Jean-Baptiste Quéméneur
Romain Vilaça

Projet CO_2

I. Introduction

Relation entre le CO_2 et le Climat

Le système climatique de la Terre évolue depuis plusieurs millions d'années. Les preuves tirées des archives naturelles offrent la possibilité d'étudier ce climat à long terme, mais également de réaliser des perspectives sur les changements observés et les changements prévus pour les décennies et siècles à venir. Ces reconstructions du climat montrent notamment la présence d'un lien fort entre les concentrations de CO_2 dans l'atmosphère et la température à la surface de la planète (IPCC et al., 2021). Or, depuis la révolution industrielle, les émissions de gaz à effet de serre et particulièrement de CO_2 n'ont cessé d'augmenter et ont provoqué une accélération du réchauffement climatique (U.S. Global Change Research Program, 2009). À ce jour, la température de surface moyenne a déjà augmenté de 1,1°C par rapport aux températures observées entre 1850 et 1900. Les modèles proposés par le GIEC, le groupe d'experts intergouvernemental sur l'évolution du climat, indique que la température terrestre va continuer d'augmenter dans les années à venir atteignant au minimum une augmentation de +1.5°C d'ici quelques décennies et une augmentation de plusieurs degrés Celsius à long terme (figure 1, (IPCC et al., 2021)) .

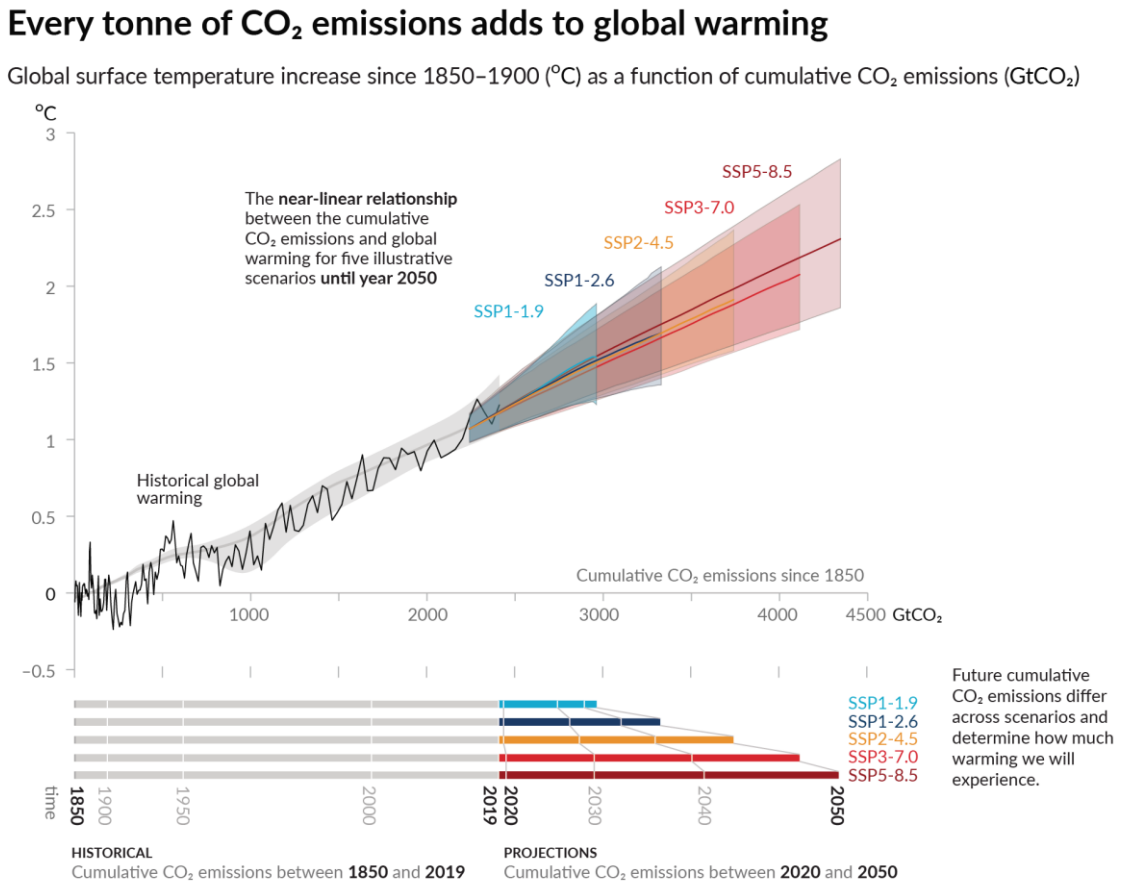


Figure 1 : Données historiques (ligne noire fine) représentant la température observée à la surface du globe en °C depuis 1850-1900 en fonction des émissions cumulées de dioxyde de carbone (CO_2) de 1850 à 2019. *La fourchette grise avec sa ligne centrale montre une estimation correspondante du réchauffement historique de la surface causé par l'homme. Les zones colorées indiquent la fourchette réaliste des projections de la température à la surface du globe, et les lignes centrales de couleur épaisse indiquent l'estimation médiane en fonction des émissions cumulées de CO_2 de 2020 à 2050 pour l'ensemble des scénarios illustratifs (SSP1-1.9, SSP1-2.6, SSP2-4.5, SSP3-7.0, et SSP5-8.5). Les projections utilisent les émissions cumulées de CO_2 de chaque scénario respectif, et le réchauffement planétaire projeté inclut la contribution de tous les forçages anthropiques. Source IPCC et al., 2021.*

Conséquences de la Relation entre le CO_2 et le Climat

Lors de son dernier rapport, publié en mars 2023, le GIEC a conclu que les changements du système climatique, y compris les changements dans les climats régionaux et extrêmes, sont attribuables à l'activité humaine (Mukherji et al., 2023). Les émissions anthropiques de CO_2 observées et prévues dans les années à venir ne sont donc pas anodines. Leurs conséquences sont multiples, à la fois sur la faune et la flore, notamment marines, mais également sur les sociétés humaines au global en impactant directement ou indirectement les récoltes, la santé et même les infrastructures (figure 2, (Mukherji et al., 2023)).

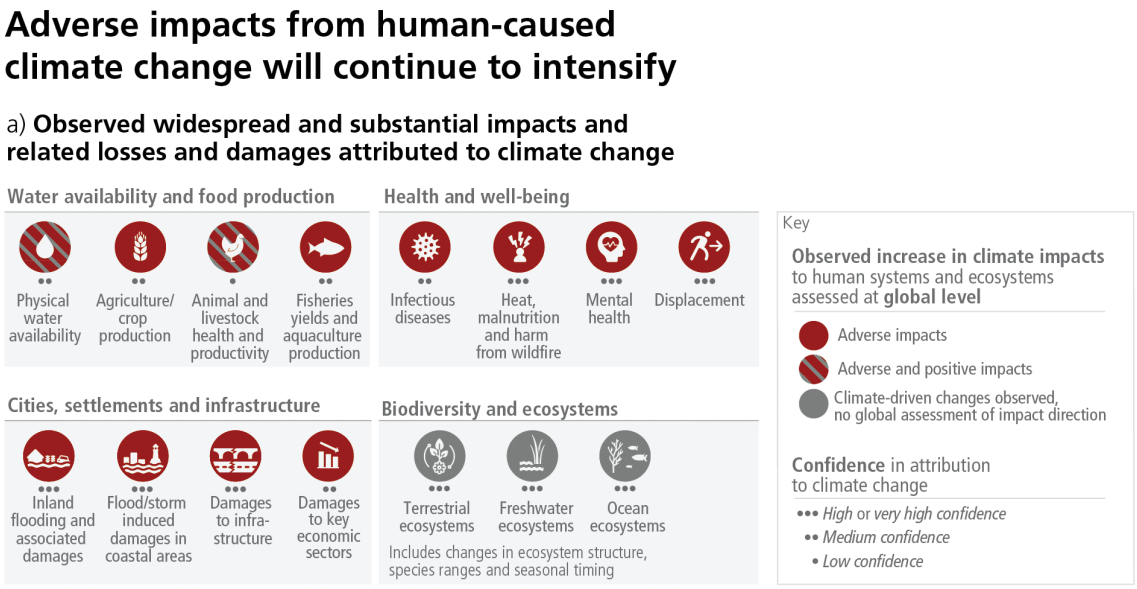


Figure 2: Présentation des répercussions et des dommages associés aux changements climatiques sur les systèmes humains.

La disponibilité physique de l'eau comprend l'équilibre de l'eau disponible à partir de différentes sources, y compris les eaux souterraines, la qualité de l'eau et la demande en eau. Les évaluations de la santé mentale et des déplacements à l'échelle mondiale ne tiennent compte que des régions évaluées. Les niveaux de confiance reflètent l'évaluation de l'attribution de l'impact observé au changement climatique. Source Mukherji et al., 2023.

Concernant l'Europe, l'Agence européenne pour l'environnement a annoncé en 2019 qu'un quart des émissions totales de CO_2 de l'Union Européenne était associé aux transports. La majeure partie de ces émissions seraient associées aux véhicules routiers motorisés, faisant ainsi du transport routier l'un des secteurs économiques responsable de fortes émissions de CO_2 (figure 3, (European Parliament, 2019)). Face à ces problèmes, les gouvernements et hautes instances gouvernementales ont pris divers engagements dans les années à venir. À titre d'exemple, l'UE vise à réduire de 90 % les émissions de gaz à effet de serre provenant des transports d'ici à 2050, par rapport à 1990. Cet objectif s'inscrit dans le cadre de ses efforts visant à atteindre la neutralité carbone d'ici 2050, conformément à la feuille de route européenne du Green Deal (European Parliament, 2019).

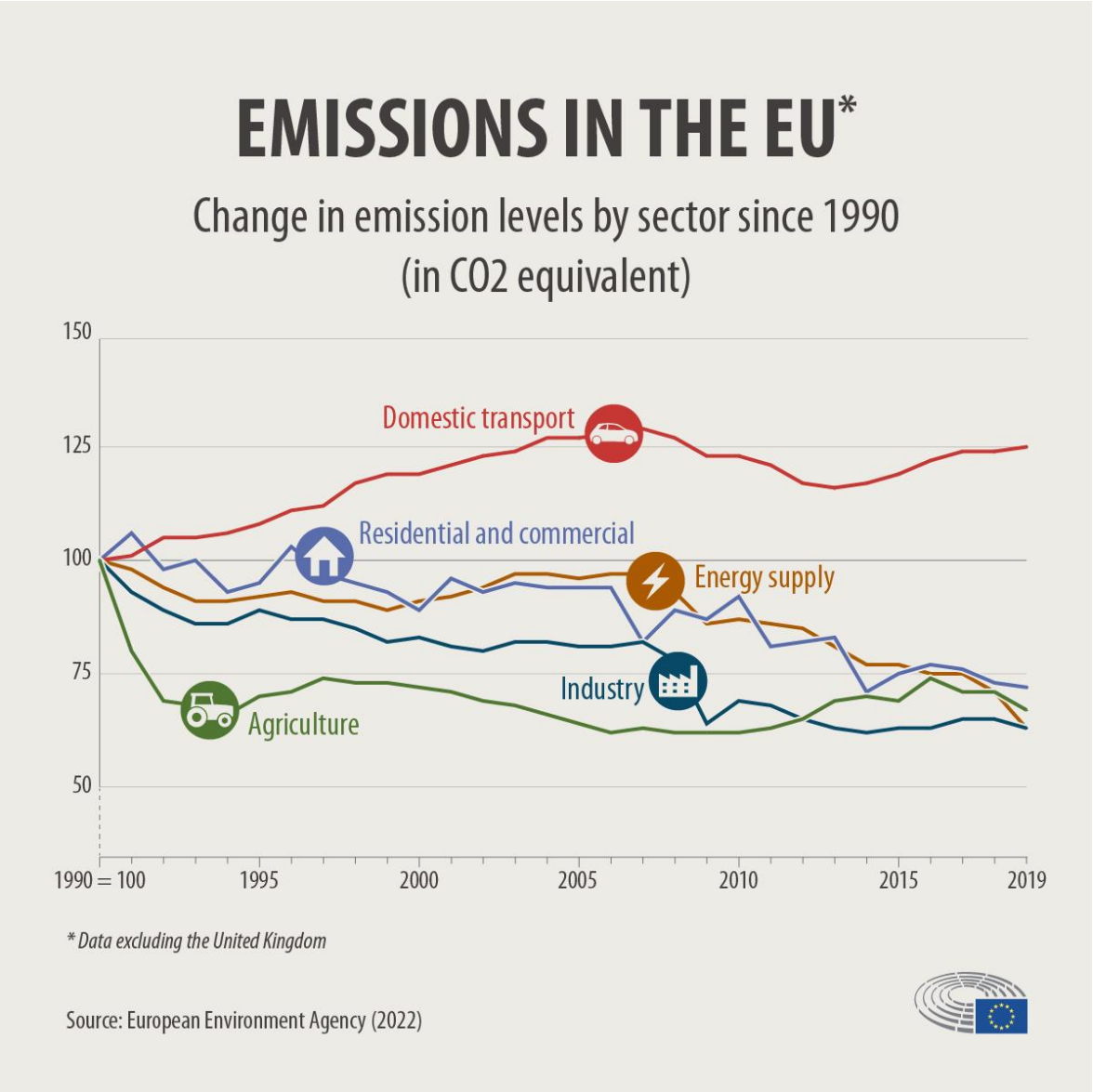


Figure 3: évolution du niveau d'émission par secteur depuis 1990 (en équivalent CO2)

Pour atteindre la neutralité carbone, l'UE se doit de diminuer ses émissions de gaz à effet de serre associés aux transports. En effet, on remarque qu'au fil des années tous les secteurs excepté celui des transports ont réussi à diminuer leur empreinte carbone (figure 3). Afin d'essayer de comprendre ce qu'il s'est passé et d'envisager l'avenir, nous allons au cours de ce rapport analyser les données fournies par le gouvernement Français sur le parc automobile français entre 2005 et 2015. Nous allons essayer de vérifier au fil des ans si les constructeurs ont diminué la consommation de leurs véhicules, mais également s'il y a eu une prise de conscience des consommateurs. Enfin, à partir des données récoltées, nous allons essayer de proposer un modèle afin d'envisager les véhicules à promouvoir dans les années à venir pour atteindre la neutralité carbone d'ici 2050.

II. Mise en forme du jeu de données

Base de données et création du dataframe

L'ADEME, l'Agence de l'environnement et de la maîtrise de l'énergie, acquiert depuis 2001, à raison d'une fois par an, les données correspondant aux émissions de CO_2 et de polluants des véhicules commercialisés en France. Ces informations sont transmises par l'Union Technique de l'Automobile du motocycle et du Cycle, UTAC (en charge de l'homologation des véhicules avant leur mise en vente), et sont en accord avec le ministère du développement durable. À noter que la dernière mise à jour des données présentes sur le site du gouvernement remonte au 15 octobre 2015.

Les informations disponibles concernant ces données sont les suivantes (Ademe):

Pour chaque véhicule les données d'origine (transmises par l'Utac) sont :

- les consommations de carburant
- les émissions de dioxyde de carbone (CO_2)
- les émissions des polluants de l'air (réglementés dans le cadre de la norme Euro)
- l'ensemble des caractéristiques techniques des véhicules (gammes, marques, modèles, n° de CNIT, type d'énergie ...)

Sur le site Carlabelling (http://carlabelling.ademe.fr), l'ADEME complète ces données avec les informations suivantes :

- les valeurs du bonus-malus et de l'étiquette Classe Energie - CO_2
- les résultats d'expertises tels le coût annuel de la consommation de carburant sur 15 000 km.
  Elle établit également des classements pour distinguer les véhicules « les plus propres en CO_2 et les plus économes en énergie ».

Pour répondre à nos questions, nous avons choisi de suivre l'évolution des données sur 10 ans et avons récupéré les dataframes de 2005 à 2015. Après visualisation des données, une première étape de mise en forme et d'harmonisation était nécessaire. Cette étape a été réalisée sur Excel et est présentée ci-dessous. Pour information, un aperçu des différentes variables brutes par années est disponible en Annexe (Annexe 1).

Nettoyage et harmonisation des différents dataframes avant concaténation

a. Suppression des variables inutiles à l'analyse

Après avoir exploré les différents jeux de données, on constate que certaines variables peuvent être directement écartées car redondantes ou inutiles pour les futures analyses.

- Variable " type 2"

Cette variable n'est présente que sur le jeu de données de 2008 et correspond aux 8 premiers numéros du code "CNIT". Nous choisissons donc de la supprimer du jeu de données final.

- Variable “TVV”

La variable “TVV” ou “Type Variante Version” est présente sur les jeux de données allant de 2011 à 2015. Elle correspond à un code propre au véhicule que l’on retrouve sur la carte grise. Ce code est directement lié au code “CNIT”. Il n’est donc pas nécessaire de garder ces deux variables et nous décidons de supprimer la variable “TVV” et de conserver “CNIT” qui est déjà présent dans tous les jeux de données.

- Variable “Date_maj”

Cette variable est présente sur les jeux de données allant de 2011 à 2015. Elle correspond à la date à laquelle les données ont été mises à jour. Nous avons choisi de remplacer cette variable par la variable “annee_df” qui indique seulement l’année du jeu de données.

b. Ajout de variable dans certains dataframes

Après avoir exploré les différents jeux de données, on constate que certaines variables sont présentes seulement sur certains jeux de données. En fonction de leurs pertinences, nous avons fait le choix d’ajouter certaines de ces variables sur l’ensemble des jeux de données.

- Variable “annee_df”

Cette variable correspond à l’année d’acquisition du jeu de données par l’ADEME. Elle remplace la variable “Date_maj” présente sur les jeux de données allant de 2011 à 2015 en indiquant seulement l’année. Elle a également été ajoutée aux jeux de données allant de 2005 à 2010.

- Variable “hybride”

Cette variable apporte une information sur le type de véhicule : hybride, oui ou non. Elle est initialement présente dans les jeux de données allant de 2012 à 2015. Elle a été ajoutée aux autres jeux de données en vérifiant si l’intitulé “hybride” est présent dans la variable “lib_mod” qui correspond aux caractéristiques du véhicule.

- Variable “co_typ_1”

Cette variable apporte une information sur le niveau d’émissions de monoxyde de carbone (g/km) d’un véhicule. Elle est initialement présente dans les jeux de données allant de 2012 à 2015 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “hc”

Cette variable apporte une information sur le niveau d’émissions d’hydrocarbures imbrûlés (g/km) d’un véhicule. Elle est initialement présente dans les jeux de données allant de 2012 à 2015 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “nox”

Cette variable apporte une information sur le niveau d’émissions d'oxydes d'azote (g/km) d’un véhicule. Elle est initialement présente dans les jeux de données allant de 2012 à 2015 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “hcnox”

Cette variable apporte une information sur le niveau d’émissions d’hydrocarbures imbrûlés associé au niveau d’émissions d'oxydes d'azote (g/km) d’un véhicule. Elle est initialement présente dans les jeux de données allant de 2012 à 2015 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “ptcl”

Cette variable apporte une information sur le niveau d’émissions de particules (g/km) d’un véhicule. Elle est initialement présente dans les jeux de données allant de 2012 à 2015 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “masse_ordma_min”

Cette variable apporte une information sur la masse d'un véhicule non chargé (kg). Elle est initialement présente dans les jeux de données allant de 2012 à 2015 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “masse_ordma_max”

Cette variable apporte une information sur la masse d'un véhicule chargé (kg). Elle est initialement présente dans les jeux de données allant de 2012 à 2015 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “champ_v9”

Cette variable apporte une information sur le certificat d’immatriculation qui contient la norme EURO. Elle est initialement présente dans les jeux de données allant de 2011 à 2015 a été ajoutée aux autres jeux de données en remplissant les cellules par des NA. Nous souhaitons par la suite conserver uniquement la norme EURO dans le jeu de données.

- Variable “carrosserie”

Cette variable apporte une information sur la carrosserie d’un véhicule. Elle est initialement présente dans les jeux de données allant de 2012 à 2014 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “gamme”

Cette variable apporte une information sur la catégorie ou la gamme d’un véhicule. Elle est initialement présente dans les jeux de données allant de 2012 à 2014 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA.

- Variable “etiq_energ”

Cette variable apporte une information sur la catégorie énergétique d’un véhicule (A-G). Elle est initialement présente dans les jeux de données allant de 2007 à 2010 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA. Les informations manquantes ont été ajoutées par la suite en croisant les informations d’étiquetage énergétique transmises par le gouvernement (ARISTIZABAL (V) et al., 2019) avec le niveau d'émission de CO_2 d'un véhicule pour 100 kilomètres parcourus.

- Variable “bonus_malus”

Cette variable apporte une information sur le bonus ou le malus (€) proposé à l’achat d’un véhicule en fonction de sa catégorie énergétique. Elle est initialement présente dans les jeux de données allant de 2008 à 2010 et a été ajoutée aux autres jeux de données en remplissant les cellules par des NA. Les informations manquantes auraient pu être trouvées sur le site de Carlabelling, mais cela aurait demandé un travail trop conséquent et nous avons décidé de conserver uniquement les informations déjà à notre disposition.

c. Harmonisation des noms des variables

L’objectif est de réunir les jeux de données allant de 2005 à 2015 en un seul dataframe. Pour cela, une étape d’harmonisation du nom des variables est nécessaire. En effet, on remarque une grande hétérogénéité dans le nom des variables entre les différents dataframes (Tableau 1). Il est donc nécessaire de définir un nom commun pour chaque variable et de mettre à jour les différents jeux de données. Pour définir le nom à garder, nous nous sommes référés au dictionnaire des variables, disponible sur le portail du gouvernement (Annexe 1).

Le tableau 1 représente, pour chaque variable, l’intitulé final et la liste des différents intitulés que prenait cette variable dans les différents jeux de données. Se référer à l’annexe 2 pour visualiser les intitulés initiaux de chaque dataframe (Annexe 2).

d. Complément des données

Pour éviter certains problèmes par la suite, les lignes correspondantes aux véhicules électriques (cod cbr = EL) ont dû être implémentées de 0 pour les variables de consommations et d’émissions de gaz.

Tableau 1 : Harmonisation des noms de chaque variable

Intitulé final	Intitulés des variables dataset 2005-2015
lib_mrq_utac	Marque, MARQUE, lib_mrq, mrq_utac
lib_mod	MODELE VERSION, Modèle dossier, Modèle UTAC, Désignation commerciale, lib_mrq_doss, lib_mod_doss, lib_mod, dscom
cnit	CNIT, cnit
cod_cbr	cod_cbr, typ_cbr, energ, ENERGIE, Carburant, carburant
hybride	Hybride, hybride
puiss_admin_98	puiss_admin_98, Puissance administrative, puissance fiscale
puiss_max	puiss_max, Puissance maximale (kW), puissance reelle
typ_boite_nb_rapp	typ_boite_nb_rapp, bv, Boîte de vitesse
conso_urb	conso_urb, urb, Comsommation urbaine (l/100km)
conso_exurb	conso_exurb, ex-urb, Consommation extra-urbaine (l/100km)
conso_mixte	conso_mixte, mixte, Consommation mixte (l/100km)
co2	co2, CO2, CO2 (g/km)
co_typ_1	co_typ_1, CO type I (g/km)
hc	hc, HC (g/km)
nox	nox, NOX (g/km)
hcnox	hcnox, HC+NOX(g/km)
ptcl	ptcl, Particules (g/km)
masse_ordma_min	masse_ordma_min, masse vide euro min (kg),
masse_ordma_max	masse_ordma_max, masse vide euro max (kg),
champ_v9	champ_v9, Champ V9
carrosserie	Carrosserie
gamme	gamme
etiq_energ	ETIQUETTE ENERGIE
bonus_malus	BONUS(-)/MALUS(+)

La définition de chaque intitulé est présente dans le dictionnaire des variables (Annexe 1).

Nettoyage et harmonisation du dataframe concaténé

Suite à la réunification des différents dataframes en un seul jeu de données, plusieurs étapes d'harmonisation ont été réalisées.

a. Suppression des valeurs manquantes

Dans un premier temps, les valeurs manquantes (NA) des variables "cnit", "co2" et "conso_urb", "conso_exurb" et "conso_mixte" ont été supprimées pour différentes raisons. L'absence du numéro CNIT ne permet pas d'identifier et de référencer un véhicule. La variable "co2" correspond à notre variable d'intérêt et son absence ne nous apporte aucune information sur nos problématiques. Enfin, les variables "conso_urb", "conso_exurb" et "conso_mixte" sont directement liées à la variable "co2" et il est donc important qu'elles ne présentent pas de NA.

Concernant les autres variables, nous avons fait le choix de ne pas supprimer les lignes contenant des NA afin de ne pas perdre d'informations pour la visualisation des données. Ces données seront toutefois supprimées avant l'étape de modélisation.

b. Suppression des doublons

Une recherche de doublon a également été effectuée sur la variable « cnit ». En théorie, aucun doublon ne devrait être présent dans cette variable puisqu'elle correspond au numéro d'identification d'un véhicule. En pratique, plus de 130 000 doublons ont été trouvé soit quasiment la moitié des données brutes récoltées sur le site du gouvernement.

c. Vérification du type des variables

Suite à l'importation et à la concaténation des jeux de données, le type « object » a été attribué à la majorité des variables numériques. Une étape de transformation du type en "float" ou "int" a donc été nécessaire.

d. Harmonisation des variables catégorielles

La réunification des différents dataframe en un seul jeu de données a également créé plusieurs problèmes au sein des variables catégorielles. La nomenclature des différentes modalités au sein des variables n'était pas forcément la même au cours du temps et des erreurs de saisie étaient parfois présentes. Une étape d'harmonisation des modalités catégorielles a donc été nécessaire.

III. Visualisation des données

Suite au nettoyage et à l'harmonisation des données, nous avons pu investiguer sur nos problématiques à travers des analyses graphiques couplées à des analyses statistiques. On remarque notamment que le niveau d'émissions de CO_2 moyen augmente significativement et progressivement pour atteindre un plateau en 2007 (~ 230 g de CO_2 par km). Puis, à partir de 2011, une diminution progressive et significative est observée chaque année pour atteindre le minimum d'émissions moyen en 2015 (~ 155 g de CO_2 par km) (figure 4). Ainsi, il semblerait que les constructeurs ou les consommateurs ont commencé à s'intéresser à des voitures moins polluantes à partir de 2011. Cela semble notamment corrélé avec un durcissement du bonus/malus proposé par l'état Français à partir de 2011 (Robequain, 2010).

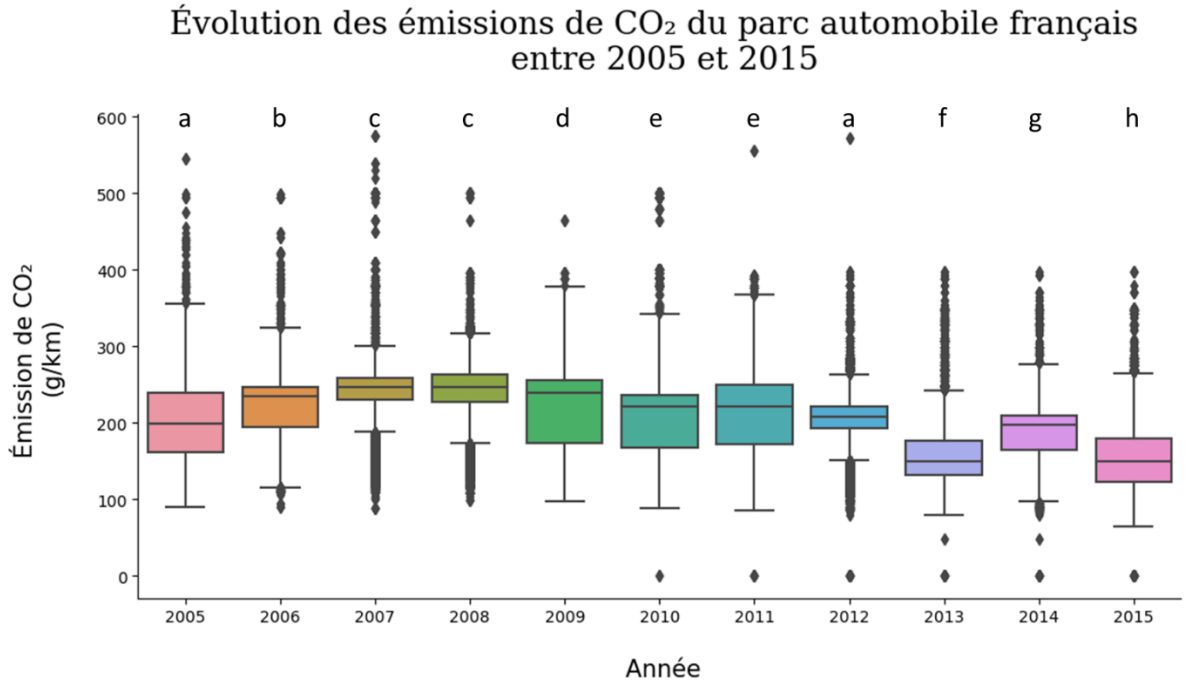


Figure 4 : Évolution des émissions de CO_2 du parc automobile français entre 2005 et 2015
Les données d'émissions de CO_2 médianes sont présentées indépendamment du constructeur et du type de véhicules. Les différences statistiques sont indiquées par un changement de lettre (p-value anova = 0.00). Les données sont initialement transmises par l'ADEME et disponible sur le site du gouvernement français.

En s'intéressant un peu plus en détail aux constructeurs (figure 5), on remarque que la majorité d'entre eux produisent des voitures de moins en moins polluantes, et ce dès les premières années de notre jeu de données. Seuls Mercedes-Benz, Nissan et Lamborghini ne semblent pas suivre ce constat. Le niveau d'émissions de CO_2 augmente même pour le constructeur Mercedes entre 2005 et 2007 pour ne redescendre qu'à partir de 2011. On remarque également que les profils d'émissions de ces trois constructeurs et principalement ceux de Mercedes-Benz et de Nissan sont relativement proches du profil d'émission global présenté plutôt dans la figure 4. Pour essayer d'éclaircir cette influence que semblent exercer ces deux constructeurs sur les données globales, nous avons quantifié le nombre de vente total des constructeurs entre 2005 et 2015 (figure 6).

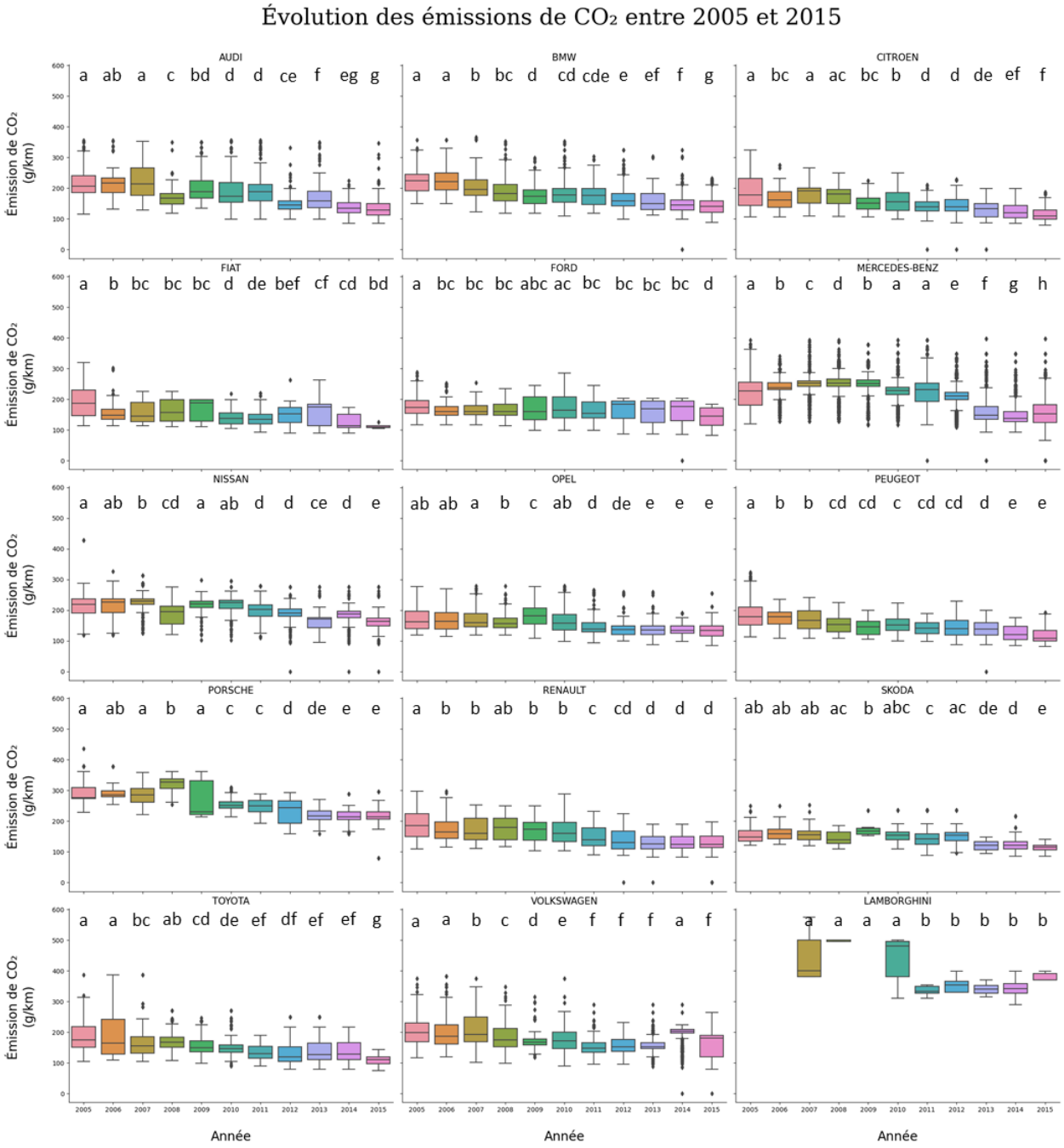


Figure 5 : Évolution des émissions de CO_2 par constructeurs entre 2005 et 2015

Les données d'émissions de CO_2 médianes sont présentées indépendamment du type de véhicules. Les différences statistiques sont indiquées par un changement de lettre (p-value anova = 0.00). Les données sont initialement transmises par l'ADEME et disponible sur le site du gouvernement français.

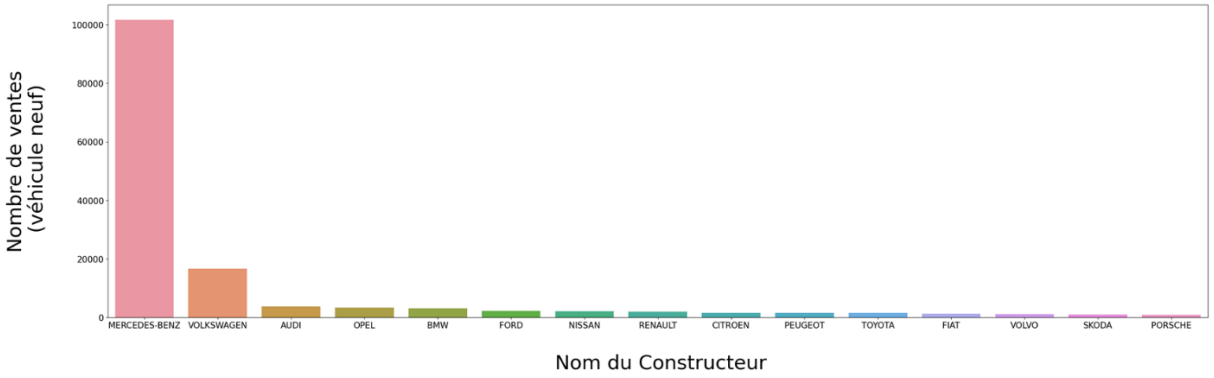


Figure 6 : Top 15 des constructeurs par ventes de véhicules entre 2005 et 2015
Données initialement transmises par l'ADEME et disponibles sur le site du gouvernement français.

Sans surprise, le constructeur Mercedes-Benz semble dominer largement le marché ce qui explique son influence sur le profil d'émissions de CO_2 global (figure 6). En effet, selon les données transmises par l'Ademe, Mercedes-Benz aurait vendu plus de 200 000 véhicules entre 2005 et 2015 contre seulement 20 000 véhicules pour le second constructeur, Volkswagen. On remarque également que les 4 premiers constructeurs sont allemands. 3 marques françaises sont dans le top 15 avec Renault, Citroën et Peugeot respectivement 7ème, 11ème et 12ème. Ces informations nous ont toutefois interpellé et en allant vérifier nos sources, il semblerait que les données disponibles sur le site du gouvernement et transmises par l'Ademe ne soient pas complètes. En effet, rien que pour l'année 2015, il semblerait que 1,9 millions de véhicules neufs ont été vendus et que les constructeurs français, Renault, Peugeot et Citroën détenaient 54,2 % de parts de marché (Le Point, 2016).

Une personne tierce a déjà signalé ce problème sur le site du gouvernement, mais aucune réponse n'a été apportée pour l'heure. Nous avons relancé la question car nous ne comprenons pas ces observations. Une de nos hypothèses est que les données transmises par l'Ademe dépendent directement des constructeurs. Ainsi, Mercedes-Benz communiquerait plus facilement ou plus efficacement leurs données associées aux ventes de leurs véhicules.

Une matrice de corrélation a ensuite été réalisée pour essayer d'avoir une vue d'ensemble des variables qui influent sur les émissions de CO_2 (figure 7). On observe notamment un cluster de corrélations positives fortes (r = 0.93 - 0.98) entre les variables conso_urb, conso_exurb, conso_mixte et co2. Cela semble logique puisque les émissions de CO_2 sont directement associées à la consommation du véhicule. Par ailleurs, on observe également une corrélation forte entre l'émission de CO_2 et l'émission d'autres polluants comme le HC NOX (r = 0.76) signifiant que les émissions du CO_2 vont de pair avec les émissions d'autres gazs de combustion : l'oxyde d'azote (NOX) et les hydrocarbures imbrûlés (HC). Enfin, on remarque également des corrélations plus modérées entre les émissions de CO_2 et la masse vide et chargée (respectivement r = 0.73 ; 0.70). Plus un véhicule est lourd, plus il émet du CO_2.

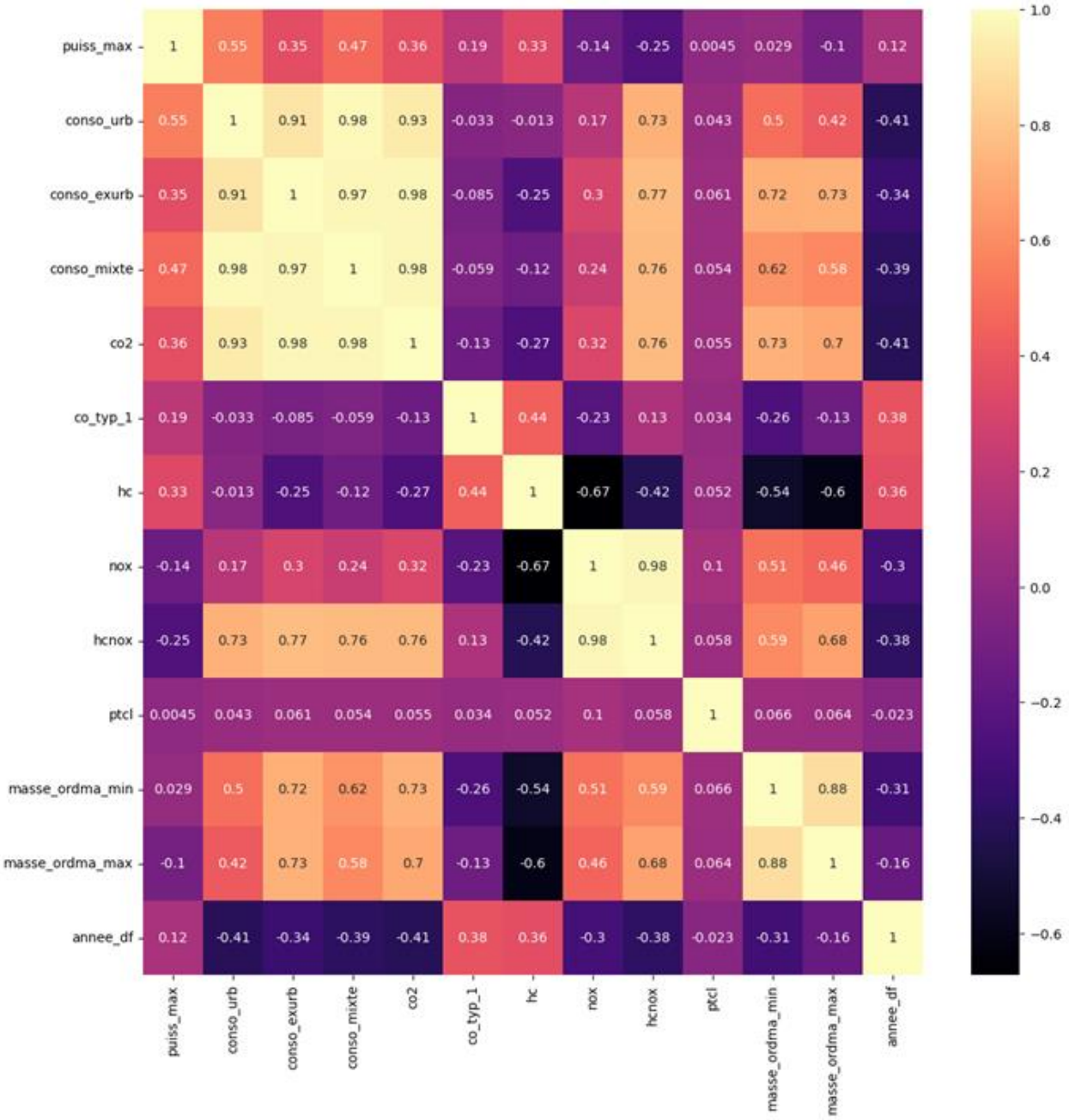


Figure 7 : Matrice de corrélation

En approfondissant nos recherches, on remarque que la relation entre les émissions CO_2 et le transport repose principalement sur l'utilisation des véhicules thermiques. En effet, les véhicules essences comme les véhicules diesels produisent beaucoup plus d'émissions de CO_2 que les véhicules dits écologiquement propres (Figure 8). A noter toutefois que le graphique présenté ci-dessous ne tient pas compte de l'évolution temporelle des émissions de CO_2 des véhicules thermiques qui de surcroît ont diminué entre 2005 et 2015 grâce à la mise en place du bonus/malus écologique.

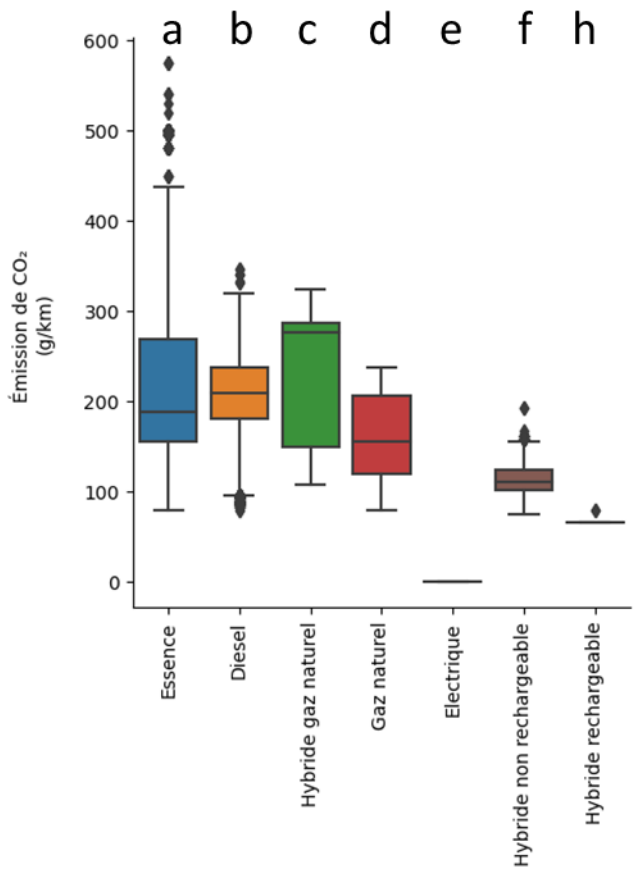


Figure 8 : Émissions de CO_2 en fonction du carburant

Les données d'émissions de CO_2 médianes sont présentées en fonction du type de carburant du véhicule. Les différences statistiques sont indiquées par un changement de lettre (p-value anova = 0.00). Les données sont initialement transmises par l'ADEME et disponibles sur le site du gouvernement français.

Toutefois, les ventes des véhicules propres restent dérisoires comparativement aux véhicules thermiques (Figure 9). Par ailleurs, l'achat d'une voiture électrique ou hybride rechargeable entre 2005 et 2015 pouvait être limité par d'autres facteurs comme la disponibilité des bornes de recharge ou l'autonomie de ces types de véhicules. Il serait ainsi intéressant d'effectuer des analyses similaires sur un pool de données plus récent.

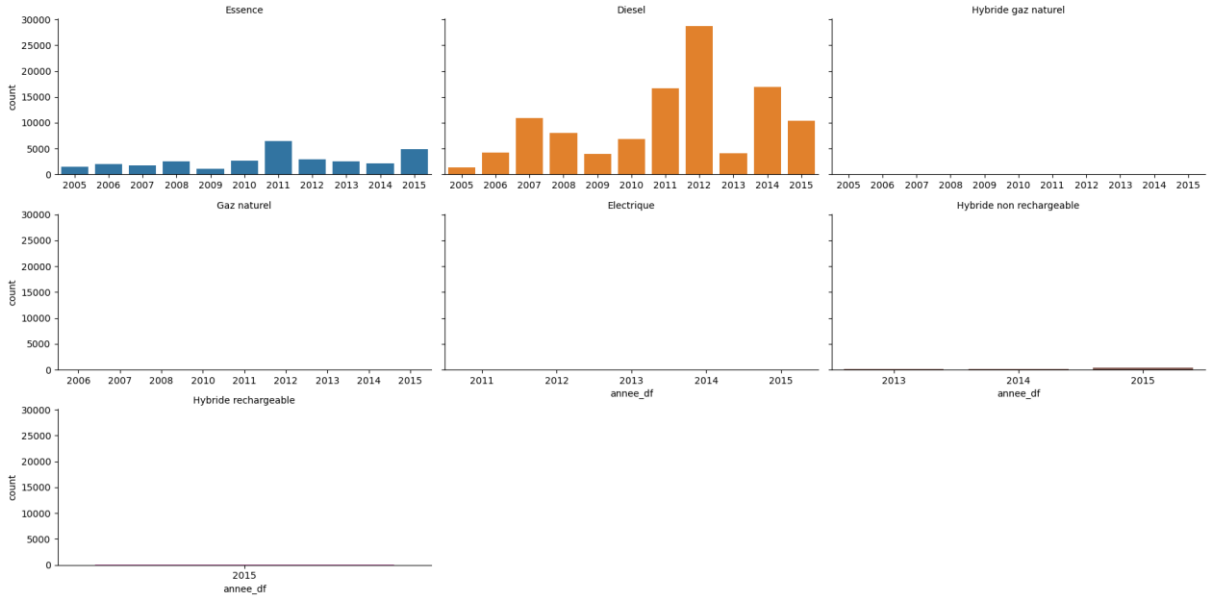


Figure 9 : Nombre de vente des véhicules en fonction de leur carburant entre 2005 et 2015

Le nombre de ventes des véhicules est présenté en fonction du type de carburant entre 2005 et 2015. L'échelle de l'axe Y est partagée entre tous les graphiques. (voir Annexe 3 pour plus de détails)

IV. Modélisation

À partir des données recueillies, nous allons essayer de modéliser les variables qui influencent les émissions de CO_2 des véhicules. Pour cela, nous nous sommes basés à la fois sur une variable cible numérique : la variable CO_2 qui correspond au niveau d'émission de CO_2 d'un véhicule, mais également sur une variable cible catégorielle : la variable étiquette énergétique qui résume les caractéristiques d'un véhicule en fonction de son niveau d'émission de CO_2.

Préparation du jeu de données

a. Data Cleaning

Les colonnes contenant majoritairement des données manquantes, telles que 'hc' et 'bonus_malus', ont été supprimées du jeu de données. Nous avons également choisi de supprimer les lignes contenant des valeurs manquantes afin d'éviter d'induire des biais dans les résultats de nos modélisations. Enfin, pour éviter certains problèmes de colinéarité, les variables associées à la consommation d'un véhicule (*i.e.* variables utilisées pour calculer les émissions de CO_2) ont été supprimées du jeu de données ainsi que la variable 'nox' qui est fortement corrélée à la variable 'hcnox'.

b. Pré-encodage

Certaines variables catégorielles comme la marque, le modèle ou le numéro cnit d'un véhicule contiennent beaucoup de modalités différentes, ce qui rend leur encodage compliqué. Elles ont donc été supprimées du jeu de données. D'autres variables comme le type de carburant, le type de boîte de vitesses, le champ V9, ou encore le type de carrosserie, possèdent un nombre de modalité plus restreint, mais ont toutefois nécessité une compilation de ces modalités afin de diminuer leur nombre.

c. Encodage

Toutes les variables catégorielles excepté la variable étiquette énergétique ont reçu un encodage de type "one hot encoding" en supprimant la première colonne afin de limiter les problèmes de colinéarité. La variable étiquette énergétique a quant-à-elle reçu un encodage ordinal avec un score allant de 0 à 6 pour les modalités de "G" à "A".

d. Pré-processing

Les variables numériques ont toutes été normalisées afin d'éviter des impacts négatifs sur la qualité de nos modèles associés aux différentes échelles des données.

Modèles réalisés

L'étude des émissions de CO_2 a été réalisée avec deux types de modélisations distincts. Le premier volet de modélisations a pour objectif de déterminer la valeur d'émission de CO_2 d'un véhicule et repose sur des modélisations de type régression. Le second volet de modélisations a pour objectif de déterminer la classe énergétique d'un véhicule et repose sur des modélisations de type classification.

a. Modélisations des émissions de CO_2

Les émissions de CO_2 d'un véhicule ont été modélisées sous trois modèles de machine learning :

- Régression linéaire
- Arbre de régression
- Régression Random Forest

Les résultats de ces modélisations sont présentés dans le tableau 2. Par souci de lisibilité, la profondeur des arbres a été limitée à cinq nœuds. À noter toutefois que sans cette limitation, le coefficient de détermination des arbres de régression et de random forest dépasse 0.98.

Tableau 2 : Résultats et métriques des modèles de régression sur CO_2

Modèle	R^2	MAE	MSE	RMSE
Régression linéaire	train : 0.9315	6.69	74.79	8.65
	test : 0.9322	6.72	74.85	8.65
Arbre de Régression	train : 0.9328	6.21	73.46	8.57
	test : 0.9306	6.32	76.60	8.75
Régression Random Forest	train : 0.9378	6.07	68.00	8.25
	test : 0.9358	6.19	70.89	8.42

D'après les différentes métriques, le modèle représentant le plus fidèlement les émissions de CO_2 est la régression associée au random forest. En effet, cette régression affiche le R^2 le plus élevé (R^2 = 0.9358, tableau 2), ainsi que l'erreur absolue moyenne la plus faible (MAE = 6.19, tableau 2). À noter toutefois que les trois modèles sont pertinents (all RMSE < 10 pour un effectif de 12 124, tableau 2) et qu'ils minimisent l'erreur de prédiction à moins de 7g de CO_2 émis par véhicule (all MAE < 7, tableau 2).

Interprétation du modèle

Le modèle de régression associée au random forest a ensuite été analysé à l'aide de la méthode des valeurs Shapley (figure 10). Les variables sont ordonnées de la plus importante (*i.e.* 'masse_ordma_max') à la moins importante (*i.e.* 'carrosserie_BREAK) pour le modèle d'entraînement. Ainsi, la masse en charge d'un véhicule semble-être la caractéristique majeure déterminant son taux d'émission de CO_2. Plus cette masse est élevée, plus le véhicule émet du CO_2 et inversement. Cela concorde notamment avec les deux premières modalités de carrosseries retrouvées dans le classement : « Monospace » et « tout terrain et chemins » qui influence également positivement, mais dans une moindre mesure, les émissions de CO_2. La seconde variable la plus importante concerne les émissions d'hydrocarbures imbrulés et d'oxyde d'azote. Cette variable est positivement corrélée aux émissions de CO_2. Plus une voiture émet du CO_2, plus elle émet également d'autres polluants associés à la combustion des moteurs thermiques. Cette caractéristique et ses résultats se retrouvent notamment à travers la

représentation de la modalité euro6 du champ v9 qui est négativement corrélée aux émissions de CO_2 et qui correspond au moteur les moins polluants vis-à-vis des oxydes d'azote, du monoxyde de carbone, des hydrocarbures imbrulés et des particules fines. Il est également intéressant de noter que la troisième variable la plus importante selon les valeurs shapley représente les boîtes de vitesse (Manuelle : oui ou non). Cette variable est négativement corrélée aux émissions de CO_2. Cela implique que les véhicules possédant une boîte de vitesse automatique semblent émettre en moyenne plus de CO_2 que les véhicules manuels.

Ainsi, notre premier modèle nous indique que pour limiter les émissions de CO_2, le véhicule le plus intéressant serait un petit véhicule manuel qui permettrait par la même occasion de diminuer les émissions d'autres polluants de l'air.

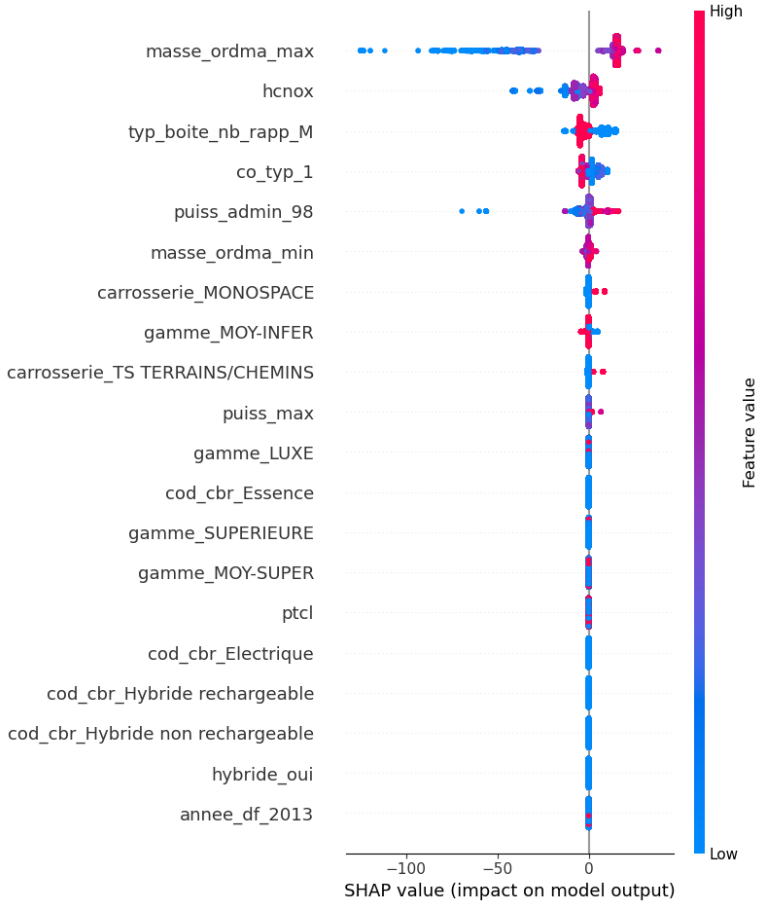


Figure 10 : Summary plot des valeurs de shapley par variable du modèle de régression associé au random forest pour la modélisation de la classe énergétique

Les valeurs de shapley sont représentées pour chaque variable en fonction de leur ordre d'importance dans la bonne prédiction de la valeur d'émission de CO_2 d'un véhicule. Pour chaque variable, on peut observer si son poids influence négativement ou positivement le taux d'émission de CO_2.

Amélioration du modèle

Afin d'améliorer notre modèle, nous avons essayé de réduire ses dimensions en ne conservant que les variables les plus importantes à savoir la masse en charge maximale, le type de boîte de vitesse et le niveau de hcnox (déterminé à l'aide du features importance, Annexe 4,A). Cependant, les résultats du modèle réduit sont moins pertinents que le modèle général (tableau 3).

Tableau 3 : Résultats et métriques du modèle de régression réduit sur CO_2

Modèle	R^2	MAE	MSE	RMSE
Régression Random Forest réduit	train : 0.9180	6.76	89.63	9.47
	test : 0.9186	6.84	89.77	9.47

b. Modélisation de la classe énergétique

La classe énergétique d'un véhicule a été modélisée sous trois modèles de machine learning :

- Régression logistique
- Arbre de classification
- Classification Random Forest

Les résultats de ces modélisations sont présentés dans le tableau 4. Par souci de lisibilité, la profondeur des arbres a été limitée à cinq nœuds. À noter toutefois que sans cette limitation, le coefficient de détermination des arbres de classification et de random forest sont supérieurs à 0.87.

D'après les différentes métriques, les trois modèles de classification affichent des résultats très proches. Selon l'accuracy, le modèle représentant le plus fidèlement la classe énergétique d'un véhicule semble-être l'arbre de classification (accuracy = 0.8200, tableau 4). Cependant, les trois modèles sont pertinents et en s'intéressant aux autres métriques (tableau 4), notamment le F1-score, il est difficile d'affirmer qu'un modèle est plus fidèle qu'un autre.

Par ailleurs, les visualisations des prédictions de chaque modèle (figure 11) nous apprennent que les trois modèles affichent encore une fois des résultats similaires. Cependant, ici, seul les modèles de régression logistique et d'arbre de classification parviennent à classer des véhicules en classe énergétique A (A=6) (figure 11). Cette classe ne contient que très peu de véhicules, il est donc compliqué de bien les classer. On observe également que les modèles sont relativement précis dans le choix des classes des véhicules et que lorsqu'ils se trompent, ils classent les véhicules dans une classe adjacente. Enfin, on observe que les véhicules les mieux classés sont ceux qui appartiennent aux classes les plus représentées (classes 1 et 2 = étiquettes énergétiques E et F).

Tableau 4 : Résultats et métriques des modèles de classification sur la classe énergétique

Modèle	Accuracy
Régression Logistique	train : 0.7873 test : 0.7871
Arbre de Classification	train : 0.8247 test : 0.8200
Classification Random Forest	train : 0.8246 test : 0.8185

	Precision			Recall			F1-score			
Classe	Rég. Log.	Arb. Clas.	Clas. Rd. Forest	Rég. Log.	Arb. Clas.	Clas. Rd. Forest	Rég. Log.	Arb. Clas.	Clas. Rd. Forest	n
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3
F	0.86	0.92	0.92	0.87	0.87	0.88	0.86	0.89	0.90	5962
E	0.75	0.79	0.81	0.73	0.83	0.83	0.74	0.81	0.82	3802
D	0.54	0.57	0.70	0.49	0.51	0.33	0.52	0.54	0.45	642
C	0.72	0.65	0.58	0.79	0.79	0.85	0.75	0.72	0.69	1055
B	0.67	0.67	0.57	0.76	0.69	0.69	0.71	0.68	0.63	541
A	0.60	0.81	1.00	0.39	0.43	0.11	0.47	0.56	0.20	119

A

Prédiction / Realité	1	2	3	4	5	6
0	0	3	0	0	0	0
1	5165	795	2	0	0	0
2	843	2776	163	20	0	0
3	0	97	317	214	14	0
4	0	14	88	830	120	3
5	0	0	13	91	409	28
6	0	0	1	1	71	46

B

Prédiction / Realité	1	2	3	4	5	6
0	3	0	0	0	0	0
1	5180	781	0	1	0	0
2	458	3171	116	54	3	0
3	7	43	328	243	21	0
4	0	3	98	837	117	0
5	0	0	29	125	375	12
6	0	0	2	19	47	51

C

Prédiction / Realité	1	2	3	4	5	6
0	3	0	0	0	0	0
1	5265	697	0	0	0	0
2	447	3163	83	104	5	0
3	0	40	215	363	24	0
4	0	0	9	893	153	0
5	0	0	0	166	375	0
6	0	0	1	9	96	13

Figure 11 : Visualisation de la fidélité des prédictions des modèles
A : Modèle régression logistique, B : Arbre de Classification, C : Classification Random Forest

Interprétation du modèle

L'arbre de classification a ensuite été analysé à l'aide de la méthode des valeurs Shapley (figure 12). Les variables sont ordonnées de la plus importante (*i.e.* 'masse_ordma_max') à la moins importante (*i.e.* 'carrosserie_MINISPACE') pour le modèle d'entraînement en discriminant l'importance de chaque variable dans la classification des classes énergétiques. Ainsi, comme pour les modèles de régression, la masse du véhicule en charge semble également être la variable la plus importante pour déterminer l'ensemble des classes énergétiques. La seconde variable la plus importante du modèle correspond au type de boîte de vitesse et semble permettre de différencier principalement les véhicules des classes E, F et C. Enfin, on retrouve les émissions de deux autres polluants, le monoxyde de carbone et les hydrocarbures suivi par la puissance administrative du véhicule.

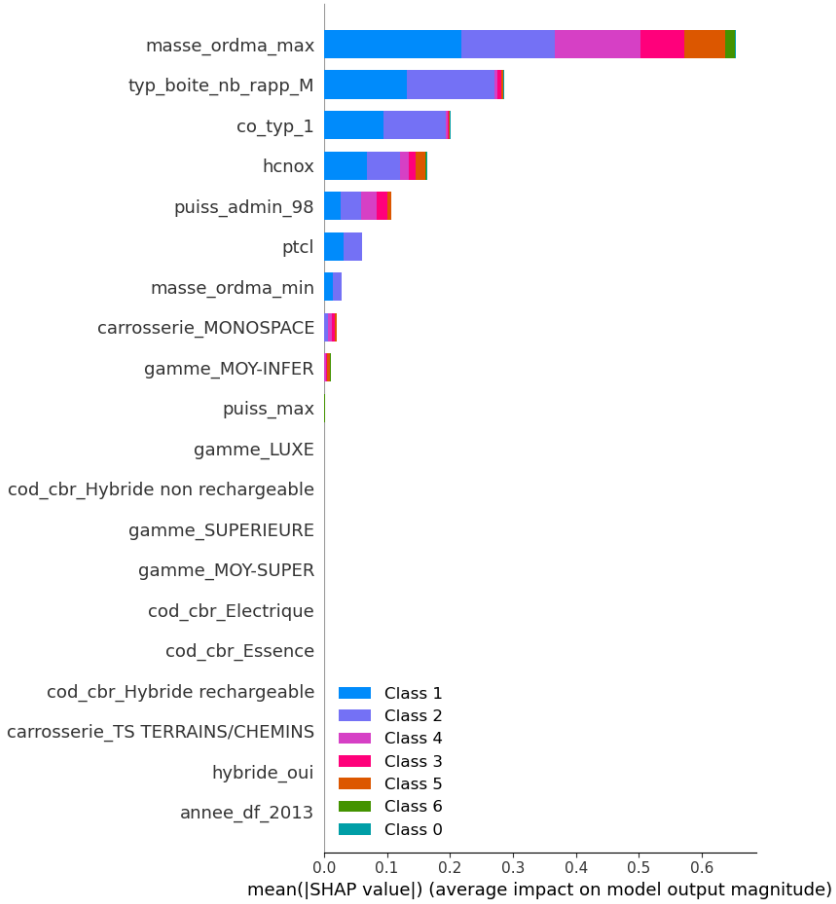


Figure 12 : Summary plot des valeurs de shapley par variable du modèle de régression associé au random forest

Les valeurs de shapley sont représentées pour chaque variable en fonction de leur ordre d'importance dans la bonne prédiction de la classe énergétique du véhicule. Class 6 : Classe énergétique A, Class 5 : Classe énergétique B, Class 4 : Classe énergétique C, Class 3 : Classe énergétique D, Class 2 : Classe énergétique E, Class 1 : Classe énergétique F, Class 0 : Classe énergétique G

Pour essayer d'aller plus loin, nous avons observé les différences entre les summary plot pour chaque classe énergétique. Dans ce rapport, nous nous attarderons sur les summary plot des classes F et B (figure 13).

La classe F est la classe majoritaire dans notre jeu de données. Elle représente à elle seule plus de 50% des véhicules. Il semblerait que cette classe soit majoritairement composée de véhicules automatiques avec une masse en charge relativement élevée comme en témoigne les deux variables les plus importantes du summary plot (figure 13 F). Il semblerait également que ces véhicules ont tendance à émettre d'autres polluants et qu'ils ne correspondent pas à des véhicules hybrides. À l'inverse, les véhicules énergétiquement catégorisés "B" semblent-être des véhicules manuels de gamme moyenne-inférieure ayant une masse en charge faible (figure 13 A). Ces véhicules ne semblent pas émettre beaucoup d'autres polluants et possèderaient également une puissance administrative relativement faible.

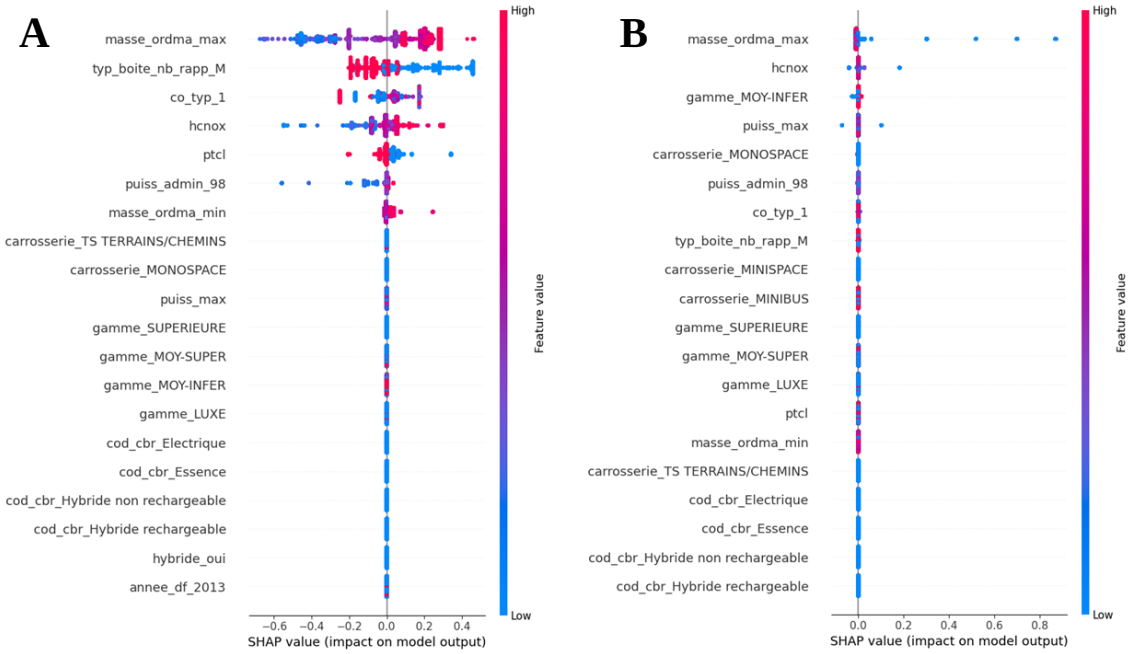


Figure 13 : Summary plot des valeurs de shapley par variable du modèle de régression associé au random forest pour les classes F (graphique A) et A (graphique B)

Les résultats obtenus suite à la modélisation des classes énergétiques sont relativement similaires à ceux obtenus suite à la modélisation des émissions de CO_2. Les véhicules qui émettent le moins de CO_2 sont de petits véhicules manuels.

<u>Amélioration du modèle</u>

Afin d'améliorer notre modèle, nous avons essayé de réduire ses dimensions en ne conservant que les variables les plus importantes à savoir le type de boîte de vitesse, la masse en charge, la carrosserie "minibus" et le niveau d'émission de monoxyde de carbone (déterminé à l'aide du features importance, Annexe 4,B). Cependant, les résultats du modèle réduit sont moins pertinents que le modèle général (tableau 5).

Tableau 5 : Résultats et métriques du modèle de classification réduit sur la classe énergétique

Modèle	Accuracy
Arbre de Classification réduit	train : 0.8157 test : 0.8086

Avant de conclure ce rapport, nous tenions à vérifier nos modèles à l'aide d'une troisième analyse. En effet, nous trouvons particulier le fait que le modèle le plus fidèle pour déterminer les émissions de CO_2 soit un modèle de régression et non un modèle de classification. Nous avons donc réalisé des nouveaux modèles à partir des quantiles de la variable CO_2 (tableau 6). Les résultats de ces modèles sont moins probants que ceux préalablement réalisés. Notre modèle le plus fidèle semble donc bel et bien être un modèle de régression.

Tableau 6 : Résultats et métriques des modèles de classification sur les quantiles de CO_2

Modèle	Accuracy
Régression Logistique	train : 0.7053 test : 0.7063
Arbre de Classification	train : 0.8013 test : 0.8025
Classification Random Forest	train : 0.7654 test : 0.7648

Classe	Precision			Recall			F1-score			
	Rég. Log.	Arb. Clas.	Clas. Rd. Forest	Rég. Log.	Arb. Clas.	Clas. Rd. Forest	Rég. Log.	Arb. Clas.	Clas. Rd. Forest	n
75-100	0.64	0.71	0.97	0.61	0.97	0.51	0.63	0.82	0.67	2028
50-75	0.55	0.96	0.66	0.56	0.56	0.80	0.56	0.71	0.72	2751
25-50	0.68	0.70	0.67	0.84	0.97	0.94	0.75	0.81	0.78	3502
0-25	0.93	0.99	0.98	0.74	0.73	0.71	0.82	0.84	0.82	3830

V. Conclusion :

Un des points essentiels au bon déroulement de ce projet a été le nettoyage et l'harmonisation des jeux de données disponibles sur le site du gouvernement. En effet, nous souhaitions utiliser un maximum de données disponibles afin d'obtenir les analyses les plus pertinentes. Cependant, la structure des jeux de données étaient étonnamment très hétérogène entre les années. Il nous a donc fallu faire un gros travail de nettoyage afin de pouvoir fusionner les dataframes entre eux. Suite à cela, nous avons dû supprimer les lignes correspondant aux mêmes véhicules passant ainsi d'un dataframe de 285 830 lignes à un dataframe de 152 986 lignes. Cela nous a permis de faire des figures et des statistiques intéressantes retraçant les données transmises entre 2005 et 2015 et ce malgré l'importance du nombre de valeurs manquantes.

Pour l'étape de modélisation nous avons dû supprimer les valeurs manquantes, passant de 152 986 à 46 577 lignes et ne représentant plus que les données pour les années 2012, 2013 et 2014. Les résultats de nos modélisations restent toutefois pertinents vis-à-vis de nos données et ne semblent pas avoir été impactés par l'importance des valeurs manquantes. Cependant, nos modélisations constituent également le problème majeur de notre projet. En effet, le but d'une modélisation est de pouvoir anticiper et prédire, à travers des calculs, des évènements ou des choix à venir. Or, la pertinence d'une modélisation dépend initialement des données utilisées. Dans notre cas, nous savons que les données utilisées et transmises par l'ademe ne correspondent pas aux véritables données du parc automobile français. Pour rappel, d'après les données transmises par cet organisme, le parc automobile français serait constitué d'un peu moins de 300 000 véhicules achetés entre 2005 et 2015. Le principal constructeur serait Mercedes-Benz avec 65% de parts de marché, soit plus de six véhicules Mercedes-Benz pour dix véhicules toutes marques confondues. Or, un article publié dans le Point en 2016 indique que 1,9 millions de véhicules neufs ont été vendus rien qu'en 2015 et que les constructeurs français, Renault, Peugeot et Citroën détenaient 54,2 % de parts de marché (Le Point, 2016). Ainsi, il est fortement probable que les résultats de nos modélisations ne soient pas tout à fait fidèles aux données réelles. On peut toutefois s'attendre à ce que la masse d'un véhicule ainsi que le type de boîte de vitesses jouent toujours un rôle prépondérant dans le niveau d'émissions de CO_2. Il serait également intéressant de réaliser ces analyses avec un jeu de données actualisé afin de prendre en compte l'apparition sur le marché des voitures 100% électriques rechargeables.

Nous tenions à terminer ce rapport sur une observation, une réflexion et une critique que nous nous sommes faites. Si le but de ce projet est d'évaluer l'impact d'un véhicule sur le climat, alors il nous semble nécessaire de comptabiliser également les émissions de CO_2 émises lors de la conception des véhicules. De la même manière, il serait intéressant de prendre en compte la source d'énergie utilisée pour alimenter les voitures électriques et hybrides. En effet, une voiture 100% électrique émet indirectement bien plus de CO_2 en Allemagne (principale source d'électricité = centrale thermique au charbon) qu'en France (principale source d'électricité = centrale nucléaire) (Canals Casals et al., 2016). Enfin, il semblerait que le niveau d'émissions de CO_2 d'un véhicule dépend également de son utilisation (figure 14). On remarque notamment que plusieurs véhicules devraient être énergétiquement déclassés en fonction de leurs émissions urbaines, extra-urbaines ou un mixte des deux (outliers supérieurs). Il serait donc intéressant de mettre en place une attribution plus pertinente et fidèle pour représenter les classes énergétiques.

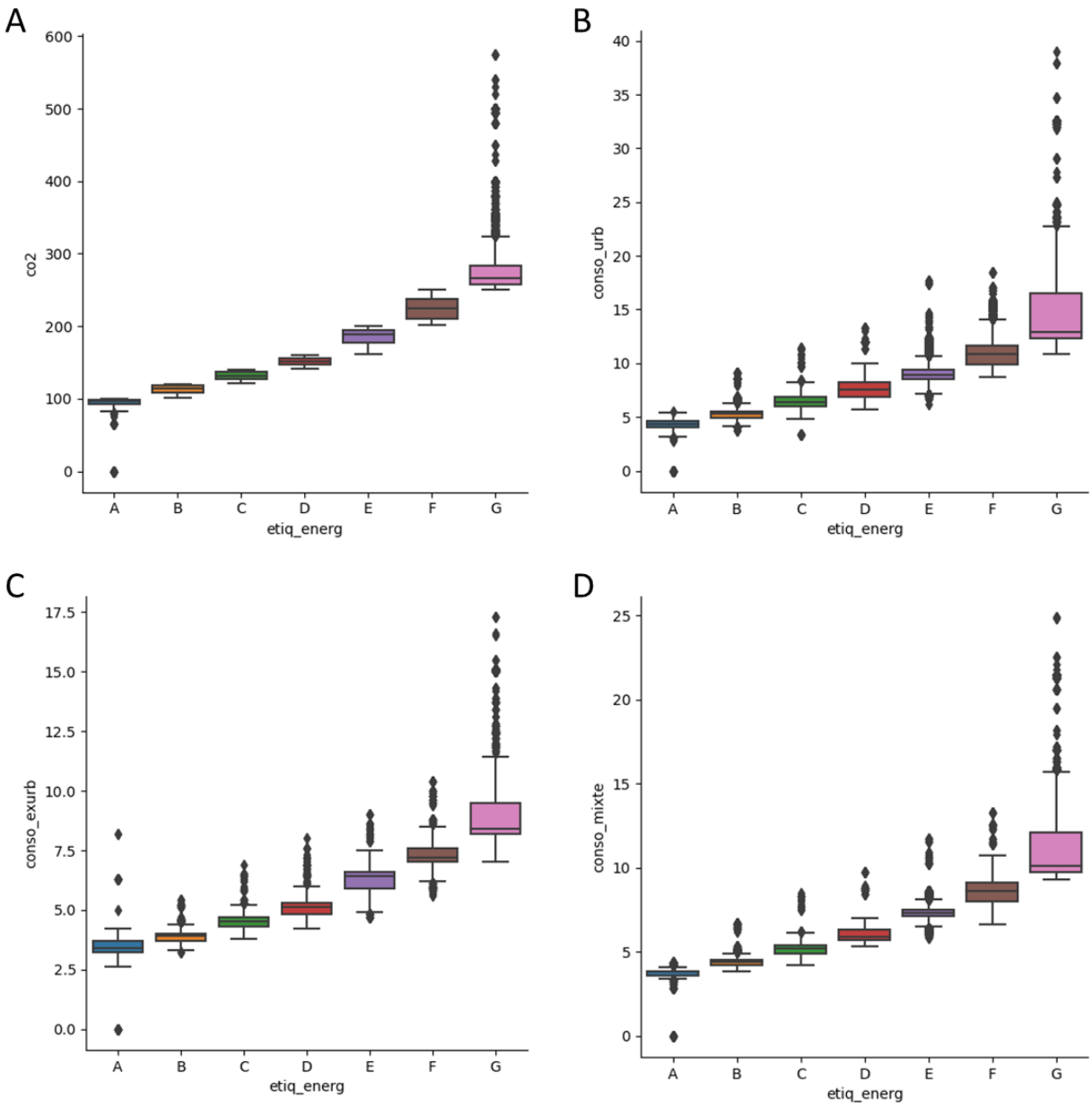


Figure 14 : Représentation des classes énergétiques en fonction des émissions de CO_2 et de la consommation d'un véhicule

A : Émissions de CO_2 ; B : Consommation urbaine d'un véhicule ; C : Consommation ex-urbaine d'un véhicule ; D : Consommation mixte d'un véhicule. Il est intéressant de noter que les émissions de CO_2 dépendent directement de la consommation d'un véhicule, mais que la classe énergétique de certains véhicules ne correspond pas à sa consommation (présence d'outliers max).

Annexe 1

Dictionnaire des variables utilisées

Intitulé final	Définition
lib_mrq_utac	marque du véhicule
lib_mod	modèle commercial
cnit	Code National d'Identification du Type. Ce code est propre à chaque véhicule
cod_cbr	type de carburant
hybride	information permettant d'identifier les véhicules hybrides (O/N)
puiss_admin_98	puissance administrative du véhicule, en chevaux dits fiscaux (CV)
puiss_max	puissance maximale du véhicule (en kW)
typ_boite_nb_rapp	type de boîte de vitesse et le nombre de rapports
conso_urb	consommation urbaine de carburant (en l/100km)
conso_exurb	consommation extra urbaine de carburant (en l/100km)
conso_mixte	consommation mixte de carburant (en l/100km)
co2	émission de CO2 (en g/km)
co_typ_1	résultat d'essai de CO type I (en g/km)
hc	résultats d'essai HC (gaz imbrulés, en g/km)
nox	résultats d'essai NOx (oxyde d'azote, en g/km)
hcnox	résultats d'essai HC+NOX (en g/km)
ptcl	résultat d'essai de particules (g/km)
masse_ordma_min	masse en ordre de marche mini (kg),
masse_ordma_max	masse en ordre de marche maxi (kg),
champ_v9	champ V9 du certificat d'immatriculation qui contient la norme EURO
carrosserie	type de carrosserie (berline, coupé, cabriolet, etc)
gamme	gamme du véhicule (économique, moy-sup, luxe, etc)
etiq_energ	classification en fonction des émissions de CO_2 du véhicule
bonus_malus	bonus ou malus écologique. S'applique sur l'achat d'un véhicule neuf en fonction de ces émissions de CO_2

Annexe 2

Intitulés initiaux des variables pour chaque dataframe de 2005 à 2015

Présence (OUI) ou absence (NON) De la colonne dans chaque jeu de données de 2005 à 2015

Année	lib_mrq_utac	Marque	MARQUE	MODELE VERSION	Modèle dossier	Modèle UTAC	Désignation commerciale	lib_mrq_doss	lib_mrq	mrq_utac	lib_mod_doss
2005	NON	NON	OUI	OUI	NON	NON	NON	NON	NON	NON	NON
2006	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2007	NON	NON	OUI	OUI	NON	NON	NON	NON	NON	NON	NON
2008	NON	NON	OUI	OUI	NON	NON	NON	NON	NON	NON	NON
2009	NON	NON	OUI	OUI	NON	NON	NON	NON	NON	NON	NON
2010	NON	NON	OUI	OUI	NON	NON	NON	NON	NON	NON	NON
2011	NON	NON	NON	NON	NON	NON	NON	NON	OUI	NON	OUI
2012	NON	NON	NON	NON	NON	NON	NON	NON	OUI	NON	OUI
2013	NON	OUI	NON	NON	OUI	OUI	OUI	NON	NON	NON	NON
2014	NON	NON	NON	NON	NON	NON	NON	NON	OUI	NON	OUI
2015	NON	NON	NON	NON	NON	NON	NON	OUI	NON	OUI	OUI

Année	hybride	Hybride	puiss_admin_98	puiss_max	Puissance administrative	puissance fiscale	Puissance maximale (kW)	puissance reelle	typ_boite_nb_rapp	bv	Boîte de vitesse
2005	NON	NON	NON	NON	NON	OUI	NON	OUI	NON	OUI	NON
2006	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2007	NON	NON	NON	NON	NON	OUI	NON	OUI	NON	OUI	NON
2008	NON	NON	NON	NON	NON	OUI	NON	OUI	NON	OUI	NON
2009	NON	NON	NON	NON	NON	OUI	NON	OUI	NON	OUI	NON
2010	NON	NON	NON	NON	NON	OUI	NON	OUI	NON	OUI	NON
2011	NON	NON	OUI	OUI	NON	NON	NON	NON	OUI	NON	NON
2012	OUI	NON	OUI	OUI	NON	NON	NON	NON	OUI	NON	NON
2013	NON	OUI	NON	NON	OUI	NON	OUI	NON	NON	NON	OUI
2014	OUI	NON	OUI	OUI	NON	NON	NON	NON	OUI	NON	NON
2015	OUI	NON	OUI	OUI	NON	NON	NON	NON	OUI	NON	NON

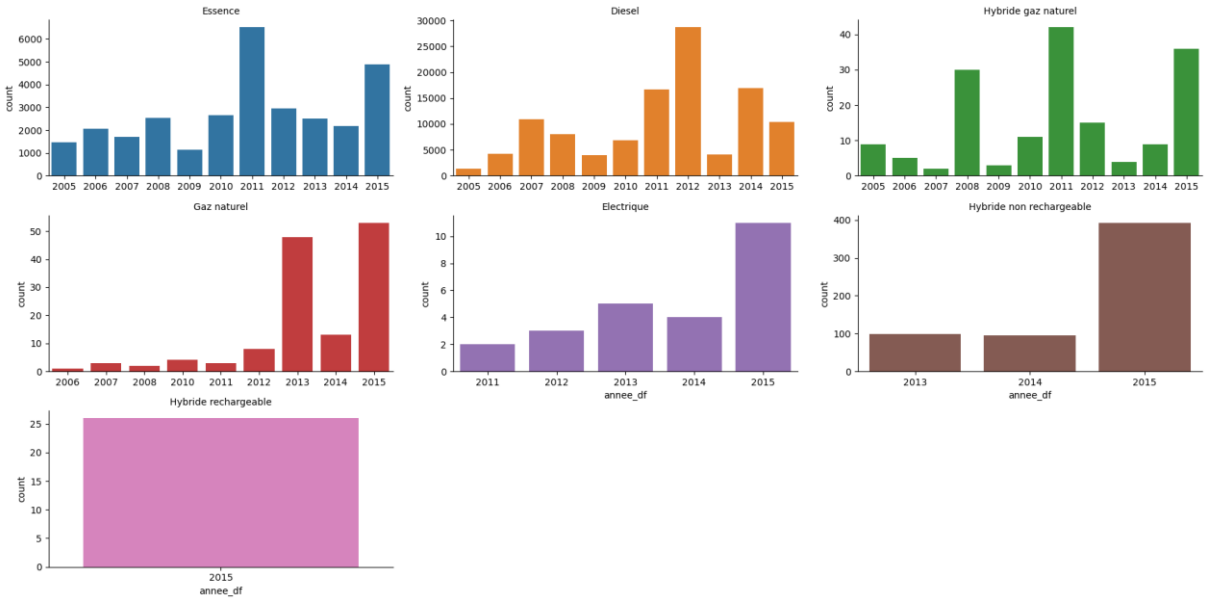
Année	co_typ_1	CO type I (g/km)	hc	HC (g/km)	nox	NOX (g/km)	hcnox	HC+NOX (g/km)	ptcl	Particules (g/km)	masse_ordma_min
2005	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2006	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2007	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2008	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2009	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2010	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2011	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2012	OUI	NON	OUI	NON	OUI	NON	OUI	NON	OUI	NON	OUI
2013	NON	OUI	NON	OUI	NON	OUI	NON	OUI	NON	OUI	NON
2014	OUI	NON	OUI	NON	OUI	NON	OUI	NON	OUI	NON	OUI
2015	OUI	NON	OUI	NON	OUI	NON	OUI	NON	OUI	NON	OUI

Année	lib_mod	dscom	cnit	CNIT	tvv	Type Variante Version (TVV)	cod_cbr	typ_cbr	energ	ENERGIE	Carburant	carburant
2005	NON	NON	NON	OUI	NON	NON	NON	NON	NON	OUI	NON	NON
2006	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2007	NON	NON	NON	OUI	NON	NON	NON	NON	NON	NON	NON	OUI
2008	NON	NON	NON	OUI	NON	NON	NON	NON	NON	NON	NON	OUI
2009	NON	NON	NON	OUI	NON	NON	NON	NON	NON	NON	NON	OUI
2010	NON	NON	NON	OUI	NON	NON	NON	NON	NON	NON	NON	OUI
2011	OUI	OUI	OUI	NON	OUI	NON	NON	OUI	NON	NON	NON	NON
2012	OUI	OUI	OUI	NON	OUI	NON	NON	OUI	NON	NON	NON	NON
2013	NON	NON	NON	OUI	NON	OUI	NON	NON	NON	NON	OUI	NON
2014	OUI	OUI	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	NON
2015	NON	OUI	OUI	NON	OUI	NON	NON	NON	OUI	NON	NON	NON

Année	conso_urb	urb	conso_exurb	ex-urb	conso_mixte	mixte	Consommation urbaine (l/100km)	Consommation extra-urbaine (l/100km)	Consommation mixte (l/100km)	co2	CO2 (g/km)	CO2
2005	NON	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	NON	OUI
2006	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2007	NON	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	NON	OUI
2008	NON	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	NON	OUI
2009	NON	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	NON	OUI
2010	NON	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	NON	OUI
2011	OUI	NON	OUI	NON	NON	NON	NON	NON	NON	OUI	NON	NON
2012	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	OUI	NON	NON
2013	NON	NON	NON	NON	NON	NON	OUI	OUI	OUI	NON	OUI	NON
2014	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	OUI	NON	NON
2015	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	OUI	NON	NON

Année	masse vide euro min (kg)	masse_ordma_max	masse vide euro max (kg)	champ_v9	Champ V9	date_maj	Date de mise à jour	Carrosserie	gamme	ETIQUETTE ENERGIE	BONUS(-)/ MALUS(+)	TYPE 2
2005	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2006	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON	NON
2007	NON	NON	NON	NON	NON	NON	NON	NON	NON	OUI	NON	NON
2008	NON	NON	NON	NON	NON	NON	NON	NON	NON	OUI	OUI	OUI
2009	NON	NON	NON	NON	NON	NON	NON	NON	NON	OUI	OUI	NON
2010	NON	NON	NON	NON	NON	NON	NON	NON	NON	OUI	OUI	NON
2011	NON	NON	NON	OUI	NON	OUI	NON	NON	NON	NON	NON	NON
2012	NON	OUI	NON	OUI	NON	OUI	NON	OUI	OUI	NON	NON	NON
2013	OUI	NON	OUI	NON	OUI	NON	OUI	OUI	OUI	NON	NON	NON
2014	NON	OUI	NON	OUI	NON	OUI	NON	OUI	OUI	NON	NON	NON
2015	NON	OUI	NON	OUI	NON	OUI	NON	NON	NON	NON	NON	NON

Annexe 3



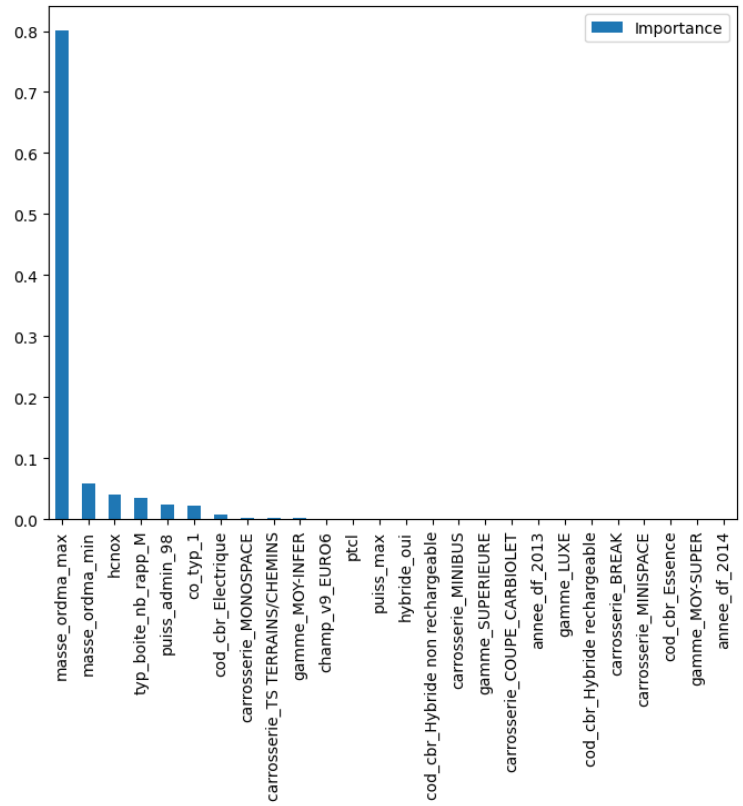
Nombre de vente des véhicules en fonction de leur carburant entre 2005 et 2015

Le nombre de ventes des véhicules est présenté en fonction du type de carburant entre 2005 et 2015. L'échelle de l'axe Y n'est pas partagée entre tous les graphiques.

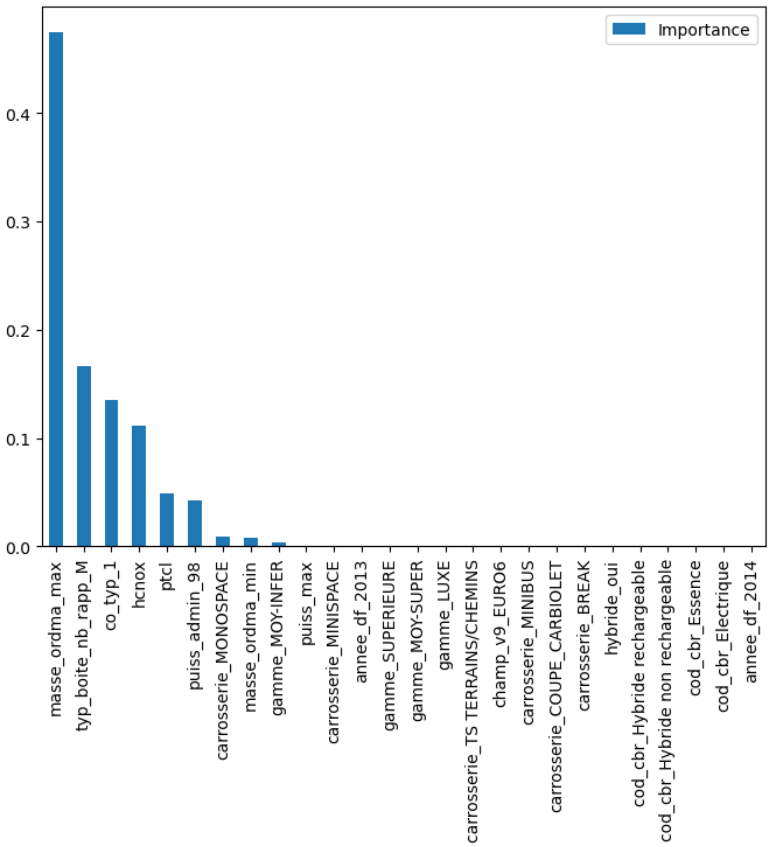
Annexe 4

Feature importances des modèles CO_2 (A) et étiquettes énergétiques (B)

A



B



Bibliographie

Ademe Emissions de CO2 et de polluants des véhicules commercialisés en France - data.gouv.fr.

ARISTIZABAL (V), S., MEURISSE, B., SOLIDAIRE, M. D. L. T. E. E. and CGDD Service de l'économie, de l'évaluation et de l'intégration du développement durable (2019). *Dispositifs d'étiquetage : bonnes pratiques et écueils à éviter. Cas des produits polluant l'air intérieur.* Ministère de la transition écologique. Paris.

Canals Casals, L., Martinez-Laserna, E., Amante García, B. and Nieto, N. (2016). Sustainability analysis of the electric vehicle use in Europe for CO2 emissions reduction. *Journal of Cleaner Production* **127**, 425–437.

European Parliament (2019). CO2 emissions from cars: facts and figures.

IPCC, Allan, R. P., Cassou, C., Chen, D., Cherchi, A., Connors, L., Doblas-Reyes, F. J., Douville, H., Driouech, F., Edwards, T. L., et al. (2021). Summary for Policymakers. *Climate Change 2021: The Physical Science Basis.*

Le Point (2016). Le marché auto français décolle en 2015. *Le Point*.

Mukherji, A., Thorne, P., Cheung, W. W. L., Connors, S. L., Garschagen, M., Geden, O., Hayward, B., Simpson, N. P., Totin, E., Blok, K., et al. (2023). Synthesis report of the ipcc sixth assessment report (ar6).

Robequain, L. (2010). Très coûteux, le bonus-malus automobile sera durci en 2011. *Les Echos*.

U.S. Global Change Research Program ed. (2009). *Global climate change impacts in the United States: a state of knowledge report*. Cambridge [England] ; New York: Cambridge University Press.