



# **RAPPORT FINAL PROJET SCIENTIFIQUE COLLECTIF ECO 05**

**Optimisation du malus écologique**

2023-2024

---

Maria Baz - Capucine Burel - Antoine Maechler - Etienne Marécaux - Eva Zribi



INSTITUT  
POLYTECHNIQUE  
DE PARIS

## TABLE DES MATIÈRES

<b>Remerciements</b>	<b>3</b>
<b>1 Introduction et littérature</b>	<b>4</b>
1.1 Le malus écologique et ses défauts . . . . .	4
1.2 Littérature . . . . .	4
1.3 Problématique . . . . .	5
1.4 Objectifs et cheminements du PSC . . . . .	5
<b>2 Présentation des données et validation du sujet</b>	<b>6</b>
2.1 Présentation des bases utilisées . . . . .	6
2.2 Traitement des données et détails statistiques . . . . .	7
2.3 Des émissions de CO2 imparfaitement estimées lors de la vente . . . . .	8
<b>3 Estimation des émissions de CO2 : Modèle et résultats</b>	<b>11</b>
3.1 Le machine learning au service de la mesure des émissions . . . . .	11
3.2 Sélection d'un modèle . . . . .	12
3.3 Approche simple : la régression linéaire . . . . .	13
3.4 Les améliorations de la régularisation Lasso . . . . .	15
3.5 Les améliorations du modèle KNN . . . . .	18
<b>4 Effets de cette nouvelle estimation sur la demande de véhicules et sur les émissions résultantes</b>	<b>20</b>
4.1 Modèle économique . . . . .	20
4.2 Mise en évidence expérimentale des améliorations permises par un calcul plus juste des émissions	23
<b>5 Réduire encore les émissions par la création d'une grille de taxation optimisée</b>	<b>29</b>
5.1 Approche générale d'optimisation du montant de la taxe . . . . .	29
5.2 Présentation et discussion autour de la construction de taxations alternatives . . . . .	31
<b>6 Conclusion</b>	<b>35</b>
6.1 Bilan de notre travail - retour sur nos objectifs . . . . .	35
6.2 Prolongements . . . . .	36
6.3 Organisation du travail . . . . .	36
<b>7 Appendice</b>	<b>38</b>

## REMERCIEMENTS

Nous tenons en premier lieu à remercier Monsieur Clément Malgouyres, enseignant-chercheur au Centre de Recherche en Economie et Statistique (CREST), qui nous a apporté une aide précieuse tout au long du projet. Il nous a encadré et fourni une importante documentation ainsi que les bases de données sur lesquelles nous avons fondé notre travail. Ses remarques pertinentes nous ont permis de cerner les enjeux de notre sujet, fixé nos objectifs et contribué à l'avancement de notre projet. Il a toujours été à nos côtés pour nous conseiller dans nos démarches afin de parvenir à des résultats satisfaisants.

Nous tenons également à remercier Monsieur Julien Combe, coordinateur de l'ensemble des PSC d'économie et enseignant-chercheur au département d'économie de l'École polytechnique, qui nous a accompagné durant l'intégralité de ce projet. Son support nous a permis de surmonter plusieurs obstacles et d'être toujours sur la bonne voie.

Enfin, nous remercions l'Ecole polytechnique de nous avoir offert l'opportunité de travailler sur un tel projet instructif et bénéfique, tant au niveau personnel qu'au niveau professionnel.

# 1

## INTRODUCTION ET LITTÉRATURE

### 1.1 LE MALUS ÉCOLOGIQUE ET SES DÉFAUTS

Entré en vigueur en France le 1er janvier 2008, le malus écologique sur la vente de véhicules neufs a eu un impact certain mais encore imparfait. S'il est indéniable que cette politique a eu un effet positif sur les ventes de véhicules neufs, elle comporte toujours plusieurs biais, en particulier sur l'évaluation du CO<sub>2</sub> émis par les voitures, ainsi que les niveaux de taxation proposés par l'État en fonction du CO<sub>2</sub> émis.

Actuellement, le malus écologique est défini à partir d'une évaluation protocolaire des émissions de CO<sub>2</sub> qu'engendre la circulation du véhicule étudié. L'ensemble des tests conduits est mené avant la mise en vente du véhicule et vise à évaluer le niveau d'émission du produit par des mises en condition diversifiées. Le protocole suivi définit alors une norme CO<sub>2</sub>. Depuis 2020, la norme en vigueur est le protocole Worldwide harmonized Light vehicle Test Procedure (WLTP), dont l'objectif était d'améliorer le réalisme des mesures en effectuant des tests en laboratoire et sur route. Bien que cette norme ait été reconnue plus fiable que le protocole précédent (le new European Driving Cycle, institué en 1990), la norme WLTP demeure imparfaite. Les émissions de CO<sub>2</sub> qu'elle prédit, appelées "CO<sub>2</sub> théorique" dans la suite de notre rapport, sont une bonne première approximation qu'il serait intéressant d'optimiser pour accroître la justesse du malus écologique en vigueur.

### 1.2 LITTÉRATURE

Tout au long de notre projet, une riche littérature nous a permis de saisir tous les enjeux liés à notre problématique, de construire pas à pas les modèles employés pour y répondre, et de mesurer, en s'y référant, les progrès apportés par nos modélisations et résultats.

Aujourd'hui, le secteur des transports s'avère être le plus grand contributeur aux émissions de gaz à effet de serre en France (31%). La Stratégie Nationale à Faible Teneur en Carbone (MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE ET SOLIDAIRE, 2015) prévoit une France neutre en carbone en 2050. En 2007, l'État français a adopté une mesure forte en appliquant à l'achat de véhicules neufs une taxe en fonction de leurs émissions : bonus sur les véhicules à basses émissions de carbone, et malus graduel sur les véhicules à hautes émissions.

Cette politique a ainsi contribué de façon remarquable à la baisse des émissions à court terme, notamment en France, Norvège, Royaume-Uni et aux Pays-Bas. Sur l'exemple de la Norvège, les effets concluants de l'implantation du malus en 2007 sont détaillés dans l'article *Greening the vehicle fleet : Norway's CO<sub>2</sub>-differentiated registration tax* (YAN et GUNNAR S., 2018).

Une première évaluation de la politique de bonus/malus par I. Durrmeyer, A. Guillouzouic, C. Malgouyres, T. Mayer, M. Tô, 2023, *Évaluation des mesures de soutien aux véhicules propres* (DURRMEYER et al., 2024) fut une belle entrée en matière à notre projet. En analysant des scénarios contrefactuels, l'étude manipule les concepts-clés de modélisation qui nous ont été fort utiles. En particulier, en s'appuyant sur une modélisation logit ou nested logit pour l'effet de la politique sur la demande, le choix de notre propre modèle de demande n'en a été que plus évident. La lecture du livre de Kenneth E. Train, *Discrete Choice Methods with Simulation* (TRAIN, 2009) a complété et approfondi nos connaissances et notre maîtrise d'une modélisation prix-dépendante de la demande.

Cependant, ce mécanisme de taxation n'est pas parfait. De nombreux scientifiques ont dénoncé le fait que cette politique ne suffirait pas à atteindre les objectifs de la France en termes d'émissions pour le transport, comme dans le rapport de KESSLER et al., 2023, qui conclue que des mesures supplémentaires doivent être prises dans le secteur des transports. Nous prenons bien en considération le fait que l'augmentation du prix, parfois conséquente, de certains véhicules, diminue le surplus du consommateur à l'achat d'une voiture neuve. Ainsi, l'un des objectifs de ce rapport sera également de construire une nouvelle politique plus réductrice d'émissions et maximisatrice du surplus du consommateur.

Et ce n'est pas le seul défaut de cette politique automobile. Une étude par D'HAULTFOEUILLE et al., 2014 met en exergue que la politique de bonus/malus, dans ses premières années de mise en place, a en réalité augmenté les émissions liées aux voitures en France ! Le mécanisme à l'oeuvre est assez simple : trop permissive et même incitative (bonus à l'achat), les ventes de véhicules ont connu une forte augmentation après l'introduction de la mesure. Les véhicules plébiscités sont certes les moins émetteurs, mais leurs émissions cumulées, au regard de leur quantité, ont dépassé les émissions de l'avant bonus/malus.

Cet effet prend d'autant plus de sens à la lecture de l'étude *From laboratory to road, a 2018 update of official and "real-world" fuel consumption and CO2 values for passenger cars in Europe* (UWE et al., 2019). En effet, le rapport, très complet, montre que les émissions utilisées actuellement ne sont pas représentatives de l'effet réel des véhicules sur l'environnement et la pollution de l'air (voir 1.1 pour les explications). Ce constat alarmant a été l'élément déclencheur de notre volonté de recalculer les émissions de CO2 de manière plus juste. Il nous est alors apparu pertinent de redéfinir le calcul des émissions réelles à travers de nombreux paramètres des modèles.

### 1.3 PROBLÉMATIQUE

La littérature révèle que la norme d'évaluation actuelle des émissions de CO2 des véhicules est imparfaite, voire dans certains cas éloignée de la réalité. Partant de ce constat, nous nous attacherons à répondre aux questions suivantes : Quels critères ou paramètres prendre en compte pour aboutir à un malus écologique fondé sur une évaluation des émissions plus juste ? La prise en compte de paramètres tangibles et constitutifs du véhicule est-elle pertinente ? Dans quelle mesure ces normes pourraient-elles permettre d'optimiser la taxe carbone ?

En somme, comment optimiser la justesse de la taxation pour mieux contrôler les émissions ?

### 1.4 OBJECTIFS ET CHEMINEMENTS DU PSC

L'objectif de notre PSC consiste à modéliser un nouveau malus écologique, qui prenne mieux en compte les émissions réelles, qui soit calculable avec fiabilité dès la mise en vente du véhicule, et qui permette de moduler la demande en fonction d'un seuil d'émissions choisi par le gouvernement. Pour ce faire, nous avons travaillé en plusieurs étapes :

1. Nous définissons d'abord une nouvelle fonction de calcul des émissions de CO2 de chaque véhicule, dans le but de fonder notre malus écologique sur des paramètres plus proches des émissions réelles. Nous utilisons alors des bases de données regroupant les caractéristiques de nombreux véhicules vendus au cours de l'année 2021 puis par régression et machine learning nous construisons une fonction de calcul des émissions. La prise en compte des paramètres du véhicule couplée aux résultats obtenus en laboratoire permet d'affiner les estimations pour que celles-ci s'approchent des émissions réelles, connues seulement plusieurs mois après la mise en vente.
2. En construisant un modèle de demande, nous étudions ensuite l'effet de notre nouveau calcul des émissions sur les parts de marchés de chaque véhicule : confrontés à des prix différents résultant de la réévaluation

des émissions de chaque véhicules, les consommateurs vont en effet orienter leurs choix différemment. Nous mesurerons ainsi l'effet de notre calcul initial des émissions sur les émissions moyennes observées qui seront la conséquences de nouvelles tendances de ventes.

3. Nous travaillons enfin à la définition d'une nouvelle fonction de malus, en résolvant un problème d'optimisation sous contrainte. Il s'agit de définir le malus ajouté au prix de chaque véhicule, qui conduise à une redistribution des parts de marché permettant de garantir un niveau moyen de CO<sub>2</sub> émis fixé par le gouvernement.

## 2

# PRÉSENTATION DES DONNÉES ET VALIDATION DU SUJET

## 2.1 PRÉSENTATION DES BASES UTILISÉES

Pour répondre à notre problématique, nous conduisons un travail d'analyse de données, permettant de dégager de premières conclusions et de construire un modèle optimisé. Nous nous sommes appuyés sur deux bases de données complémentaires, fournies par notre tuteur et rassemblant des informations sur un grand nombre de véhicules vendus entre 2005 et 2021. Il s'agit du Répertoire statistique des véhicules routiers (RSVeRo) et d'une base fournie par l'ADEME.

### 2.1.1 • RÉPERTOIRE STATISTIQUE DES VÉHICULES ROUTIERS (BASE RSVeRo)

La base Rsvero contient deux fichiers qui recensent les voitures vendues entre 2005 et 2021 en France. 84 000 véhicules y sont répertoriés. La base RSVeRo a été produite à partir de la base de données Sprit, d'abord éditée lors de la publication du rapport *From Laboratory to Road* UWE et al., 2019 et obtenue par scraping du site [spritmonitor.de](https://www.spritmonitor.de). Cette base RSVeRo rassemble :

- les données d'identification du véhicule. Il s'agit de sa marque, son modèle, son année de mise en circulation et son Code National d'Identification du Type (CNIT). L'identifiant CNIT est un code alphanumérique composé de 12 à 15 caractères, synthétisant les caractéristiques du véhicule, et garantissant l'identité des véhicules d'une base de donnée à l'autre.
- des éléments relatifs aux émissions du véhicule. La base donne ainsi pour chaque modèle, son niveau d'émission théorique de CO<sub>2</sub>, obtenu à l'issu des tests initiaux, ainsi que les émissions réelles obtenues quelques mois après la mise en circulation.
- des caractéristiques techniques, telles que la catégorie énergétique du véhicule, sa puissance, son poids, sa carrosserie, ou encore son nombre de cylindrées.

Un tableau est fourni en appendice, lisant exhaustivement les grandeurs données par la base (table 8).

Cette base ne précise cependant ni le prix des véhicules, ni le volume des ventes. D'où la nécessité de compléter ces données par une seconde base.

### 2.1.2 • BASE FOURNIE PAR L'ADEME

La seconde base de données que nous avons utilisé a été éditée par l'Agence De l'Environnement et de la Maîtrise de l'Energie (ADEME), disponible sur le site internet de l'Agence de la transition écologique. Cette base répertorie les véhicules commercialisés en France en 2021. Chaque modèle présenté est accompagné de données concernant ses caractéristiques techniques. Elle complète la base RSVeRo car elle précise la présence ou non d'un bonus/malus, le prix de vente, ainsi que des indices de consommation. Un tableau est fourni en appendice, listant exhaustivement les grandeurs données par la base (table 8).

## 2.2 TRAITEMENT DES DONNÉES ET DÉTAILS STATISTIQUES

Nous avons traité ces deux bases pour produire une base de données de travail, rassemblant des informations complémentaires. Il s'agit de manipuler ces données afin de les rendre exploitables pour la suite de notre travail. En particulier :

1. On fusionne ces bases de données en utilisant l'identifiant CNIT.
2. On ne retient que les véhicules vendus à plus de 20 exemplaires, les autres correspondant souvent à des prototypes avec des caractéristiques extrêmes.
3. On retire les paramètres qu'on n'utilisera pas, ainsi que les véhicules électriques, non concernés par le malus et avec des émissions de CO<sub>2</sub> toujours nulles quels que soit les paramètres.
4. On supprime les véhicules dont les données ne sont pas ou mal remplies.
5. Afin de pouvoir intégrer les variables qualitatives de notre base de données à nos modèles de prédictions de CO<sub>2</sub>, on les transforme en variables quantitatives binaires. Intéressons nous par exemple à la variable "énergie" qui prend des valeurs telles que 'essence' ou 'électrique'. Afin de pouvoir l'intégrer à notre calcul, on crée une variable pour chaque catégorie d'énergie (une variable 'essence' par exemple) auxquelles on associe le nombre 0 si le véhicule ne fait pas partie de cette catégorie ou 1 si le véhicule y appartient. Nous avons effectué ce même changement sur la variable "Carrosserie" qui indiquait la catégorie du véhicule.

Finalement on aboutit à une base de données de 39755 véhicules, sur 54529 à l'origine. La base comporte 27 paramètres.

La table 1 présente les statistiques de notre base de donnée de travail. Nous affichons les statistiques permettant de mieux appréhender les ordres de grandeur sur lesquels nous travaillons. Ce tableau permet en outre de visualiser les grandeurs chiffrées que détaille notre base de travail.

	Moyenne	Écart-type	Minimum	Maximum
CO <sub>2</sub> réel (g/km)	141.42	26.5	0.0	405.89
CO <sub>2</sub> théorique (g/km)	119.15	30.70	0.0	373.0
Cylindrée	1404.3	356.3	874.1	5000.0
Poids total roulant autorisé (kg)	3062.1	687.4	0.0	6850.0
Charge utile (kg)	551.27	73.1	224.07	1243.0
Puissance (en KW)	90.68	28.9	26.3	478.0
Nombre de décibels	74.1	4.1	56.0	96.0
Prix (euros)	30147.6	12984.2	8890.0	205715

TABLE 1 – Statistiques descriptives de notre base de données de travail.

## 2.3 DES ÉMISSIONS DE CO<sub>2</sub> IMPARFAITEMENT ESTIMÉES LORS DE LA VENTE

Le tableau de statistiques descriptives de notre base de travail révèle bien le problème des écarts entre émissions réelles et émissions théoriques. En effet, alors que les émissions moyennes de CO<sub>2</sub> réel, pondérées par le nombre de ventes de chaque véhicule, s'établissent à 141.2 g/km, nous notons que la moyenne des émissions théoriques s'élève seulement à 119.15 g/km.

Traçons la distribution de l'erreur afin de confirmer ce résultat et de mieux l'illustrer.

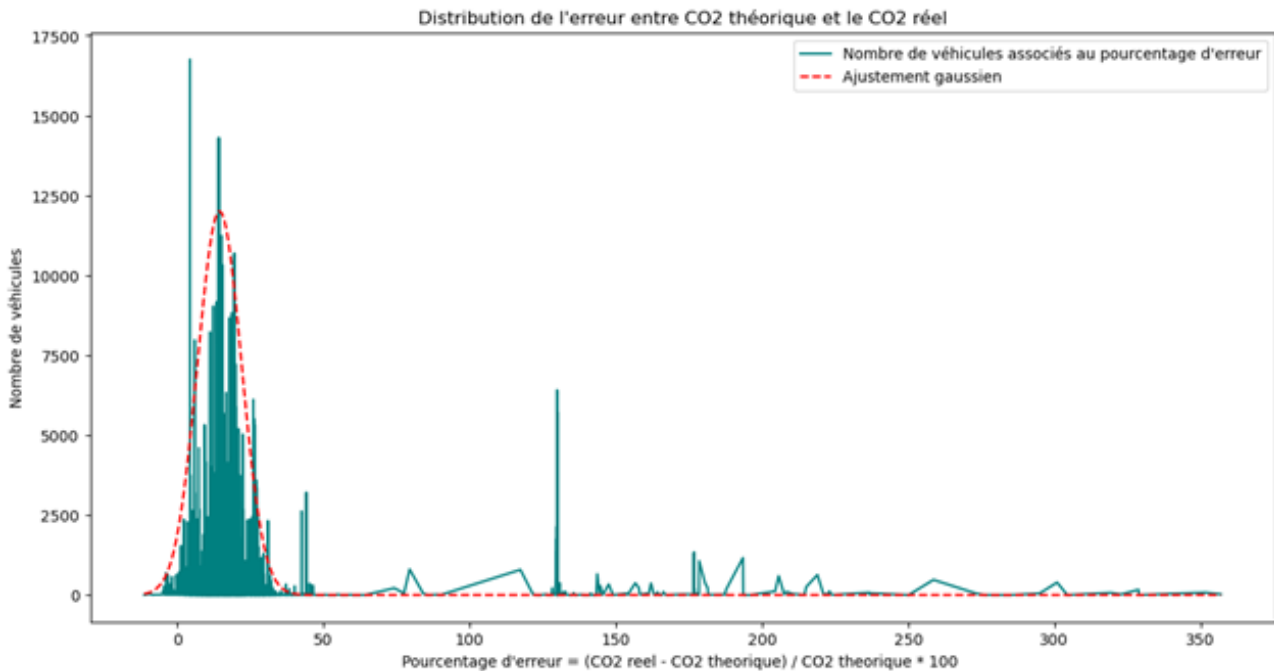


FIGURE 1 – Distribution de l'erreur entre les émissions réelles de CO<sub>2</sub> et les émissions théoriques

Ce tracé, en illustrant l'erreur dans le calcul du CO<sub>2</sub> théorique relativement au CO<sub>2</sub> réel, révèle bien que les estimations sont globalement faussées. La majorité des véhicules se trouve dans une enveloppe allant de 10 à 50% d'erreur, tandis que l'on observe également un pic de ventes correspondant à des estimations pour laquelle l'erreur s'élève à 130%.

Ce constat confirme donc la pertinence de notre entreprise, qui consiste à produire une nouvelle fonction de calcul des émissions avant la mise en vente. Mettons donc en évidence par une première analyse crue de notre base, les liens entre émissions réelles et caractéristiques du véhicules. Ceci nous permettra par exemple de comprendre quels véhicules se trouvent dans le pic de ventes situé à un niveau d'erreur de 130%.

### 2.3.1 • OBSERVATION DES LIENS ENTRE ÉMISSIONS RÉELLES ET CARACTÉRISTIQUES DU VÉHICULE

Analysons le lien entre caractéristiques et émissions de CO<sub>2</sub>.

#### Corrélation entre variables quantitatives et émission de CO<sub>2</sub>

Afin de cerner les caractéristiques techniques quantitatives que nous pourrions par la suite utiliser pour affiner l'estimation du CO<sub>2</sub> produit par un véhicule, calculons le coefficient de corrélation de Pearson qui unit



statistiquement les paramètres des véhicules au CO2 réel.

Le coefficient de corrélation de Pearson  $r$  entre deux variables  $x$  et  $y$  est calculé selon la formule suivante :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Où :

- $x_i$  et  $y_i$  sont les valeurs des observations pour les variables  $x$  et  $y$  respectivement.
- $\bar{x}$  et  $\bar{y}$  sont les moyennes des échantillons pour les variables  $x$  et  $y$  respectivement.

Ce coefficient compris entre -1 et 1 peut s'interpréter de la manière suivante :

- Un coefficient proche de 1 indique une forte corrélation positive entre les variables.
- Un coefficient proche de -1 indique une forte corrélation négative entre les variables.
- Un coefficient proche de 0 indique une faible corrélation (ou pas de corrélation) entre les variables.

	Corrélation avec le CO2 réel
CO2 théorique (g/km)	0.88
Puissance (en kW)	0.36
Charge utile (kg)	0.30
Cylindrée	0.27
Poids total roulant autorisé (kg)	0.24
Poids total autorisé en charge (kg)	0.17

TABLE 2 – Corrélation entre différentes caractéristiques techniques des véhicules et CO2 réel mesuré.

Les résultats présentés dans ce tableau révèlent plusieurs points :

- Sans surprise, le CO2 théorique est dans nos données, fortement corrélé au CO2 réel. Le coefficient n'est cependant pas égal à 1, comme les valeurs statistiques l'annonçaient déjà.
- Le poids du véhicule, lié au poids total roulant autorisé et au poids total autorisé en charge, a une influence non négligeable sur le CO2 réel mesuré, ce qui ouvre la piste d'un calcul des émissions faisant intervenir le poids.
- La puissance également a une influence, comme l'indique le coefficient de 0.36 calculé.

Plus globalement, ce tableau nous confirme qu'il est possible d'enrichir le calcul du CO2 réel en s'appuyant sur d'autres paramètres techniques qui semblent corrélés. Ceci semble d'autant plus important que l'erreur commise sur l'estimation du CO2 est inégale suivant les caractéristiques du véhicule. Analysons donc le lien entre les émissions de CO2 et nos deux variables non quantitatives (type énergétique du véhicule et carrosserie).

### Energie et émission de CO2

Les variations d'erreurs commises en fonction du type d'énergie du véhicule sont non négligeables, comme l'indique la table 3. Celle-ci fournit les statistiques éclairant la répartition des types énergétiques dans notre base. On peut ainsi y lire que les véhicules hybrides, représentant 9.12% des véhicules vendus dans nos données, émettent bien plus que ce que les tests initiaux prédisent : l'erreur s'élève à 132%. Une explication peut être que ces véhicules, plus lourds que les voitures simples, sont bien souvent utilisées excessivement sur leur mode essence : elles consomment alors davantage qu'une voiture classique, fait que le test initial en laboratoire et sur route n'anticipe pas. Inversement, les émissions des véhicules roulant grâce à une bicarburation essence GPL sont plutôt bien évaluées, l'erreur ne montant qu'à 5.6%.

Énergie	Volume (en %)	CO2 théorique	CO2 réel	Erreur (en %)
Hybride essence électricité (R)	9.12	38.43	89.48	132.82
Bicarburant essence-GPL	3.39	115.54	122.02	5.60
Hybride essence-électricité (NR)	16.60	116.86	134.39	15.00
Électricité	0.12	0.00	0.00	0.00
Bicarburant essence-gaz naturel	0.00	106.18	94.19	-11.29
Essence	40.98	131.02	154.14	17.64
Superéthanol	1.69	125.32	143.87	14.80
Hybride superéthanol-électricité (NR)	0.44	147.30	169.71	15.22
Gazole-électricité hybride (NR)	2.37	146.56	164.80	12.45
Gazole-électricité hybride (R)	0.51	41.10	115.66	181.39
Gaz naturel	0.00	109.89	100.06	-8.94
Gazole ou diesel	24.78	129.90	145.18	11.76

TABLE 3 – Volume de ventes moyen (en %) de chaque type énergétique dans la base, émission moyenne de CO2 (en g/km) théorique et réelle, et pourcentage d'erreur moyen entre CO2 théorique et CO2 réel. On note NR les véhicules non rechargeables et R les véhicules rechargeables.

Afin de rendre ces résultats plus visibles, nous avons tracé le diagramme de la Figure 2. Les niveaux de CO2 réel et théorique y sont tracés en fonction du type énergétique. Il est en particulier notable que l'erreur est considérable sur les véhicules hybrides.

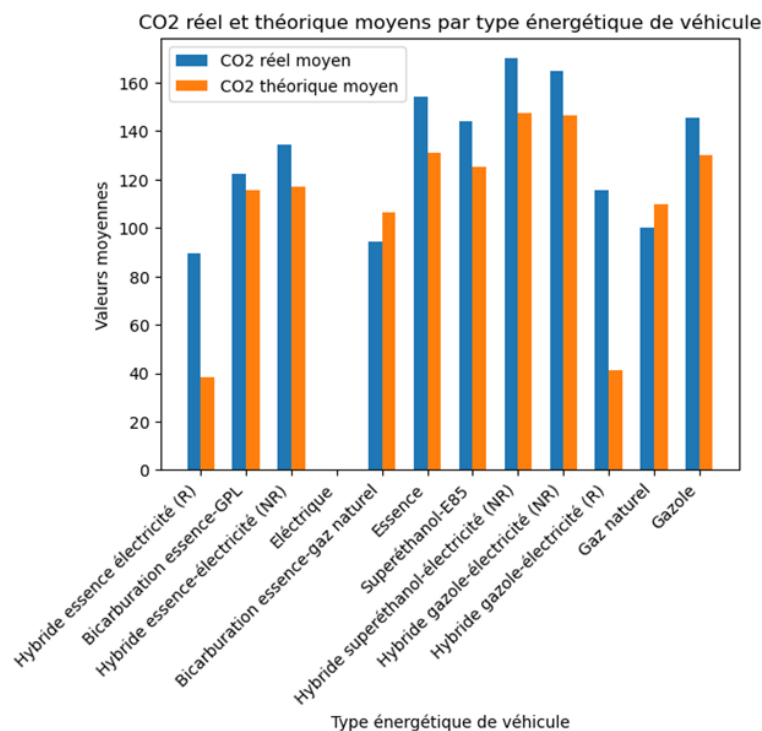


FIGURE 2 – Erreur moyenne entre CO2 théorique et CO2 réel par type d'énergie. Graphe réalisé à partir des données de notre base de travail

## Carrosserie et émission de CO2

La table 4 achève de présenter les statistiques de notre base de données, en précisant les différents types de carrosseries qui composent nos bases de données, ainsi que leur proportion respective. Là encore, nous percevons qu'il existe des différences dans l'erreur du calcul des émissions théoriques relativement aux données de CO2 réel. Par exemple, les émissions des minispaces semblent plutôt bien approchées par les tests initiaux, l'écart relatif entre les deux grandeurs s'élevant à seulement 2,96%. Au contraire, les véhicules tous terrains, voient l'erreur du calcul des émissions s'élever à 21,36%.

Carrosserie	Volume (%)	CO2 théorique (g/km)	CO2 réel (g/km)	Erreur (%)
BERLINE	43.33	116.83	135.96	16.37
BREAK	3.74	111.66	133.76	19.79
CABRIOLET	0.18	156.77	170.96	9.06
COMBISPACE	1.17	142.98	161.79	13.15
COUPE	0.12	146.96	178.53	21.48
MINIBUS	0.52	176.01	188.23	6.94
MINISPACE	0.06	110.59	113.85	2.96
MONOSPACE	0.09	170.28	185.69	9.05
MONOSPACE COMPACT	2.10	138.44	155.82	12.55
TS TERRAINS/CHEMINS	48.69	119.48	145.00	21.36

TABLE 4 – Volume de ventes moyen (en %) par carrosserie, moyenne de CO2 théorique et réel (en g/km), et pourcentage d'erreur entre CO2 théorique et CO2 réel.

Ces différentes figures nous confirment donc l'importance de construire une fonction d'estimation du CO2 améliorée et utilisable lors de la mise en circulation du véhicule, afin de pouvoir par la suite baser les politiques de taxation sur des chiffres plus représentatifs de la réalité et plus justes. Elles nous révèlent les erreurs commises, en précisant l'importance de prendre en compte la nature du véhicule pour mieux en approcher les émissions réelles. L'utilisation de paramètres techniques dans la création d'une fonction de calcul se révèle donc prometteuse.

## 3

# ESTIMATION DES ÉMISSIONS DE CO2 : MODÈLE ET RÉSULTATS

Dans cette section, nous allons nous attacher à construire un modèle de prédiction du CO2 réel émis par un véhicule plus efficace que celui utilisant simplement le CO2 théorique.

## 3.1 LE MACHINE LEARNING AU SERVICE DE LA MESURE DES ÉMISSIONS

Pour construire un estimateur du CO2 réel, nous allons utiliser l'approche du Machine Learning. L'estimation des émissions réelles de CO2 correspond en effet à un **problème de régression** : on cherche à prédire la valeur d'une variable continue. Grâce à notre base de données regroupant caractéristiques et émission réelle de CO2 pour chacun de nos modèles, on peut ainsi le résoudre par **apprentissage supervisé**.

En nous appuyant sur les observations faites de notre base de données de travail, nous retenons les paramètres suivants :

- le CO<sub>2</sub> théorique : alors que le malus actuel ne se fonde que sur cette mesure du CO<sub>2</sub>, il restera dans notre modèle le facteur déterminant.
- des facteurs intrinsèques, quantitatifs et mesurés précisément, dont la cylindrée, le poids total roulant, la charge utile, la puissance et le nombre de décibels.
- des facteurs de catégories : le type d'énergie utilisé (essence, hybride...) et le type de carrosserie (berline, tous-terrains...). Ces paramètres seront traduits sous forme binaire pour pouvoir être utilisés par les algorithmes de régression.

On peut désormais chercher à créer de nouveaux estimateurs du CO<sub>2</sub> réel en appliquant différents modèles de Machine Learning.

## 3.2 SÉLECTION D'UN MODÈLE

L'apprentissage supervisé se divise en trois parties : le choix du modèle, l'entraînement du modèle sur nos données puis l'évaluation du modèle.

Afin de distinguer quel modèle est le plus pertinent dans notre estimation des émissions de CO<sub>2</sub> réelles, il est nécessaire de pouvoir quantifier les performances de chaque modèle.

Cependant, si l'on évalue un modèle sur les données sur lesquelles l'entraînement a été effectué, il y a un risque de **surentraînement** : le modèle pourrait s'adapter parfaitement aux données d'entraînement, mais sans être capable de fournir des résultats satisfaisants sur de nouvelles données. C'est pourquoi pour évaluer les performances d'un modèle de machine learning, il est nécessaire de le tester sur des données qui lui sont inconnues. On distingue donc les **données d'entraînement** (*training set*) et les **données de test** (*test set*). Nous avons décidé dans notre approche de réserver aléatoirement 80% de nos données à l'entraînement de notre modèle et 20% de nos données à sa validation.

La performance d'un modèle de machine learning est aussi dépendante de ses hyperparamètres comme le paramètre alpha dans la régression Lasso et le nombre de voisins dans la méthode KNeighbors. Pour améliorer les performances d'un modèle, nous cherchons donc à en **optimiser les hyperparamètres** pour obtenir les meilleures performances sur les données de test. Pour éviter d'être de nouveau biaisé par le choix de nos données de test dans l'optimisation de notre paramètre, il est alors nécessaire d'avoir une troisième catégorie de données, les **données de validation** (*validation set*).

Notre base de donnée est donc séparée en trois catégories disjointes : les données d'entraînement sur lesquelles le modèle s'entraîne, les données de test sur lesquelles il est évalué et par rapport auquel les hyperparamètres sont optimisés, et les données de validation sur lesquelles seront évaluées les différentes performances des hyperparamètres. On choisit ensuite l'hyperparamètre qui maximise les performances du modèle sur les données de validation qui lui sont totalement inconnues.

On trace alors la **courbe de validation** qui montre les évolutions des performances du modèle en fonction des valeurs de l'hyperparamètre de notre modèle. Cela nous permet alors de choisir l'hyperparamètre qui maximise les performances du modèle.

Cependant, cette démarche laisse encore la place à un biais important : il est possible que le modèle marche spécifiquement bien pour le données testées mais que sa performance générale ne soit pas satisfaisante. Autrement dit, le choix de la catégorisation en données d'entraînement, de test et de validation peut avoir un impact important sur ses performances. Pour éviter ce biais, nous procédons à la méthode de **validation croisée** : nous entraînons et validons à 5 reprises notre modèle sur des découpages différents choisis de manière aléatoire. Les différents scores utilisés dans le rapport sont ainsi les moyennes des scores obtenus sur les différents découpages afin qu'il soit plus représentatifs de la réalité. On notera alors *train score*, *test score* et *validation score*

le score du modèle respectivement sur les données d'entraînement, les données de test et les données de validation.

Une dernière question nous a aussi semblé importante dans le choix du modèle : avons-nous assez de données pour que le modèle ait fini son apprentissage ? Pour répondre à cette interrogation, nous avons tracé les **courbes d'apprentissage** qui permettent de visualiser l'évolution des performances du modèle en fonction du **nombre de données d'entraînement** qu'on lui fournit. On regarde alors les valeurs de nos scores en fonction du nombre de données d'entraînement que nous avons à disposition. Lorsque cette courbe atteint un plafond d'apprentissage, on peut alors en déduire qu'il n'est pas nécessaire de lui fournir plus de données.

Le score associé à un modèle est un nombre positif qui permet de quantifier la différence entre les valeurs prédites par le modèle et les valeurs réelles enregistrées dans notre base.

Cette méthode a ainsi été appliquée sur nos trois modèles : régression linéaire, régression Lasso et KNeighbors.

### 3.3 APPROCHE SIMPLE : LA RÉGRESSION LINÉAIRE

Une première approche consiste à utiliser une régression linéaire multivariée sur les 27 dimensions pour trouver la meilleure approximation du CO<sub>2</sub> réel.

Après avoir normalisé les données, on les sépare entre un groupe d'entraînement (80% de la base) et un groupe de test, sur lequel notre modèle n'apprendra pas et qui servira à évaluer ses performances sans impact du surapprentissage (20% de la base).

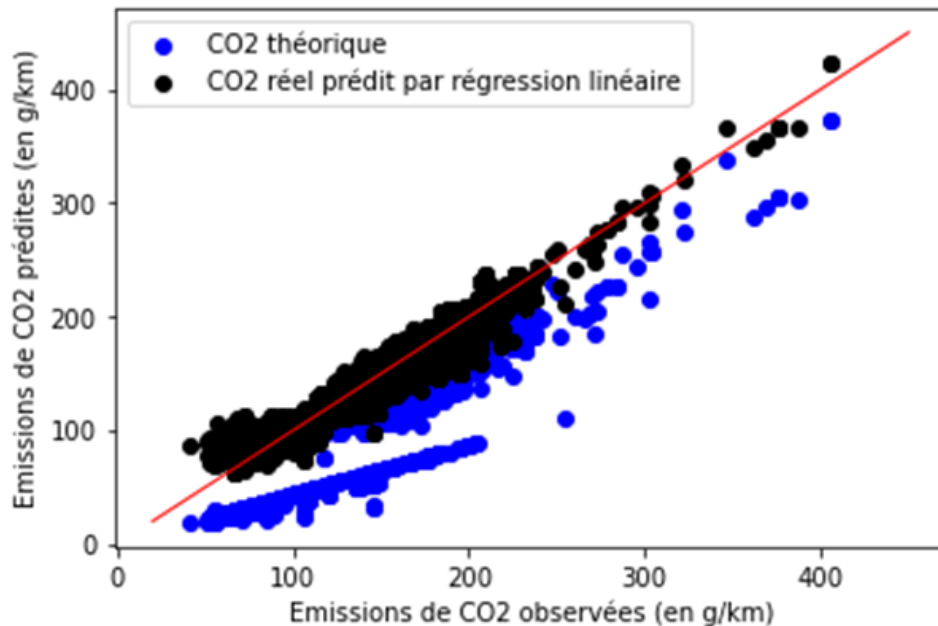


FIGURE 3 – Les prédictions du modèle et le CO<sub>2</sub> théorique par rapport aux valeurs réelles du CO<sub>2</sub> (en g/km)

Ce graphique représente en ordonnée les prédictions faites, soit par notre modèle ou par le gouvernement, avec en abscisse le CO<sub>2</sub> réel. Chaque point représente un véhicule du groupe test. La ligne rouge, l'identité permet de repérer la prédiction optimale.

En bleu sont représentées les prédictions du CO<sub>2</sub> réel faites par le gouvernement en utilisant le CO<sub>2</sub> théorique. On peut observer qu'elles sont largement en dessous de la courbe et plus éloignées par rapport aux points noirs, qui représentent les prédictions de notre modèle. De plus, on observe une ligne bleue qui se dessine avec des mesures très sous estimées du CO<sub>2</sub> réel. Une étude de la base de donnée révèle que cette ligne très mal estimée par le gouvernement est celle des véhicules hybrides, que notre modèle parvient pourtant à évaluer correctement.

Afin d'évaluer les performances du système, on va utiliser deux métriques : le coefficient  $R^2$  et le coefficient  $RMSE$ . Le coefficient  $R^2$  représente le pourcentage de variation expliqué par rapport à la variation totale, et évalue donc la pertinence de la régression linéaire. Le coefficient  $RMSE$  représente la racine de l'erreur moyenne au carré (mean-squared-error) et évaluera donc la précision des prédictions. Le coefficient  $R^2$  est celui qui est utilisé dans le calcul des *train score* et *test score* des graphes des Figure 4, Figure 6, Figure 8, et Figure 9.

Les coefficients  $R^2$  et  $RMSE$  sont respectivement donnés par les formules :

$$R^2 = 1 - \frac{\sum_{i=1}^n (CO2_i - \hat{CO2}_i)^2}{\sum_{i=1}^n (CO2_i - CO2_{\text{moyen}})^2} \text{ et } RMSE = \sqrt{\frac{\sum_{i=1}^n (CO2_i - \hat{CO2}_i)^2}{n}}$$

où les  $CO2_i$  sont les émissions réelles (observées) et les  $\hat{CO2}_i$  les émissions prédites par notre modèle.

#### Modèle de l'État :

Pour le modèle proposé par l'état fondé sur le CO<sub>2</sub> théorique, on obtient :

- $R^2 = 0,02$
- $RMSE = 29,04$

La valeur du  $R^2$  est ici peu significative, car on ne s'attend pas précisément à une relation linéaire. Cependant, le  $RMSE$  est très élevée (pour des valeurs de CO<sub>2</sub> généralement de 200 g/km) avec une erreur relative d'environ 15%.

#### Notre premier modèle :

Pour notre premier modèle, on obtient :

- $R^2 = 0,902$
- $RMSE = 9,21$

Ceci est cohérent avec les statistiques présentées dans le tableau 1.

On observe immédiatement une nette amélioration des performances : non seulement le modèle de la régression linéaire semble cohérent, mais on obtient un  $RMSE$  de 9,21, soit une erreur relative d'environ 5% et un modèle trois fois plus performant que celui de l'État.

En outre, ce modèle parvient à très bien prédire le CO<sub>2</sub> émis par les véhicules hybrides, dont les émissions sont d'ordinaire largement sous-estimées.

À titre informatif, les coefficients des paramètres calculés par le modèle sont donnés ci-dessous dans la table 5.

Cylindrée moyenne	-1.285685e+00
Moyenne CO2 théorique	2.470390e+01
Poids total roulant moyen	9.841190e-01
Charge utile moyenne	1.129916e+00
Puissance moyenne (en KW)	1.034678e+01
Nombre de décibels moyen	4.224617e-01
Essence	3.801911e+12
Gazole ou diesel	3.696314e+12
Superéthanol	4.238528e+11
Essence-électricité hybride rechargeable	2.186414e+12
Essence-électricité hybride non rechargeable	2.361766e+12
Gazole-électricité hybride non rechargeable	2.238910e+12
Gaz naturel	5.573457e+10
Gazole-électricité hybride rechargeable	6.559842e+11
Bicarburant essence-GPL	5.433648e+11
Superéthanol-électricité hybride non rechargeable	6.442025e+11
Bicarburant essence-gaz naturel	3.941079e+10
BERLINE	1.082985e+13
TS TERRAINS/CHEMINS	1.178271e+13
BREAK	7.161672e+12
MINIBUS	2.844720e+12
MONOSPACE COMPACT	4.239172e+12
COMBISPACE	3.610189e+12
COUPE	2.259825e+12
MONOSPACE	1.459133e+12
MINISPACE	5.414788e+11
CABRIOLET	1.691968e+12

TABLE 5 – Coefficients des paramètres, obtenus par régression

### 3.4 LES AMÉLIORATIONS DE LA RÉGULARISATION LASSO

Afin de rendre notre modèle davantage utilisable dans la pratique, on peut utiliser une régularisation Lasso. En pénalisant l'utilisation de nombreux paramètres pendant l'apprentissage, cette régularisation va permettre de réduire le nombre de facteurs "déterminants", et ainsi donner une formule plus simple, avec moins de paramètres.

#### Optimisation de l'hyperparamètre $\alpha$

Comme nous pouvons l'observer en Figure 4, les résultats de la courbe de validation montre que les performances du modèle sont maximales lorsque  $\alpha$  vaut 0,1.

Courbe de validation : choix de l'hyperparamètre  $\alpha$  dans le modèle Lasso

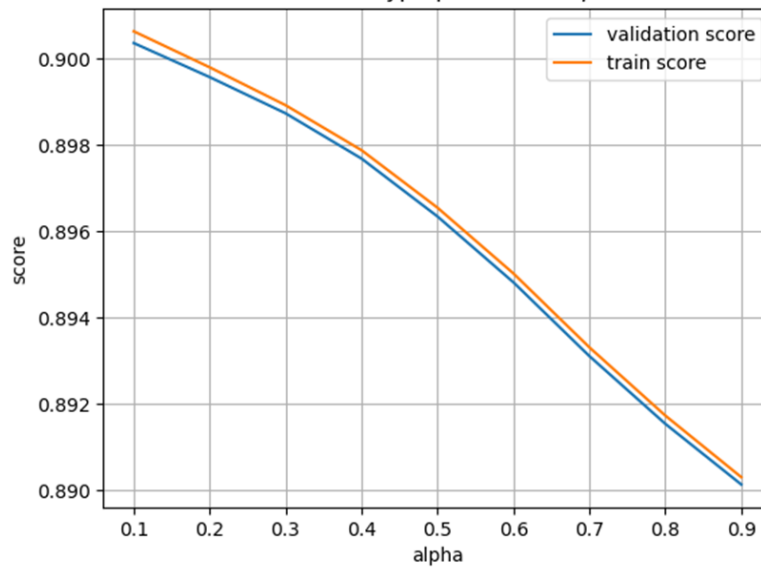


FIGURE 4 – Courbe de validation du modèle Lasso : visualisation des scores en fonction du choix du paramètre  $\alpha$ .

### Résultat du modèle Lasso

Pour ce modèle régularisé avec Lasso ( $\alpha = 0,1$ ), on obtient :

- $R^2 = 0,90$
- $RMSE = 9,29$

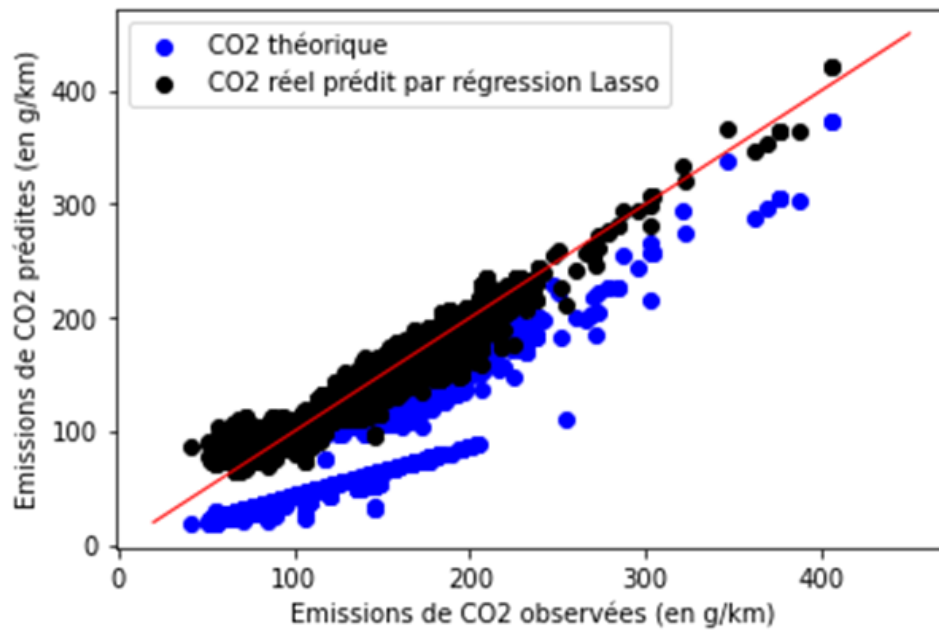


FIGURE 5 – Les prédictions du modèle avec Lasso par rapport au CO2 réel



On observe ici qu'avec désormais seulement 17 paramètres, on obtient des performances quasi similaires.

En particulier, on trouve comme coefficients :

Moyenne CO2 théorique	24.61
Poids total roulant moyen	0.29
Charge utile moyenne	1.36
Puissance moyenne (en KW)	9.57
Nombre de décibels moyen	0.33
Essence	3.53
Gazole ou diesel	-1.09
Superéthanol	0.27
Gazole-électricité hybride non rechargeable	-1.18
Gazole-électricité hybride rechargeable	0.46
Bicarburant essence-GPL	-0.43
Superéthanol-électricité hybride non rechargeable	0.23
TS TERRAINS/CHEMINS	0.77
BREAK	-0.40
MINIBUS	0.02
COMBISPACE	0.81
CABRIOLET	-0.14

TABLE 6 – Coefficients pondérant les paramètres, obtenus par régularisation Lasso

Cette régression peut se révéler utile notamment dans le cas d'une mise en pratique du modèle, pour le rendre plus intelligible et simple sans perdre en performance.

### Courbe d'apprentissage du modèle Lasso

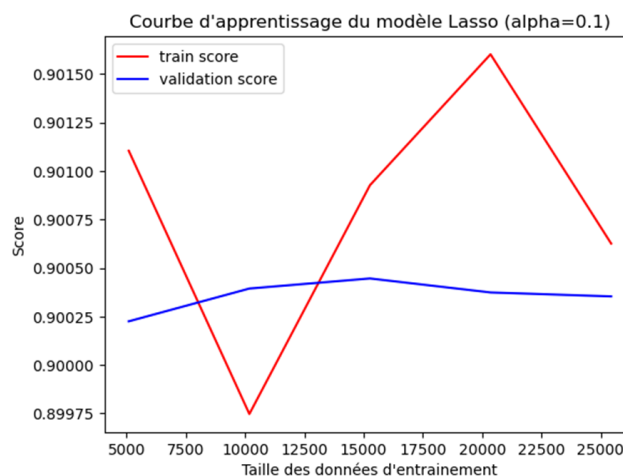


FIGURE 6 – Courbe d'apprentissage du modèle Lasso.

Comme on l'observe en Figure 6, A partir de 12 000 données, on voit que le train score fluctue légèrement mais que le validation score atteint un plateau. Comme le validation score atteint un plateau, on en déduit que le modèle est capable de généraliser et que la taille de notre données est donc suffisante pour que le modèle soit

bien calibrée.

### 3.5 LES AMÉLIORATIONS DU MODÈLE KNN

Un autre modèle puissant de régression adapté à ce problème est le modèle K-nearest-neighbors (KNN). Celui-ci consiste à prédire le CO<sub>2</sub> émis par un véhicule en fonction de ses "voisins" dans l'espace vectoriel de dimension le nombre de paramètres. En prenant la moyenne des émissions d'un nombre  $n$  de plus proches voisins d'un véhicule, on peut prédire la valeur de ses émissions.

Le choix du nombre de voisins pris en compte dans le modèle KNN impacte directement le comportement du modèle, notamment sa stabilité : un plus faible nombre de voisins permet de mieux traiter les cas rares mais est plus variable tandis qu'un grand nombre de voisin permet d'être plus robuste et fiable.

On entraîne ici le modèle sur la même base que la régression Lasso avec  $n = 3$  voisins.

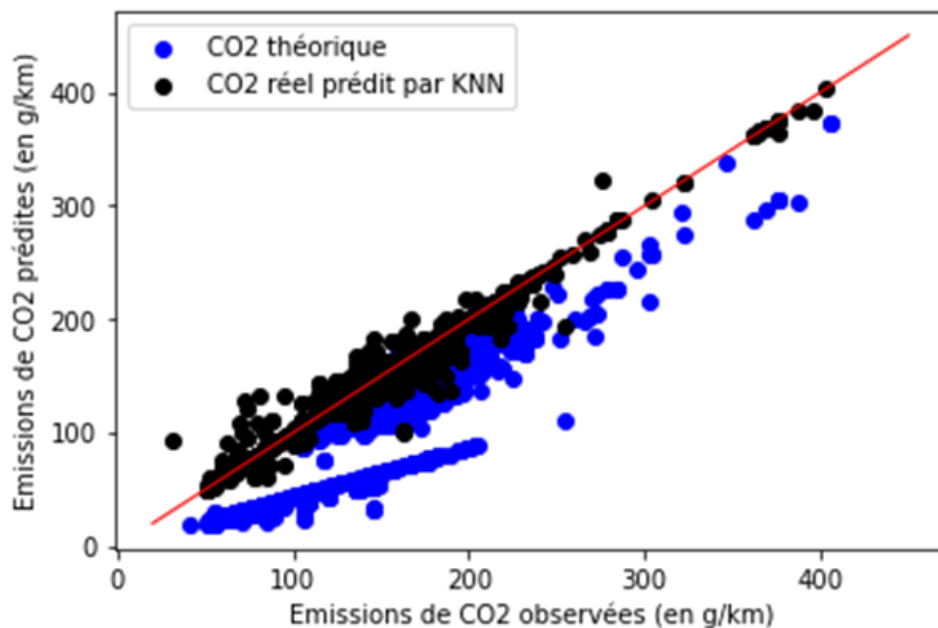


FIGURE 7 – Les prédictions du modèle KNN par rapport au CO<sub>2</sub> réel

On observe ici que le modèle KNN produit des prédictions en moyenne plus précises que les régressions. En particulier, on obtient un  $RMSE = 3.38$  et un  $R^2 = 0.987$ . Ces scores excellents indiquent que la meilleure méthode de prédiction serait la méthode KNN.

Cependant, on observe que la méthode KNN présente un plus grand nombre de valeurs aberrantes que les méthodes de type Lasso, car les valeurs anormales ont davantage tendance à s'agréger entre elles alors qu'elles sont gommées dans les régressions linéaires.

On trace la performance en fonction de l'hyperparamètre  $n$  : le nombre de voisins sur lequel l'estimation s'appuie. Ceci va nous permettre de déterminer le meilleur choix de l'hyperparamètre.

Courbe de validation : Choix de l'hyperparamètre  $n$  dans le modèle KNeighbors

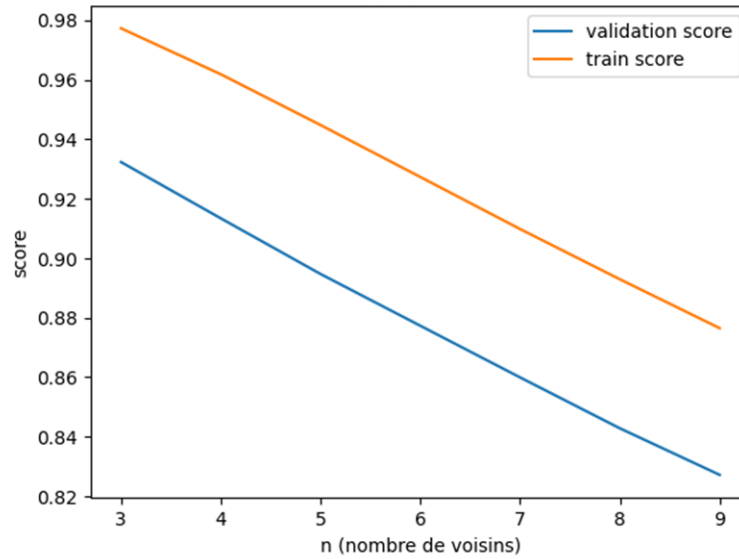


FIGURE 8 – Courbe de validation du modèle KNN : visualisation des scores en fonction du choix du paramètre  $n$ .

Ce graphe atteint son maximum en  $n = 3$ , ce qui justifie donc ce choix d'hyperparamètre pour prédire le CO2 réel.

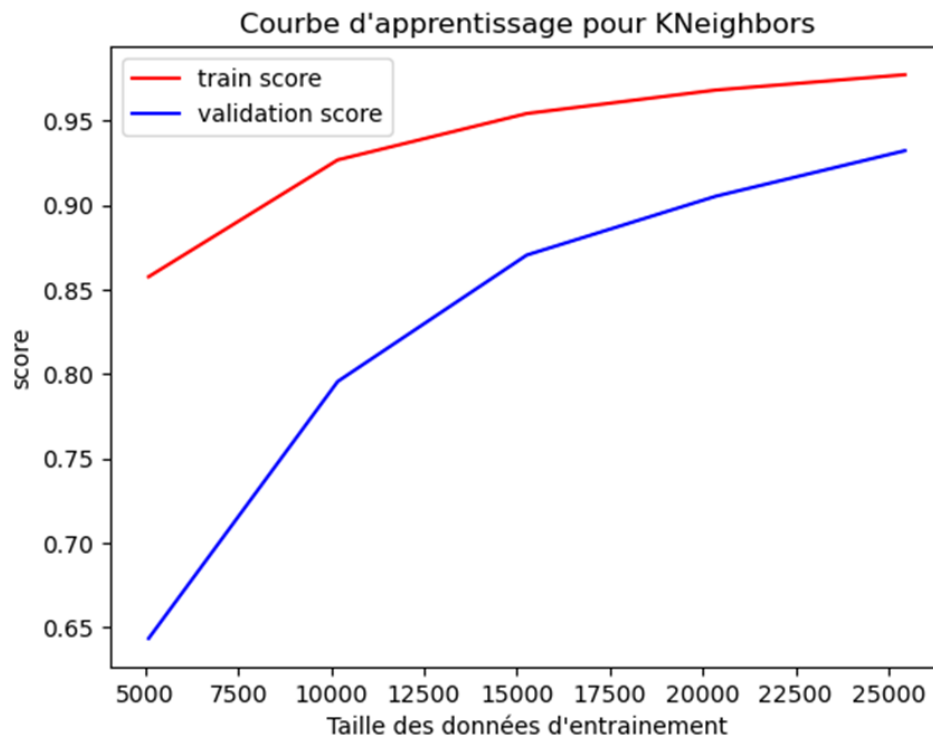


FIGURE 9 – Courbe d'apprentissage du modèle KNN.

À nouveau on observe qu'à partir de 25000 données, la courbe d'apprentissage se stabilise et atteint un plateau, ce qui montre que le modèle se généralise et qu'il est bien calibré.

Cependant, en raison de ce problème de valeurs aberrantes et face à la simplicité de la régression Lasso, on optera dans la suite pour une prédiction du CO<sub>2</sub> réel par la régression Lasso.

À l'issue de cette partie, nous avons donc réussi à créer une méthode de calcul des émissions, implémentable avant la mise en vente du véhicule, et se rapprochant très sensiblement des émissions réelles observées. Cette partie est donc un succès.

## 4

# EFFETS DE CETTE NOUVELLE ESTIMATION SUR LA DEMANDE DE VÉHICULES ET SUR LES ÉMISSIONS RÉSULTANTES

Aujourd'hui, le malus écologique du gouvernement se construit sur les émissions de CO<sub>2</sub> théoriques de chaque véhicule. Or, grâce au machine learning, nous avons réalisé une estimation plus précise des émissions *réelles* de CO<sub>2</sub> par véhicule. Nous pouvons désormais envisager de construire le malus directement sur les émissions réelles de CO<sub>2</sub> estimées par voiture. Cela devrait permettre à la fois d'avoir un malus plus juste car plus représentatif de l'impact écologique réel de chaque voiture, mais aussi plus efficace dans la réduction de la moyenne des émissions réelles par voiture. Nous avons donc cherché à étudier l'impact sur la demande qui résulterait de la correction du malus écologique, afin de vérifier l'effet positif sur les émissions finales auquel nous nous attendons.

Nous présenterons dans un premier temps notre modélisation de la demande de voitures par les consommateurs français, qui se construit sur le modèle économique de logit. Dans un second temps, nous évaluerons l'impact de notre proposition sur le malus 2021 et le malus 2024.

### 4.1 MODÈLE ÉCONOMIQUE

Pour obtenir un malus plus fiable que celui fondé sur le CO<sub>2</sub> théorique, nous confrontons les résultats de meilleure approche du CO<sub>2</sub> réel (exposés en deuxième partie) à un modèle de demande.

Une approche économique courante consiste à faire l'hypothèse que le consommateur maximise son utilité. Avant de faire un choix pour un produit, celui-ci prend en compte un set de facteurs de préférences, comme le prix, la qualité, la taille... On peut alors modéliser le choix d'un consommateur par une fonction, la fonction d'utilité, dépendant de ces paramètres préférentiels. Après lecture du livre de Kenneth E. Train, Discrete Choice Methods with Simulation (TRAIN, 2009), le choix préférentiel nous est apparu être le modèle Logit, qui estime les parts de marché dans un domaine de consommation, comme l'automobile, à partir de données récoltées ayant une influence.

#### 4.1.1 • PRINCIPE DU MODÈLE LOGIT ET ESTIMATION DES PARTS DE MARCHÉ

L'ensemble des choix proposés au consommateur doit être un ensemble satisfaisant 3 conditions : être exhaustif, admettre un nombre fini d'alternatives (que nous noterons  $N$ ), et être mutuellement exclusif. Dans notre cas, les bases de données employées satisfont bien ces 3 conditions, l'ensemble des véhicules neufs vendus

étant répertorié.

On notera  $U_j$  l'utilité du véhicule  $j$ . Il s'agit d'une grandeur abstraite qui mesure la propension du consommateur à choisir le véhicule  $j$ . Plus elle est grande, plus le consommateur a intérêt à choisir le véhicule, et plus elle est faible, moins l'intérêt est grand. Par souci de simplification, tous les consommateurs sont supposés être égaux dans leur choix de véhicule.

Nous détaillerons par la suite comment se décompose l'utilité  $U_j$  associée au véhicule  $j$ . Nous détaillons ici ce que prévoit le modèle logit : un ajustement des parts de marché en fonction de l'utilité du consommateur. En notant  $P_i$  la part de marché du véhicule  $i \in \llbracket 1; N \rrbracket$ , la modélisation prévoit

$$P_i = \frac{e^{U_i}}{\sum_{j=1}^N e^{U_j}} \quad (1)$$

Notons ici que ces parts se somment effectivement à 1. Les utilités étant, nous l'avons précisé, dépendantes du prix du véhicule, ces parts de marché sont sensibles à un changement de prix, ce qui est déterminant dans l'implémentation d'un nouveau malus tel que le veut notre étude.

#### 4.1.2 • CHOIX DE LA FONCTION D'UTILITÉ

D'après le livre de Kenneth E. Train, *Discrete Choice Methods with Simulation* (TRAIN, 2009) l'utilité du consommateur peut être écrite comme une combinaison linéaire de paramètres déterministes et non déterministes :

$$U_j = \sum_k \beta_k x_{kj} \quad (2)$$

avec  $x_{kj}$  les paramètres linéairement indépendants et influents.

Le seul paramètre préférentiel que nous pouvons manipuler facilement est le prix, les autres paramètres étant inobservés et/ou inobservables, comme la préférence de tel ou tel consommateur pour un véhicule à 5 places, pour un véhicule avec une grande autonomie... Le prix est mesuré, et chaque consommateur y fait face de la même manière en moyenne.

Sous ces restrictions, on peut réécrire :

$$U_j = \beta p_j + \delta_j \quad (3)$$

où les  $\delta_j$  renferment l'ensemble des paramètres préférentiels inobservés du consommateur.

Enfin, pour cerner l'impact réel du prix sur le consommateur, il s'agit d'estimer le paramètre bêta. Une simple régression linéaire sur les parts de marché ne serait pas très concluante ici ; en effet, il faudrait supposer que tous les modèles ont la même attractivité hors-prix pour le consommateur, ce qui n'est pas vrai. Pour pallier à ce problème, nous avons choisi une borne supérieure de ce paramètre, introduit dans le rapport *Évaluation des mesures de soutien aux véhicules propres* (DURMEYER et al., 2024),  $\beta = -0.00015$ .

#### 4.1.3 • MESURE INDIRECTE DES PRÉFÉRENCES DES CONSOMMATEURS (DELTA)

Grâce à la base de donnée fournie par l'ADEME, nous avons accès aux véhicules vendus en 2021 qui étaient alors soumis au malus 2021. Le volume des ventes de chaque modèle de véhicule reflète donc à la fois l'importance que les consommateurs accordent à son prix augmenté du malus (information reflétée par le paramètre *beta*) mais aussi la valeur intrinsèque que les consommateurs lui accordent (que l'on va enregistrer dans la variable *delta*).

Il convient pour achever l'implémentation du modèle, de déterminer la valeur de la variable *delta*, pour chaque modèle du dataset. Il est important de noter ici que l'utilité est une grandeur abstraite qu'il est impossible de déterminer sans connaître l'ensemble des grandeurs qui la définissent.

Ainsi, il est impossible de déterminer ces paramètres de préférences intrinsèques à partir de la seule formule (3). Pour implémenter les variables  $\delta_j$ , il faut remonter à l'équation (1). Les parts de marché étant connues à partir des données de l'ADEME, et le dénominateur  $\sum_{j=1}^N e^{U_j}$  une constante du modèle, il suffit de se donner un modèle de référence  $c$  pour lequel on choisit arbitrairement, et sans perte de généralité,  $\delta_c = 0$ . On écrit ensuite l'équation (1) pour tout modèle  $i$  :

$$\begin{cases} \ln(P_i) = \beta p_i + \delta_i - \ln(\sum_{j=1}^N e^{U_j}) \\ \ln(P_c) = \beta p_c + \delta_c - \ln(\sum_{j=1}^N e^{U_j}) \end{cases} \quad (4)$$

d'où en soustrayant les deux équations,

$$\delta_i = \ln\left(\frac{P_i}{P_c}\right) - \beta(p_i - p_c) \quad (5)$$

Toutes les valeurs de l'équation (5) sont des données de la base ADEME. Il suffit alors de se donner un modèle de référence, et l'implémentation en découle aisément algorithmiquement. Grâce à cette astuce de calcul, nous avons pu enregistrer l'ensemble des préférences du consommateur pour chacun des modèles : déterministes via le prix et non déterministes via les variables *delta*.

Nous montrons ici, à titre d'information, l'histogramme des valeurs prises par ces variables lors de l'implémentation du modèle. Notons que cet histogramme, en révélant la distribution et les valeurs prises par le paramètre *delta*, illustre bien l'importance de prendre en compte cette variable pour créer un modèle juste de prédiction des parts de marché.

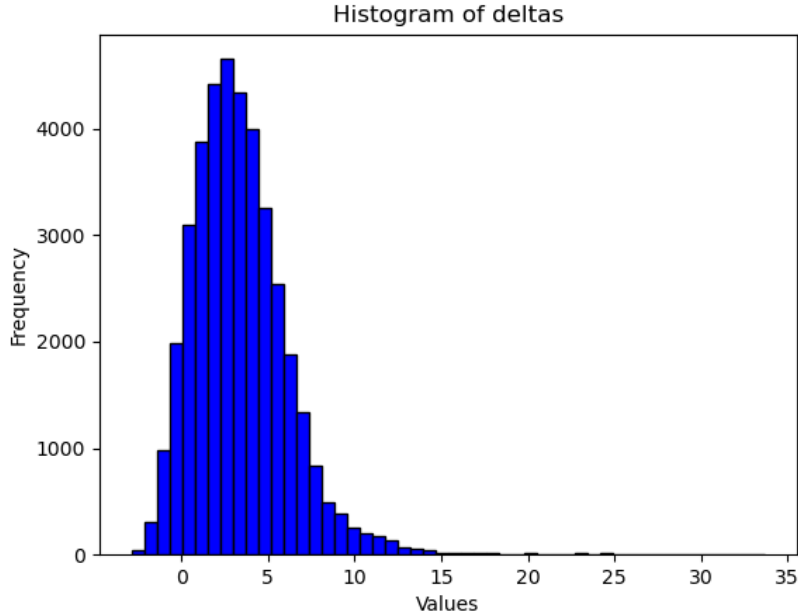


FIGURE 10 – **Histogramme des paramètres *delta***. Ce graphique présente les estimations des paramètres  $\delta$  qui capturent la qualité inobservée des véhicules. Voir l'équation (5) et le texte associé pour plus de détails.

#### 4.1.4 • ESTIMER LES ÉMISSIONS DE CO<sub>2</sub> CONSÉCUTIVES AUX CHOIX DES CONSOMMATEURS GRÂCE À LA MODÉLISATION DE LA DEMANDE

Maintenant que nous avons un modèle économique nous permettant de connaître les parts de marché de chaque modèle de véhicule en fonction de son prix, il nous est possible d'en déduire les émissions de CO<sub>2</sub> que l'on s'attend à observer en tenant compte de la répartition des ventes. Nous utilisons donc le modèle logit pour étudier l'impact de l'implémentation d'un malus sur les parts de marché. Les utilités étant sensibles au prix, on peut reconstruire grâce à l'équation (1), les émissions réelles contrefactuelles moyennes en présence de n'importe quel malus. On écrit :

$$U_j^a = \exp \beta(p_j + m_j) + \delta_j \quad (6)$$

avec  $p_j + m_j$  le prix actualisé du malus.

Les émissions moyennes réelles résultant des choix des consommateurs, sont alors données par la formule suivante :

$$CO_2^{\text{moyen réel}} = \sum_{j=1}^N P_j^a \times CO_2^{\text{réel contrefactuel}} \quad (7)$$

avec  $P_j^a$  les parts de marchés résultant de prix par véhicules influencés par la politique de taxation

## 4.2 MISE EN ÉVIDENCE EXPÉRIMENTALE DES AMÉLIORATIONS PERMISES PAR UN CALCUL PLUS JUSTE DES ÉMISSIONS

### 4.2.1 • LE MALUS MIS EN PLACE PAR L'ÉTAT EN 2021 ET EN 2024

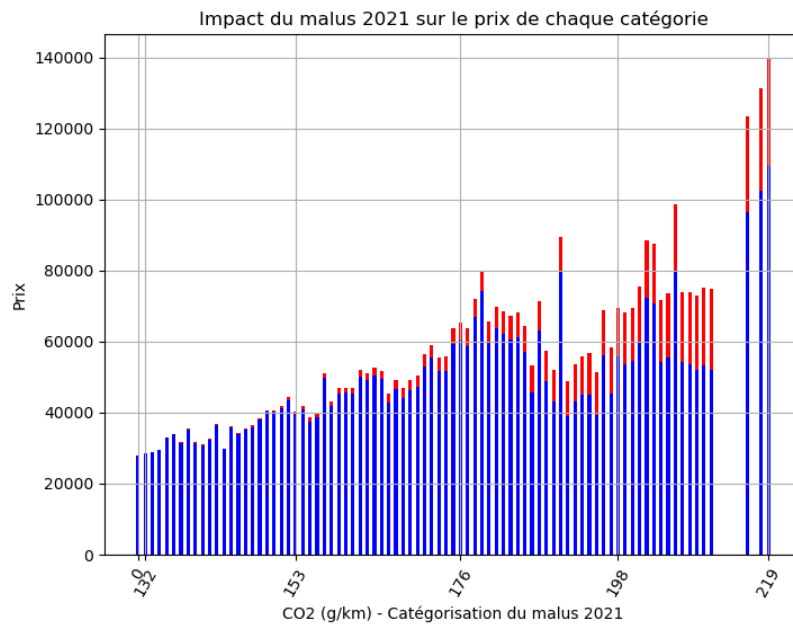
La première étape de notre démarche a été de nous procurer les grilles de taxation du malus écologique proposé en 2021 et en 2024 afin d'en étudier les effets, puis d'appliquer nos contrefactuelles à ces grilles. Celles-ci associent une taxe précise à un niveau d'émission donné, correspondant aux émissions que l'on pense être celles du véhicule avant sa mise en vente.

Nous nous sommes procuré le tableau des montants du malus 2021 sur le site de l'Argus spécialisé dans la vente de véhicules neufs et d'occasions (les montants ont été corroborés par d'autres journaux d'actualité automobile présentant tous les mêmes montants en 2021) .

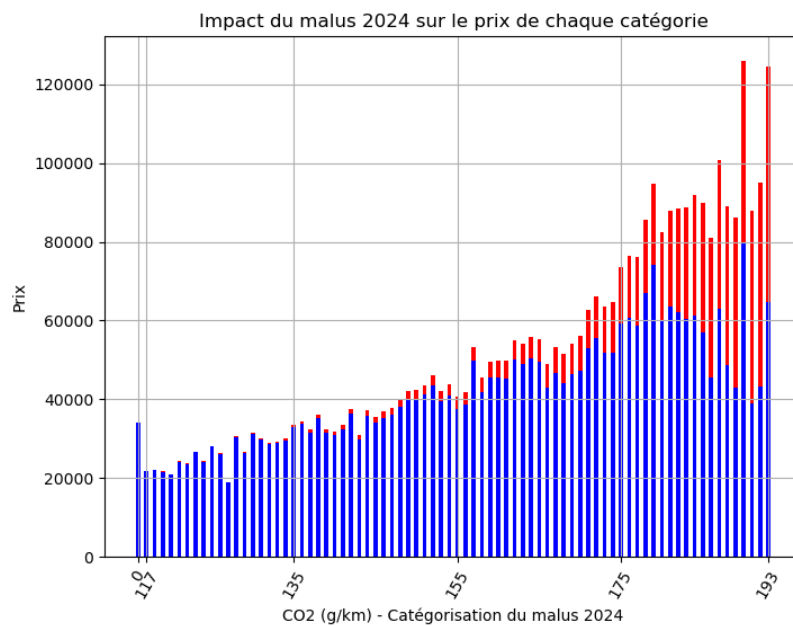
Le malus 2024 est quant à lui disponible sur le site du Service Public du gouvernement.

Une fois les deux tableaux implémentés dans nos fichiers, nous avons créé une fonction associant à chaque émission théorique de CO<sub>2</sub>, préalablement évaluée par le constructeur, son malus 2021 et son malus 2024. La figure 11 décrit ainsi l'état des lieux avant les corrections que nous souhaitons apporter. Le malus de l'état est croissant et calculé à partir des émissions théoriques de chaque modèle affichées lors de sa commercialisation.

On peut également noter, en comparant les graphes de la figure 11, de grandes différences entre les malus de 2021 et de 2024 : la taxe maximale est passée en trois ans seulement de 30000.00€ à 60000.00€, ce qui constitue, pour les catégories d'émissions les plus hautes, une multiplication du prix moyen par 2 en 2024.



(a) 2021



(b) 2024

FIGURE 11 – L'effet des malus étatiques sur le prix moyen de chaque catégorie de véhicules.

#### 4.2.2 • RÉSULTATS OBTENUS PAR NOTRE MODÉLISATION DE LA DEMANDE

Connaissant la grille de taxation, qui fournit une valeur de malus pour un niveau de CO2 estimé avant la vente, nous sommes désormais en mesure d'utiliser le modèle logit décrit en 4.1, pour évaluer les émissions que



l'on s'attend à observer dans différents scénarios : pour chacun des deux malus étatiques étudiés (2021 et 2024), sans changer la grille de tarification de l'État, nous avons observé la différence d'émissions résultantes prédites, entre un malus fondé sur les émissions théoriques et un malus fondé sur les émissions contrefactuelles issues de la partie 3. Nous avons également calculé les émissions que l'on obtiendrait si aucun malus n'était implémenté. Enfin, pour vérifier la justesse de notre modèle de demande, nous avons calculé les émissions moyennes qui ont été observées en 2021, à partir de nos données qui répertoriaient le nombre de ventes de chaque véhicule.

L'ensemble des résultats est donné dans le tableau 7. Notons que nous avons divisé le tableau en deux : les quatre premières lignes donnent des résultats d'émissions théoriques moyennes. Les acteurs publics raisonnent en effet actuellement en termes d'émissions théoriques. Les quatre lignes suivantes proposent des résultats d'émissions réelles moyennes.

	Malus 2021 appliqué aux émissions théoriques	Malus 2021 appliqué à nos émissions réelles	Malus 2024 appliqué aux émissions théoriques	Malus 2024 appliqué à nos émissions réelles
Moyenne CO2 théorique (observée)	119,149	119,419	119,419	119,419
Moyenne CO2 théorique (estimée sans malus)	127,256	127,256	127,256	127,256
Moyenne CO2 théorique (estimée avec malus)	119,149	116,592	117,197	112,409
Baisse des émissions permises grâce au malus (%)	6,804	9,147	8,584	13,209
Moyenne CO2 réel (observée)	141,420	141,420	141,420	141,420
Moyenne CO2 réel (estimée sans malus)	150,409	150,409	150,409	150,409
Moyenne CO2 réel (estimée avec malus)	141,420	138,215	139,617	133,689
Baisse des émissions permises grâce au malus (%)	6,356	8,823	7,730	12,507

TABLE 7 – **Tableau des émissions moyennes résultant de plusieurs scénarios de taxation.** A partir des grilles de tarification proposées en 2021 puis en 2024, nous modifions la demande en modifiant la valeur intrinsèque du malus de chaque véhicule.

#### 4.2.3 • ANALYSE DES RÉSULTATS

— **Validation de la pertinence du modèle**

Notons d’abord que ces résultats montrent que le modèle de demande que nous implémentons est pertinent et bien initialisé. En effet, nous observons que la valeur de la moyenne du CO<sub>2</sub> réel effectivement observé en exploitant dans nos données le nombre de ventes de chaque véhicule, est égale à la valeur du CO<sub>2</sub> réel que prédit notre modèle lorsque nous imposant un prix correspondant au malus 2021, alors en place.

Nous notons également qu’une réduction des émissions réelles comme théoriques et observée en présence du malus écologique implémenté par l’état, par rapport à ce que prévoit le modèle en l’absence de toute taxation. Ce résultat témoigne intrinsèquement de l’efficacité du malus écologique, tandis que la cohérence de ce résultat confirme que la modélisation de la demande utilisée ici est réaliste.

— **Réduction des émissions par l’amélioration de l’estimation du CO<sub>2</sub> lors de la mise en vente**

Le résultat est frappant. Avec le malus fondé sur les émissions contrefactuelles, les baisses d’émissions sont bien plus significatives qu’avec les malus étatiques calculés à partir des normes théoriques. En termes d’émissions réelles pures, les pourcentages de baisse d’émissions passent de 6.3 % à 8.82% en 2021, et de 7.73 % à 12.51 % en 2024.

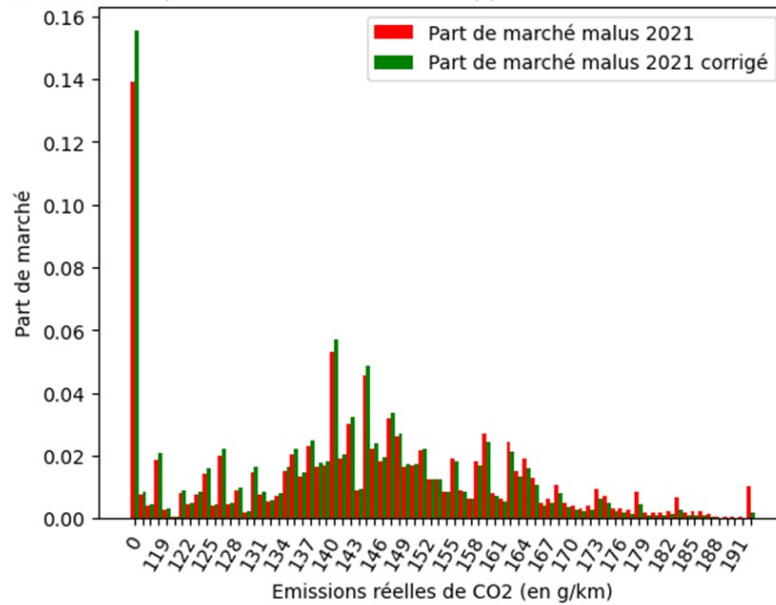
— **Un ajustement des parts de marché logit après implémentation des malus**

Il est important de noter ici la cohérence du modèle logit qui, à travers l’utilité décroissante du consommateur en fonction du prix, augmente les parts de marché des modèles peu émetteurs, moins taxés, et diminue celles des modèles émetteurs, fortement taxés. Tant en 2021 qu’en 2024, les résultats de notre modèles sont cohérents et significatifs.

#### 4.2.4 • VISUALISATION DE NOS CORRECTIONS SUR LES PRIX ET PARTS DE MARCHÉ

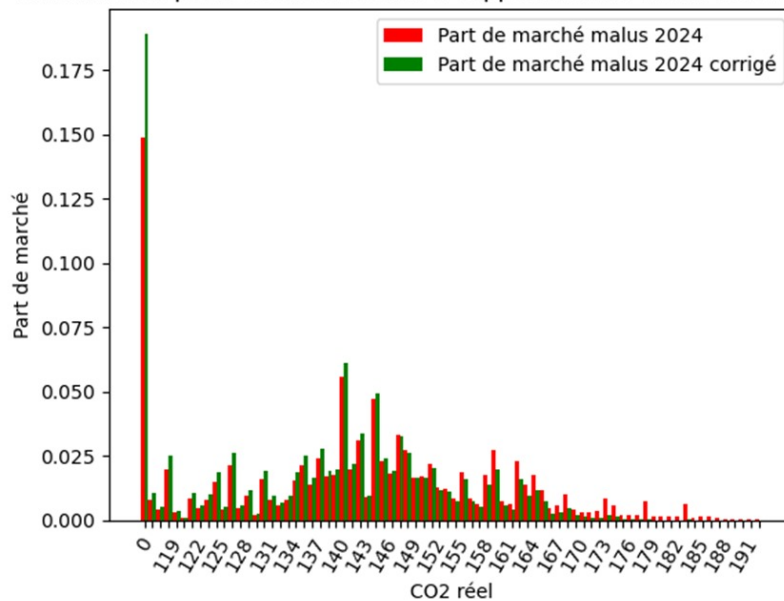
Comme nous l’attendions, notre correction sur le calcul des émissions dont est responsable un véhicule a modifié la demande de véhicules, conduisant à une diminution globale des émissions. Les graphes suivants, représentant les variations des parts de marchés dans chaque catégorie à la suite de notre correction confirme bien que la prise en compte des émissions réelles, affectant le prix individuel des véhicules, modifie les ventes en faveur des catégories les moins émettrices.

Variation des parts de marché suite à l'application du malus 2021 recalculé



(a) Application au malus 2021

Variation des parts de marché suite à l'application du malus 2024 recalculé

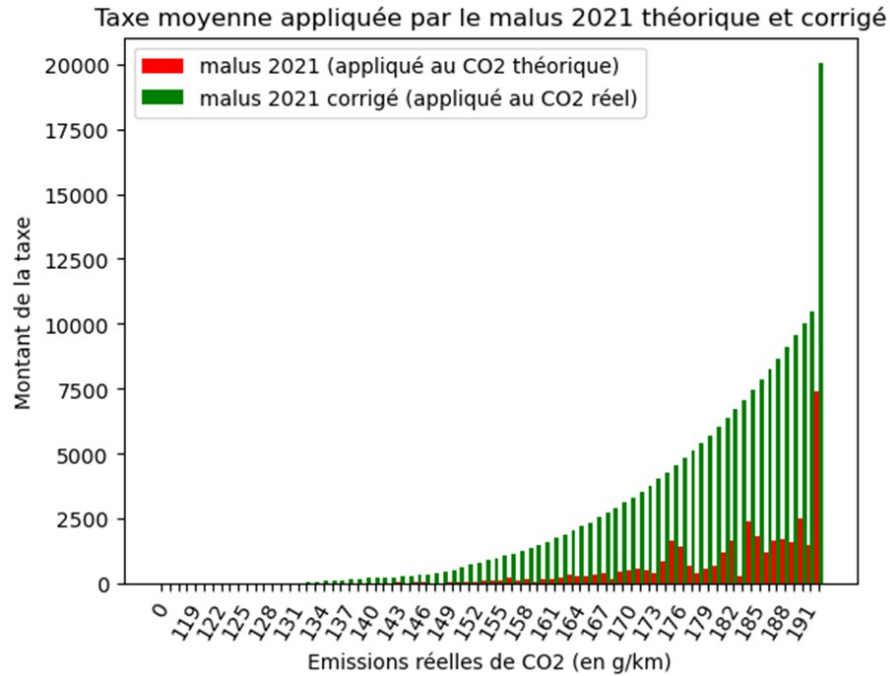


(b) Application au malus 2024

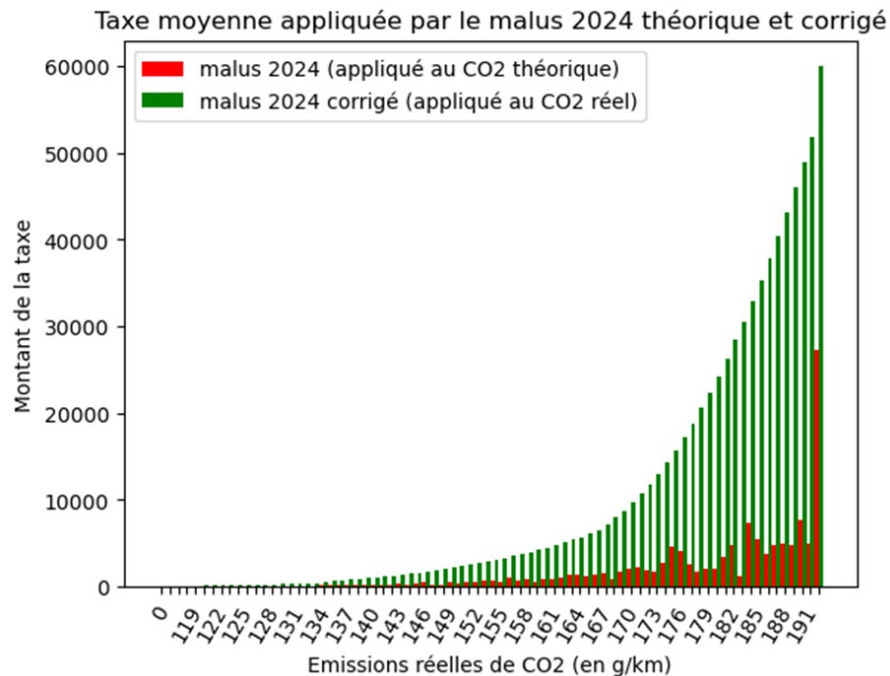
FIGURE 12 – Graphes de l'ajustement des parts de marché avec les malus 2021 et 2024 basés sur les émissions théoriques (rouge) et réelles corrigées (vert). Les catégories les moins émettrices ont systématiquement une part de marché plus grande après application des malus sur le CO2 réel corrigé.

Au vu du succès de nos précédents résultats, nous avons souhaité mesurer à quel point les modèles de véhicules étaient sous-taxés par rapport à leurs émissions réelles corrigées. Les deux graphes ci-dessous présentent la moyenne du malus appliqué aux véhicules par catégories de CO2 réel. En vert, le malus est appliqué aux émissions réelles corrigées, en rouge, aux émissions théoriques. Ces dernières considérations sont très importantes,

puisque la différence entre les bâtons vert et rouge est le manque réel de taxe induit par l'utilisation des émissions théoriques plutôt que réelles corrigées.



(a) Application au malus 2021



(b) Application au malus 2024

FIGURE 13 – Graphes des malus appliqués aux modèles en fonction de leurs émissions réelles avec les malus 2021 et 2024 basés sur les émissions théoriques (rouge) et réelles corrigées (vert)

Ces résultats sur les émissions à l'échelle du pays sont très satisfaisants. En effet, le calcul plus juste des émissions de chaque modèle permet, comme nous voulions le prouver, de réduire plus rapidement et conséquemment les émissions moyennes - et donc totales. Le but premier de notre projet est donc atteint.

## 5

# RÉDUIRE ENCORE LES ÉMISSIONS PAR LA CRÉATION D'UNE GRILLE DE TAXATION OPTIMISÉE

Nous avons vu qu'améliorer l'estimation des émissions de CO2 permettait d'avoir un effet positif sur l'impact de la taxe carbone de l'État en réduisant ces émissions. Cependant, nous avons toujours conservé la grille de taxation de l'État. Cette partie, qui se veut davantage un prolongement à notre étude se fixe pour objectif d'étudier l'effet d'une modification de cette grille de taxation sur les émissions, et de proposer différentes nouvelles pistes pour optimiser la taxe carbone.

### 5.1 APPROCHE GÉNÉRALE D'OPTIMISATION DU MONTANT DE LA TAXE

L'objectif ici est de proposer un modèle d'optimisation de la taxe qui, tout en respectant des impératifs de diminution des émissions, cherche tout de même à maximiser l'utilité des consommateurs. En effet, il serait simple de taxer très durement tous les véhicules - même légèrement polluants - pour forcer le consommateur à se déporter sur les véhicules les plus propres. Cette approche est contestable car se ferait au prix d'une grande insatisfaction du consommateur.

L'approche que nous retiendrons donc sera la suivante : nous chercherons à maximiser l'utilité générale, sous contrainte d'être en moyenne sous un certain seuil d'émissions de CO2.

#### 5.1.1 • MODÈLE D'OPTIMISATION DU MALUS

Nous utilisons ici le modèle logit, qui permet de mesurer les parts de marchés et leurs variations suivant le prix (et donc les différents malus qu'on y appliquera).

Pour chacune des  $N$  catégories de taxation (que l'on précisera par la suite), on a donc suivant le modèle logit une utilité de :

$$V_i(\mathbf{p}, \mathbf{m}) = -\exp(\delta_i + \beta(p_i + m_i)) \quad (8)$$

D'après la section 2.2 sur la *Random utility framework* du livre *Structural Econometric Modeling in Industrial Organization and Quantitative Marketing* (HORTACSU et JOONHWI, 2023), on écrit l'utilité de la population de manière globale comme :

$$V(\mathbf{p}, \mathbf{m}) = -\sum_i \exp(\delta + \beta(p_i + m_i)) \quad (9)$$

Grâce au modèle logit, on sait que les émissions de CO2 sont :  $\sum_i CO2_i \times \frac{\exp(\delta_i + \beta(p_i + m_i))}{\sum_j \exp(\delta_j + \beta(p_j + m_j))}$

On cherche alors à maximiser sous contrainte :

$$\max_{\{m_i\}_{i=1}^N} V(\mathbf{p}, \mathbf{m}) \quad \text{s.c.} \quad \sum_j CO2_j \times \exp(\delta + \beta(p_j + m_j)) \leq \text{seuil} \quad (10)$$

Il reste, et ce point est crucial, à déterminer le découpage des véhicules en catégories pour pouvoir procéder à l'optimisation de  $\mathbf{m}$ .

### 5.1.2 • CATÉGORISATION DES VÉHICULES

En raison de contraintes de puissance de calcul, on ne peut procéder à la meilleure solution, qui serait de créer et d'optimiser un malus pour chaque type de véhicule. On doit, à la manière de l'État, créer des catégories de taxation.

Après avoir expérimenté des approches en découpant en tranches de part de marché de taille égale (après avoir rangé les véhicules par CO2 émis croissant) ou en tranche de CO2 de taille constante (qui présentait le désavantage d'avoir des catégories presque vides et d'autres très pleines), nous avons décidé d'opter pour un découpage hybride. Le découpage se fait par tranche de CO2 mais de tailles variables, ce qui permet d'assurer une répartition relativement homogène des véhicules entre les catégories. Ce découpage s'appuie sur celui de l'État pour le malus 2024, qui présente une relative homogénéité des parts de marché. Nous avons cependant rajouté des catégories de malus en-dessous et au-dessus de l'intervalle de taxation de l'État, notamment pour s'adapter aux nouvelles valeurs du CO2 réel qui sont largement supérieures à celles attendues par l'État.

Nous ajoutons d'abord trois catégories supplémentaires pour les véhicules à basse émissions de CO2. Ce sont les véhicules dont les émissions réelles appartiennent respectivement aux intervalles  $[0,60]$ ,  $[60,90]$ , et  $[90, 117]$ . Nous avons ensuite gardé la catégorisation du malus 2024 qui associe à chaque unité d'émission de CO2 entre 117 g/km et 193 g/km une catégorie de véhicules. Nous avons ensuite rajouté 9 catégories supplémentaires pour les émissions de CO2 plus importantes appartenant respectivement aux intervalles  $[194, 197]$ ,  $[197, 200]$ ,  $[200, 203]$ ,  $[203, 206]$ ,  $[206, 209]$ ,  $[209, 212]$ ,  $[212, 215]$ ,  $[215, 218]$ . Nous avons ensuite mis dans une même catégorie les véhicules dont les émissions réelles de CO2 étaient supérieures à 218 g/km.

À noter que cette catégorisation avec des parts de marché relativement constantes ne l'est plus si on trace les parts de marché sans aucun malus, qui sert de base à notre optimisation. (On a en effet une énorme accumulation dans les catégories supérieures qui ne sont plus taxées, avec près de 80% des véhicules dans la catégorie maximale)

### 5.1.3 • LES PARTIS PRIS VIS-À-VIS DES RÉSULTATS EXPÉRIMENTAUX

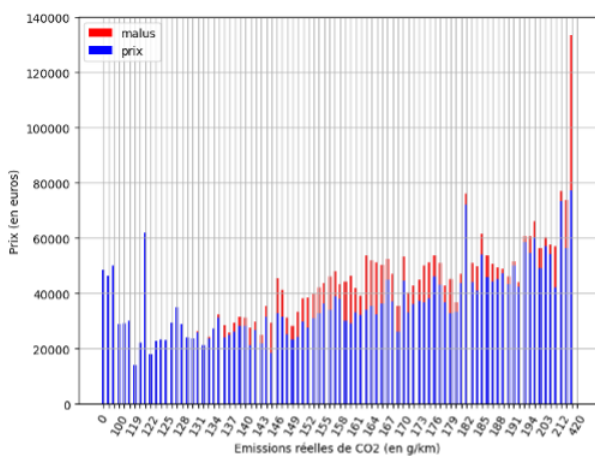
Au cours de l'élaboration de l'optimisation du malus, plusieurs partis ont été pris.

- Présence d'un maximum de taxation : Quels que soient les seuils de CO2 utilisés, il a été décidé de fixer un plafond de taxation à 120000€. En effet, l'État augmente à chaque nouveau malus ce plafond, qui est passé en 2021 de 30000€ à 60000€ en 2024, et cette nouvelle borne est un bon compromis entre progression de la taxe et prix relativement mesuré.
- Choix du seuil de CO2 moyen à 130g : en l'absence d'un parc automobile proposant suffisamment d'alternatives bas-carbone, fixer le seuil à 130g de CO2, soit 3,8g de baisse par rapport à nos estimations de 2024 semble être un objectif ambitieux et cohérent par rapport aux précédentes baisses de CO2.
- Non nécessaire croissance de  $\mathbf{m}$  : en raison de certaines valeurs de  $\delta$  importantes, l'optimisation donnait à certains endroits une décroissance du malus pour des émissions de CO2 plus grandes. Bien que nous discuterons de cette conséquence dans la suite, il a été décidé, à part dans un cas, de ne pas imposer une croissance du malus à tout prix.

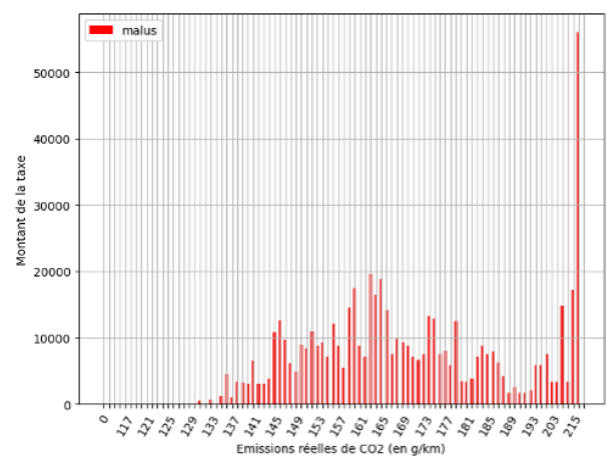
## 5.2 PRÉSENTATION ET DISCUSSION AUTOUR DE LA CONSTRUCTION DE TAXATIONS ALTERNATIVES

### 5.2.1 • APPROCHE BRUTE

Pour un seuil de 130g, et sans imposer de contrainte sur la croissance, on trouve un malus par catégorie suivant :



(a) prix+malus



(b) malus suivant les catégories de taxation

FIGURE 14 – Résultats du malus optimisé suivant les catégories de taxation.

Plusieurs résultats sont notables :

1. La taxation pour la plus haute catégorie de CO2 est proche de 60000€, ce qui est cohérent avec le seuil maximal de l'État.
2. On observe que le malus est très fluctuant, à l'exception de la dernière catégorie qui est fortement taxée car doit être évitée à tout prix. Nous notons ainsi que la taxe n'est pas croissante avec le niveau d'émission des véhicules. Cette fluctuation est due aux prix différents suivant les catégories, mais surtout aux  $\delta$ . En effet, ce  $\delta$  indique les préférences intrinsèques des consommateurs, et suffit donc selon le modèle logit à décourager les consommateurs de se reporter sur un modèle même en l'absence de malus. Bien que cela soit valide suivant le modèle logit, on obtient cependant un malus non croissant et peu parlant.
3. On note enfin que le montant de la taxe reste mesuré par rapport aux prix des véhicules, ce qui indique que le malus est raisonnable et que le seuil de 130g de CO2 est un seuil raisonnablement atteignable.

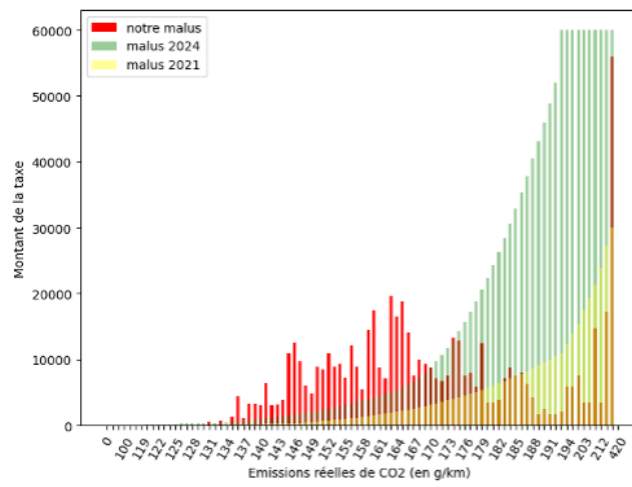


FIGURE 15 – Superposition des différents malus

Le graphe présenté en figure 15, superposant les différents malus, nous montre que l'optimisation, bien que moins régulière que les malus de 2021 et 2024, présente cependant des avantages et des résultats intéressants. Notamment, on observe que des montants bien moins élevés que ceux actuels sur les véhicules à fortes émissions sont nécessaires.

L'essentiel, et c'est le deuxième apprentissage de ce malus, est de davantage taxer les véhicules intermédiaires, qui restent relativement épargnés par le malus 2024. En effet, c'est là que se concentre la demande et le seul moyen de faire passer les émissions moyennes en dessous du seuil de 130g. De manière générale, ce nouveau malus répartit de façon plus efficiente la taxation sur les véhicules.

Si l'on trace désormais l'évolution des parts de marché résultant de ce calcul, nous obtenons le graphe présenté figure 16.

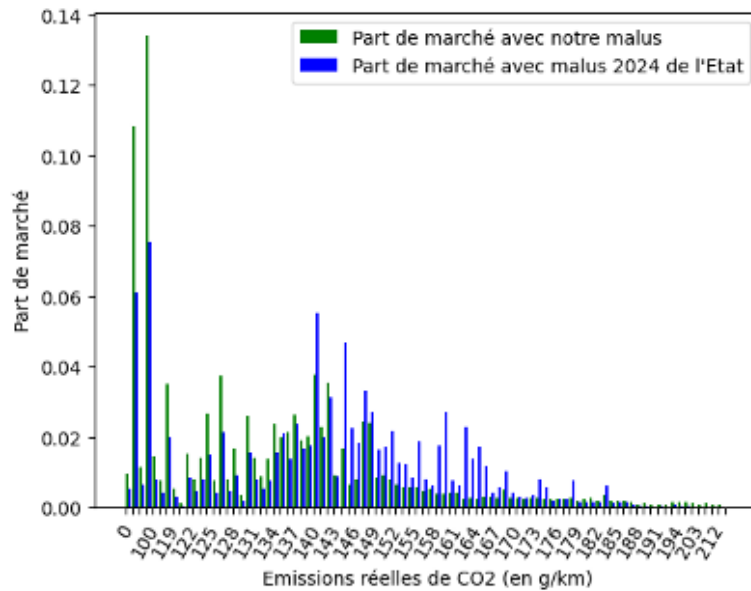


FIGURE 16 – Variation des parts de marché entre le malus 2024 et le malus optimisé.



À nouveau ce graphique montre plusieurs résultats instructifs :

1. Comme espéré, le malus optimisé décale bien les parts de marché vers la zone à faible émission. En particulier, la part de véhicules à très faibles émissions explose, avec en cumulé près de 11% de parts de marché supplémentaires dans les trois dernières catégories.
2. La zone entre 132g et 170g présente la majorité des véhicules avec le malus 2024. En effet, nous avons vu que la taxation était très faible dans cette zone ce qui entraîne cette accumulation. Le nouveau malus permet bien de résoudre ce problème avec des montants plus agressifs. Notons que la multiplication des catégories de taxation en-dessous de 117g de CO<sub>2</sub> permet de plus facilement décaler les parts de marché vers le bas avec le nouveau malus, ce qui plaide en faveur d'une extension des catégories de taxation.

Si ce malus optimisé permet de mettre en lumière plusieurs phénomènes intéressants, son application trop "mathématique" peut lui être reprochée. En effet, si le *delta* devrait - en théorie - décourager les consommateurs d'acheter des véhicules à fortes émissions dont le malus est pourtant faible, rien n'empêche des constructeurs de créer de nouveaux modèles dans ces catégories afin d'échapper aux taxes et de jouir d'un avantage concurrentiel.

Pour ces raisons, il apparaît comme nécessaire d'imposer une croissance au malus lors de l'optimisation, afin de le rendre plus cohérent et applicable au monde réel.

### 5.2.2 • CRÉATION D'UN MALUS CROISSANT OPTIMISÉ

En imposant une contrainte de croissance du malus lors de notre optimisation sous contrainte (tout en restant sous le seuil , on obtient le malus suivant :

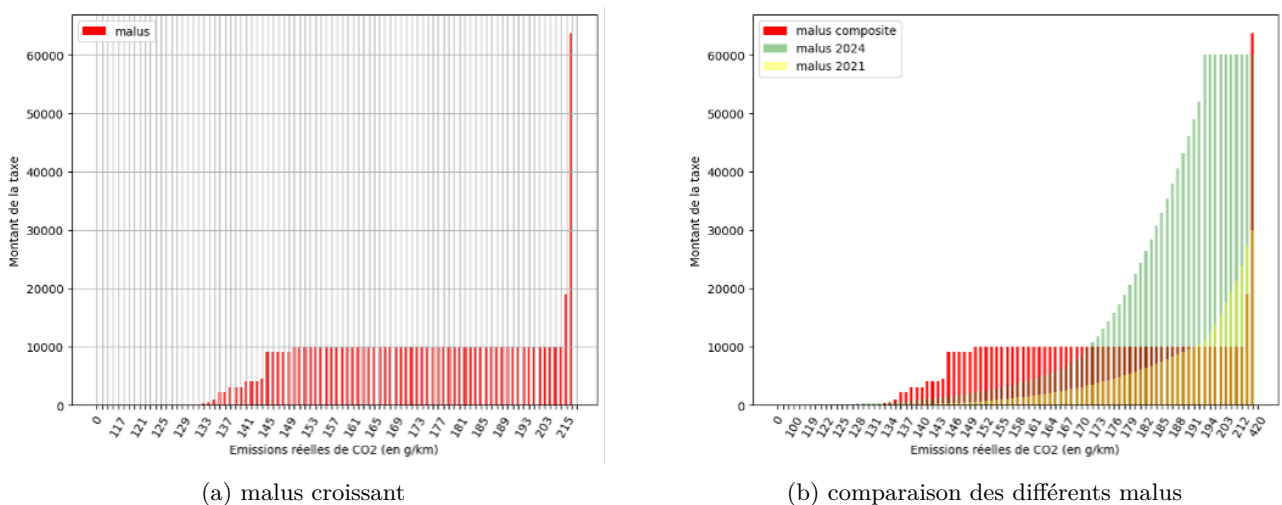


FIGURE 17 – Résultats du malus croissant optimisé suivant les catégories de taxation.

Ce malus, bien que surprenant, apporte un éclairage supplémentaire sur plusieurs points :

1. La taxation pour la plus haute catégorie de CO<sub>2</sub> n'atteint à nouveau pas le maximum fixé de 120000€ de malus, mais dépasse désormais le seuil de 60000€ du malus 2024. Ceci s'explique par la taxation supplémentaire des catégories précédentes, qui induisent donc davantage de pression sur cette catégorie (pour éviter d'y reverser certains véhicules).
2. On observe un plateau du malus de 145g à 210g de CO<sub>2</sub>. Cela s'explique par plusieurs éléments. D'abord le seuil est de 130g et tous les véhicules au-dessus sont logiquement pénalisés. Ensuite, comme vu dans le précédent malus, on a une pénalisation relativement forte sur les véhicules intermédiaires. La contrainte de croissance impose que les classes suivantes doivent se mettre à niveau sans que cela soit nécessaire.

3. La zone centrale devient alors une zone de transition entre les véhicules à fortes et faibles émissions. En définitive, ce malus quasi-constant au centre sert seulement à faire transitionner le consommateur et à ne pas le faire tomber dans un "puits" de consommation dans cette zone.

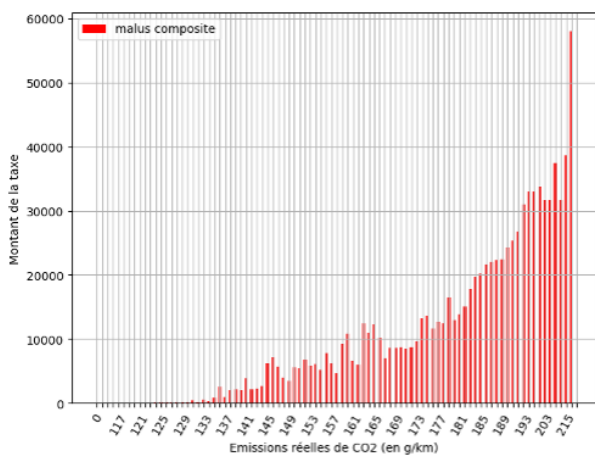
À noter que les parts de marché avec ce malus sont très proches des parts de marché du malus précédent non croissant.

### 5.2.3 • LE MALUS COMPOSITE

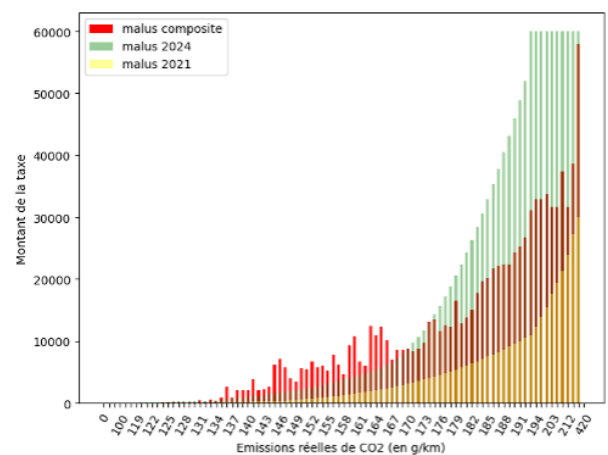
Le malus croissant devenu presque constant nous conduit à abandonner l'idée d'imposer une croissance stricte du malus. Cependant, comme vu dans le malus optimisé non croissant, de trop grandes variations du malus risque d'encourager les consommateurs et les producteurs à contourner le malus, et donc à le rendre contre-productif.

L'idée est alors de créer un malus composite entre le malus 2024 et le malus optimisé non croissant, afin de conserver les avantages de chacun : la croissance et la régularité du malus de l'État, et les taxations optimisées sur la zone intermédiaire et les montants réduits de notre malus optimisé.

On trace le malus composite qu'on compare aux malus de 2021 et 2024 :



(a) malus croissant



(b) comparaison des différents malus

FIGURE 18 – Résultats du malus composite

Ce dernier malus est le plus intéressant et permet de se rapprocher peu à peu d'une taxe applicable, bien qu'il ne soit pas encore entièrement satisfaisant :

1. On remarque d'abord qu'on a bien une quasi-croissance, et une fonction presque convexe, ce qui permet de limiter les dangers d'adaptation par les constructeurs.
2. Les montants de la taxe sont de manière générale plus faibles que ceux du malus 2024, avec cependant bien une zone de taxation pour les véhicules intermédiaires, qui va permettre de décaler les parts de marché plus efficacement.
3. Les variations de parts de marché sont très proches de celles du malus optimisé sans croissance. Cela a pour conséquence que sans maîtriser le CO<sub>2</sub> émis en fusionnant les deux malus, on aboutit à un CO<sub>2</sub> émis de 130,91g, soit moins d'un gramme de différence avec le malus optimisé. Ce résultat est fort car il montre bien l'importance de taxer d'avantage les véhicules les plus vendus, qui ne sont pas nécessairement les plus pollués mais qui sont responsables de la majorité des émissions.

4. Nous notons toutefois que la stricte croissance du malus n'est pas encore garantie et constitue une piste d'amélioration à creuser : si ceci s'explique en raison de l'inutilité de taxer davantage certaines catégories, une question éthique se pose, car il semblerait cohérent d'observer une croissance de la taxe avec les émissions dont est responsable le véhicule.

Ce malus permet ainsi de s'approcher d'un compromis entre une réduction efficace des émissions de CO<sub>2</sub>, et quasi applicabilité dans le monde réel.

En conclusion de cette partie, nous avons progressivement construit un malus en résolvant un problème d'optimisation sous contrainte. Si le malus final appelle encore des améliorations, nous aurons toutefois soulevé des points intéressants, tels que la nécessité de taxer davantage des véhicules aux émissions dans la tranche moyenne, car les parts de marché actuelles s'y concentrent massivement, ou la possibilité de moins taxer certaines classes de véhicules tout en y gagnant en terme d'émissions.

## 6

# CONCLUSION

### 6.1 BILAN DE NOTRE TRAVAIL - RETOUR SUR NOS OBJECTIFS

Notre PSC s'est proposé d'étudier et d'optimiser la politique de malus écologique sur les véhicules privés.

Il s'est structuré en quatre angles d'attaque principaux, chacun nourrissant les autres. Ces quatre facettes suivent les étapes de l'élaboration de cette taxe : une première partie explore les données que nous nous sommes procurés, pour en étudier la structure, les catégoriser et en extraire des tendances. Une deuxième partie, après avoir vu les carences de l'estimation du CO<sub>2</sub> par une simple mesure en laboratoire ou par des tests initiaux sur route, cherche à construire un nouvel estimateur du CO<sub>2</sub> en utilisant le Machine Learning à partir de paramètres intrinsèques au véhicule. Une fois cet estimateur construit, et le malus déduit à partir de celui-ci, une troisième partie cherche à évaluer l'impact de ce malus sur les ventes de véhicules et sur les émissions résultantes. Une dernière partie, encore expérimentale, vise à optimiser notre grille de taxation afin de réduire davantage les émissions.

1. Dans la première partie, nous avons pu mener une étude approfondie des données et comprendre précisément les enjeux et les politiques mises en place. Ceci est notamment passé par la compréhension d'une documentation riche, et par la recherche de relations entre les données.
2. Dans la seconde partie, nous avons d'abord cherché à formater les données afin de les rendre exploitables. Puis, après plusieurs tris dessus, nous avons entraîné un premier modèle de régression linéaire dessus. Les résultats ont été très satisfaisants, avec des performances jusqu'à trois fois supérieures à celles de l'État. Nous avons ensuite utilisé la régression Lasso pour simplifier le modèle et le rendre plus facilement implémentable sans perdre en performance. Nous avons également employé le modèle de régression KNN plus puissant pour encore gagner en performance.
3. Dans la troisième partie, nous nous sommes appuyé sur le modèle logit pour modéliser la demande de véhicules en fonction du prix des voitures. Nous avons ainsi observé les effets du malus sur la demande et par voies de conséquences, nous avons pu analyser les émissions moyennes résultant des choix des consommateurs. Ceci nous a permis de souligner l'impact positif de l'estimation initiale des émissions réelles que nous avons proposé par machine learning.
4. Enfin, bien qu'ayant atteint notre objectif initial de réduction des émissions observées par un calcul corrigé des émissions réelles, nous avons cherché à aller plus loin encore en proposant une nouvelle grille de taxation. L'enjeu a été de proposer une grille que notre modèle détermine directement en fonction du

seuil d'émissions moyennes que l'on souhaite fixer. Cette approche a ouvert de nouvelles perspectives et nous invite à poursuivre en explorant certains prolongements possibles.

Chacune des différentes parties a cherché à enrichir un projet dans sa globalité : celui de mettre en place un malus écologique plus juste, reflétant mieux les émissions de dioxyde de carbone des véhicules en circulation et contribuant à leur réduction.

## 6.2 PROLONGEMENTS

A la fin de cette année de travail autour du projet, nous identifions plusieurs points de prolongements autour des résultats émis.

- Premièrement, il serait utile, dans la partie d'optimisation du malus, de s'atteler à une taxation plus fine (notamment modèle par modèle) afin de parvenir à taxer parfaitement chaque véhicule. Outre l'amélioration logique des résultats de la taxation, cela permettrait surtout de réduire la dépendance au  $\delta$  et aux différents prix dans les catégories qui nous a pénalisés dans l'élaboration du malus croissant. Une catégorisation plus fine permettrait potentiellement de davantage se rapprocher d'un malus croissant et convexe, en même temps qu'il serait plus juste et plus précis.
- Dans le même esprit, au vu de la forme des taxations de l'État, il serait certainement intéressant d'étudier une taxation de type  $\lambda \exp(\alpha \times \text{CO}_2)$ . C'est le sens de notre proposition de malus "composite", moyenne entre le malus issu de l'optimisation et le malus de l'État, qui conserve la croissance exponentielle du malus de l'État tout en incorporant les progrès apportés par la partie optimisation de notre projet.
- Enfin, un prolongement possible pourrait être d'étudier les effets du second ordre, ou indirects, de l'implantation de notre malus. Comment réagissent les producteurs de véhicules face à cette nouvelle taxation ? Comment est, globalement et marginalement, impacté le bien-être du consommateur ? De telles questions pourraient être à se poser dans une étude complémentaire.

## 6.3 ORGANISATION DU TRAVAIL

Nous avons organisé notre travail comme suit : le projet initial était d'exploiter les bases de données de l'ADEME et du Répertoire statistique des véhicules routiers afin de cerner les voies d'amélioration du calcul des émissions de  $\text{CO}_2$  de chaque véhicule, puis de créer un outil de calcul de ces émissions qui se rapproche mieux des émissions réelles. Nous avons donc consacré la première phase de notre travail, de novembre à janvier, à ce calcul des émissions. Bien que nous travaillions collectivement, nous nous sommes assignés des sujets, sur lesquels nous avions la main. Maria était en charge de l'analyse de la littérature, Capucine était plus amplement à la tête de l'analyse initiale des bases, tandis qu'Antoine et Eva avaient le leadership sur la partie technique de manipulation des données et du machine learning. Nous espérions également étudier l'effet de cette amélioration, alors seulement hypothétique, sur la demande de véhicules et donc sur le calcul des émissions. Etienne et Maria se sont donc attelés rapidement à la prise en main des modèles de demande.

Une fois nos premiers résultats d'estimation de  $\text{CO}_2$  obtenus, à la mi-janvier, Capucine et Etienne ont poursuivi dans l'exploitation des résultats obtenus par machine learning pour étudier les effets sur la demande, tandis qu'Antoine et Eva ont commencé en parallèle un travail final d'optimisation de la grille de taxation, avant d'être rejoints par le reste de l'équipe. Ce travail n'était initialement pas prévu, mais nous est apparu comme l'aboutissement de notre démarche d'amélioration du malus écologique. Cette dernière partie nous a donné beaucoup de fil à retordre, et nous a occupé les mois de mars et avril.



Cette répartition de principe que nous avons établi pour structurer notre travail ne doit toutefois pas dissimuler la dimension collective de notre travail, notamment aux travers de fréquents points de situation collectifs.

## 7 APPENDICE

### PARAMÈTRES PRÉSENTS DANS LES BASES INITIALES

Grandeur renseignée dans la Base de l'ADEME	Présence dans la base RSVERO	Présence dans la base ADEME
type CNIT	X	X
Type énergétique	X	X
Marque	X	X
Modèle	X	X
Nombre places assises	X	
Puissance fiscale	X	X
Date minimum de mise en circulation	X	
Date maximum de mise en circulation	X	
Cylindrée	X	X
CO2 théorique	X	
CO2 réel	X	
CO2 PRF		X
Nombre de décibels	X	
PTR (poids total roulant)	X	
PTAC (poids total autorisé en charge)	X	
Poids à vide		X
Charge utile	X	
Puissance (kW)	X	X
Rapport poids puissance		X
Volume des ventes	X	X
Genre		X
Type		X
Version		X
Gamme		X
Groupe		X
Type de boîte		X
Carrosserie		X
Consommation extra-urbaine		X
Consommation mixte		X
Consommation urbaine		X
Bonus/Malus		X
Prix du véhicule		X

TABLE 8 – Liste des paramètres présents dans les bases brutes

### FICHIERS DE CODE

Lien du GitHub sur lequel se trouvent nos fichiers informatiques et lignes de code : <https://github.com/PSCfin/Fichiers-informatiques-PSCECO05>

## RÉFÉRENCES

---

- D'HAULTFOEUILLE, X., GIVORD, P., & BOUTIN, X. (2014). The environmental effect of green taxation : the case of the French bonus/malus. *The Economic Journal*, 124(578), F444-F480.
- DURRMEYER, I., GUILLOUZOUIC, A., MALGOUYRES, C., MAYER, T., & TÔ, M. (2024). Évaluation des mesures de soutien aux véhicules propres.
- HORTACSU, A., & JOONHWI, J. (2023). *Structural Econometric Modeling in Industrial Organization and Quantitative Marketing*. Princeton University Press.
- KESSLER, L., MORVILLIER, F., PERRIER, Q., & RUCHETON, K. (2023). An ex-ante evaluation of the French car feebate, *Energy Policy*. 173.
- MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE ET SOLIDAIRE. (2015). <https://www.ecologie.gouv.fr/sites/default/files/SNBC%20r%C3%A9sum%C3%A9%20EN.pdf>
- TRAIN, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.
- UWE, T., SONSOLES, D., PETER, M., ANUP, B., JAN, D., & LIGTERINK, N. (2019). *From laboratory to road, a 2018 update of official and "real-world" fuel consumption and CO2 values for passenger cars in Europe*. International Council on Clean Energy Transportation.
- YAN, S., & GUNNAR S., E. (2018). Greening the vehicle fleet : Norway's CO<sub>2</sub>-Differentiated registration tax.