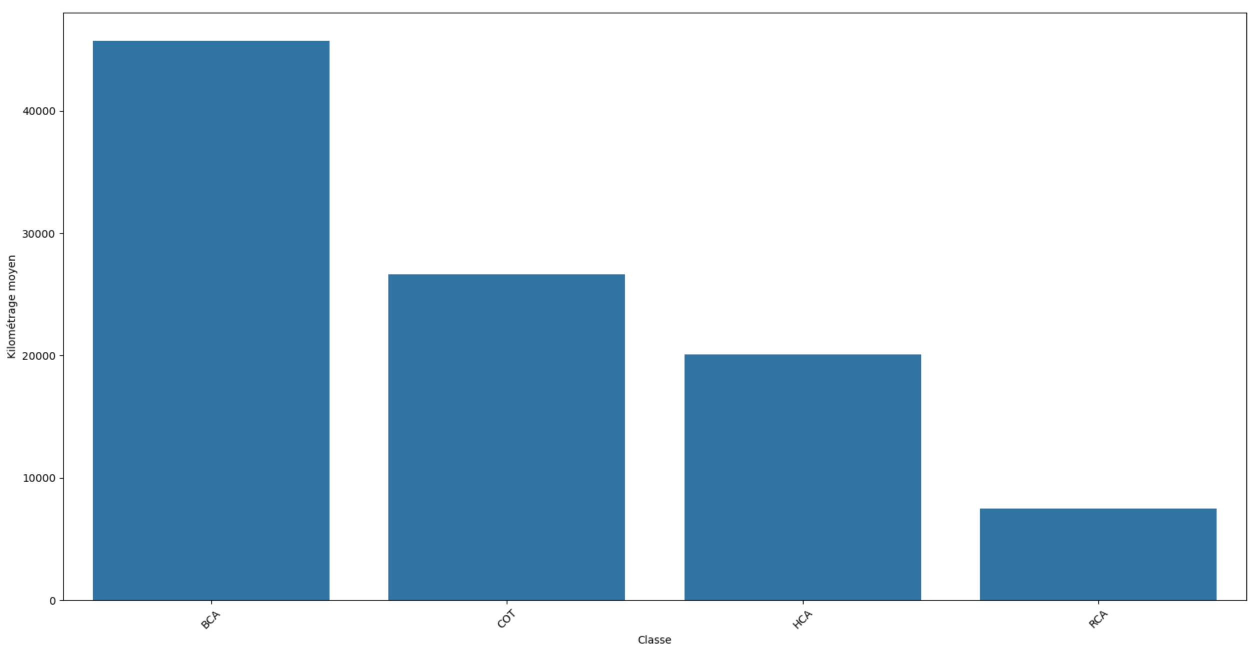

Figure 24 – Distance annuelle moyenne par classe de véhicules

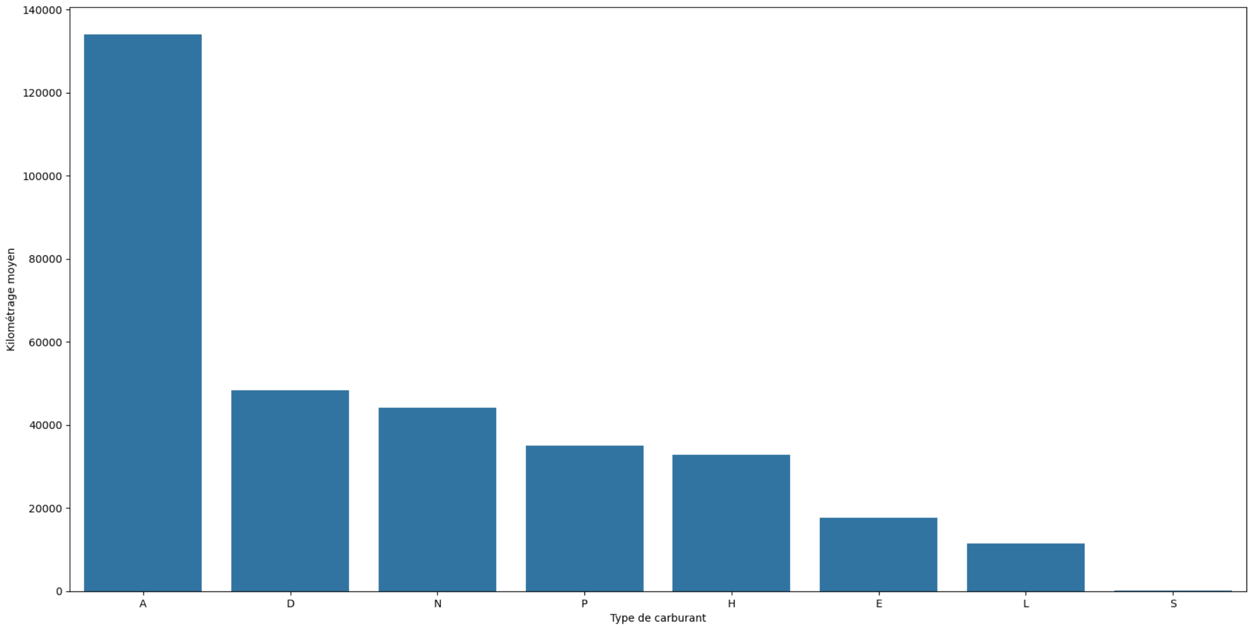

FIGURE 25 – Distance annuelle moyenne par type de carburant

```
def objective(n_estimators, min_samples_leaf, min_samples_split, max_depth):
    # Définition du modèle
    model = RandomForestRegressor(
        n_estimators=int(n_estimators),
        max_depth=int(max_depth),
        min_samples_leaf=int(min_samples_leaf),
        min_samples_split=int(min_samples_split),
        max_features='sqrt',
        n_jobs=-1,
        random_state=42
    )

    # Entrainement et prédiction
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_pred_train = model.predict(X_train)

    # Calcul des scores
    test_score = r2_score(y_test, y_pred)
    train_score = r2_score(y_train, y_pred_train)

    # Pénalisation pour limiter l'overfitting
    penalty = abs(test_score - train_score)

    # Retourner le score avec pénalité pour l'overfitting
    return test_score - penalty
```

8.4 Résultats finaux

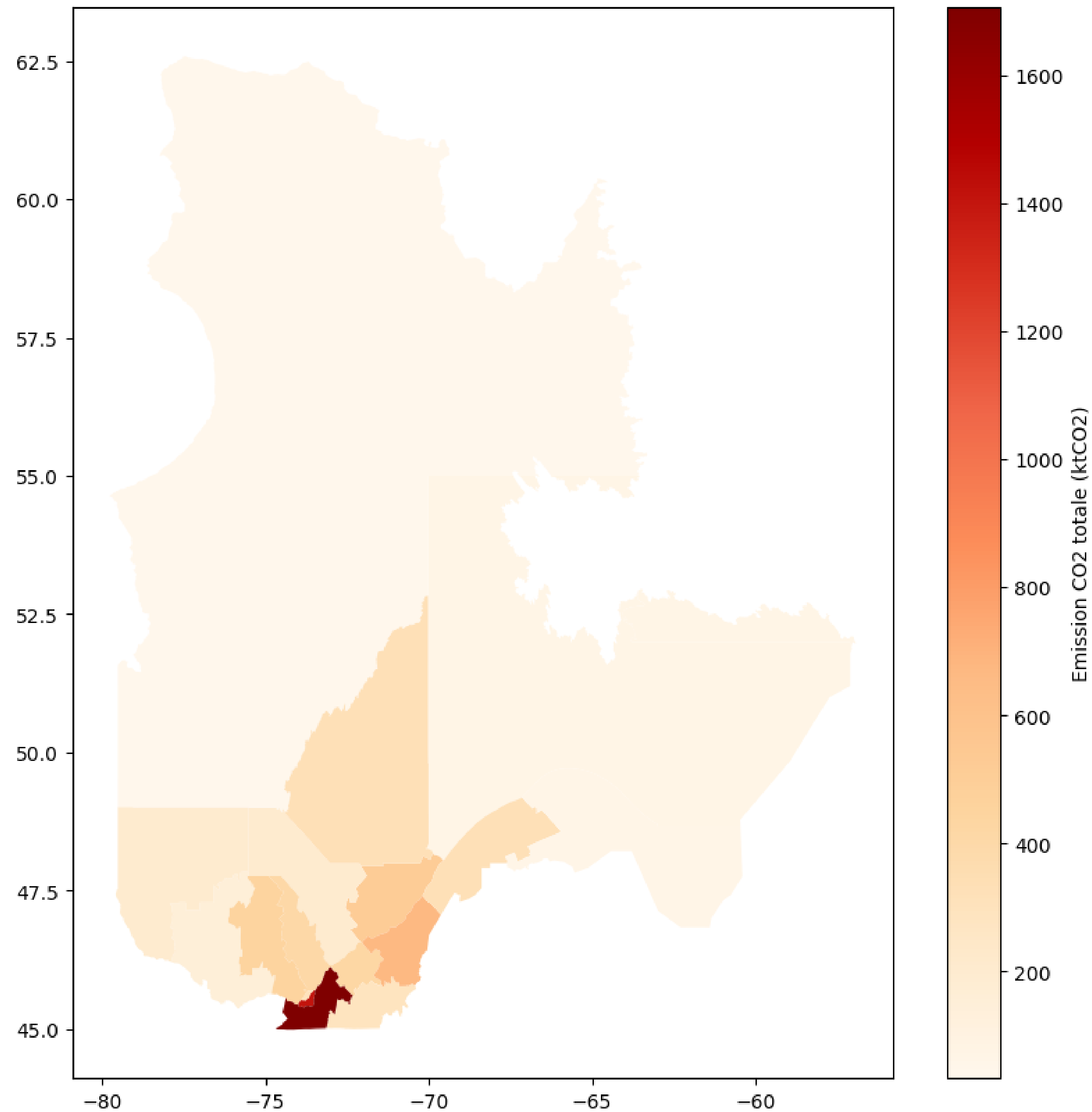

FIGURE 26 – Émissions de CO2 annuelle moyenne par région (en ktCO2)

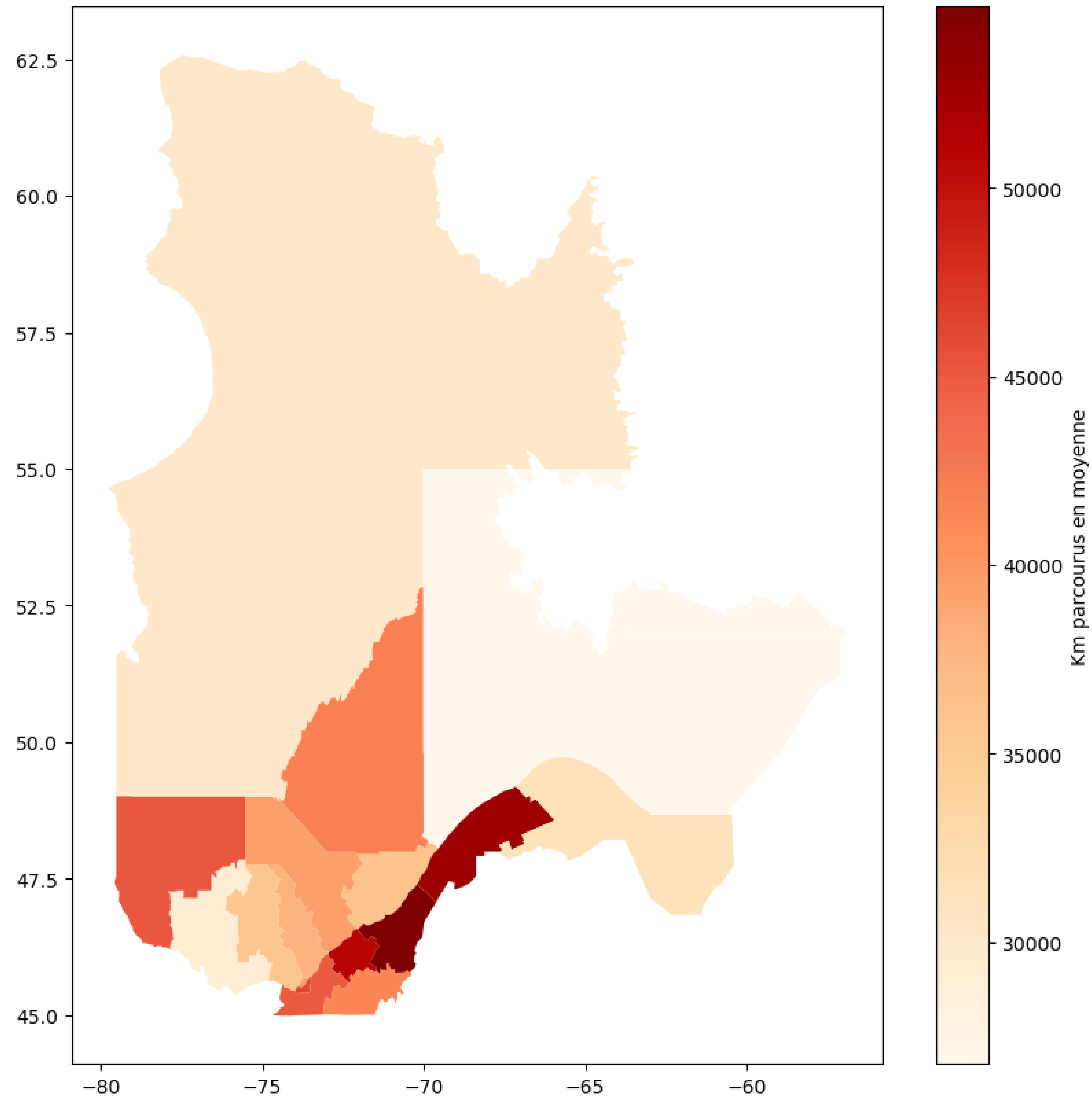

FIGURE 27 – Kilométrage moyen par région administrative

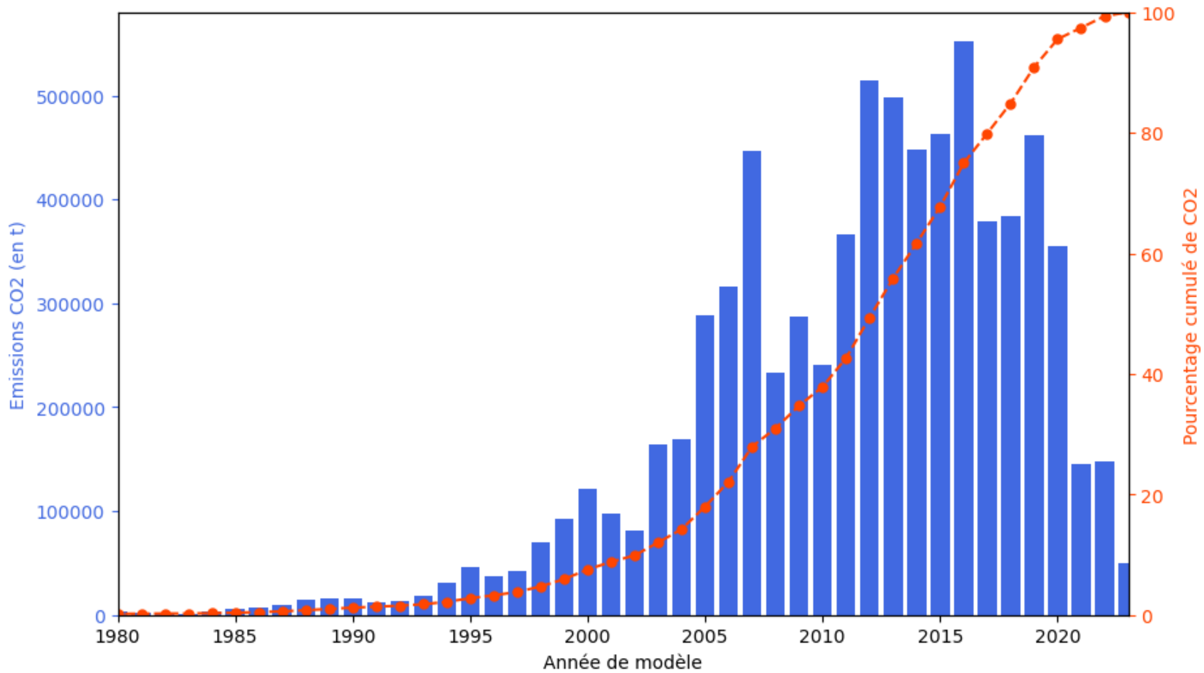

FIGURE 28 – Répartition des émissions de CO2 par année de modèle des véhicules

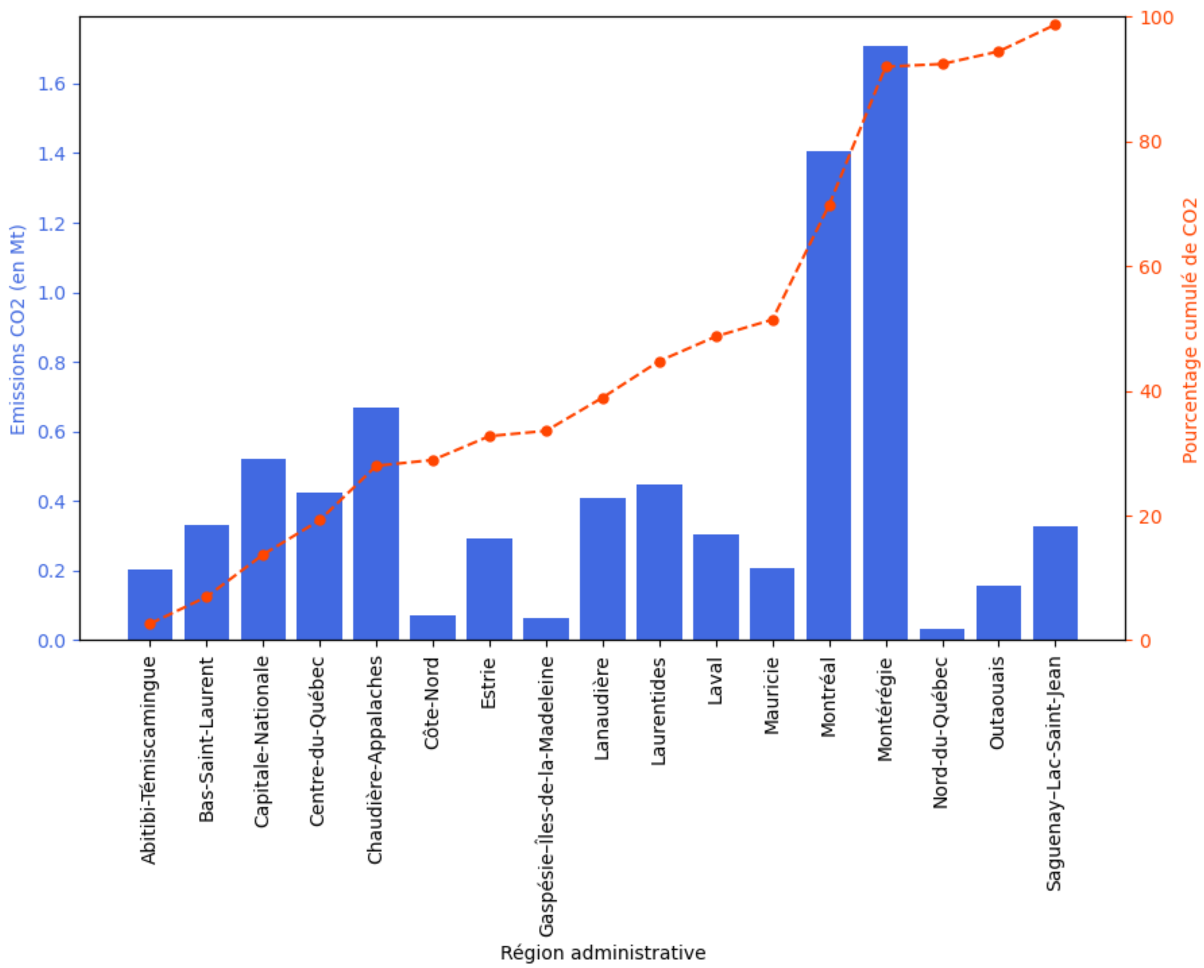

FIGURE 29 – Répartition des émissions de CO2 par région administrative

Références

[1] *Inventaire québécois des émissions de gaz à effet de serre en 2021 et leur évolution depuis 1990*. Ministère de l'environnement et de la lutte contre les changements climatiques, 2021.

[2] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 2019.

[3] Ministère de la Transition Écologique et Solidaire. Les émissions de gaz à effet de serre du secteur des transports, 2021.

[4] International Energy Agency. Energy end-uses and efficiency indicators data explorer, 2023. IEA, Paris. Available at : https://www.iea.org/data-and-statistics/data-tools/energy-end-uses-and-efficiency-indicators-data-explorer.

[5] Ministère des Transports du Québec. *Le poids lourd du Québec*. 2010.

[6] Ministère de l'Environnement et de la Lutte contre les changements climatiques. Déclaration des émissions de contaminants dans l'atmosphère provenant des carburants et combustibles.

[7] Ministère de l'Environnement et de la Lutte contre les changements climatiques. Guide de quantification des émissions de gaz à effet de serre. Guide technique, 2015.

[8] Environnement et Changement climatique Canada. Rapport d'inventaire national 1990–2021 : Sources et puits de gaz à effet de serre au canada. Technical report, Environnement et Changement climatique Canada, 2021. canada.ca/inventaire-ges, cf partie A3.1.4.2.

[9] Mobilité durable et Électrification des transports Gouvernement du Québec : Transports. Guide des normes de charges et dimensions des véhicules routiers, 2013.