Journées de méthodologie statistique de l'Insee 2022

Modélisation de l'appartenance au parc des véhicules routiers et de son utilisation

Jérémy L'Hour (*), Corentin Trevien (**)

(*) Insee (SSP Lab) et Ensae-Crest
(**) Sdes et Ensae-Crest

corentin.trevien@developpement-durable.gouv.fr

Mots-clés : Modèle de durée, prédiction, machine learning, données administratives.

Domaines : intégration des données (signes de vie), data science (machine learning)

Document provisoire,
version définitive disponible ultérieurement

Résumé

Le Sdes, SSM du ministère de la Transition écologique, a entrepris en 2019 la refonte du répertoire statistique des véhicules routiers (RSVERO). Principale innovation de ce projet, l'utilisation des données de contrôles techniques permet de s'assurer que les véhicules immatriculés sont toujours en circulation et de déterminer leur utilisation annuelle, grâce au relevé du compteur kilométrique effectué à chaque visite. Cependant, l'intégration de ces données n'est pas sans poser de questions méthodologiques : les visites interviennent à des dates variables, parfois avec du retard, et il n'est jamais certain qu'un véhicule repassera une visite de contrôle dans le futur, il existe donc une incertitude sur le fait qu'il soit toujours en circulation, notamment en cas de retard.

Cette document de travail présente les conclusions d'une collaboration du Sdes et du SSP Lab permettant d'estimer, pour toutes les voitures du répertoire, la probabilité qu'elles soient toujours en circulation à une date donnée et, le cas échéant, la distance annuelle parcourue. Plus largement, ce projet s'inscrit dans le mouvement de développement de méthodologies de traitement des sources administratives, que la mission conjointe de la DMCSI et de l'IG de l'Insee de 2019 « L'exploitation généralisée des données administratives, un changement significatif pour l'Insee » avait appelé de ses vœux.

1 Introduction

Le Service des données et études statistiques (Sdes), service statistique du ministère de la Transition écologique, est en charge du répertoire statistique des véhicules routiers (RSVERO), qui recense la totalité des voitures, deux-roues motorisés, utilitaires, poids louds, cars, bus, etc. immatriculés en France. La nouvelle version du RSVERO, lancée fin 2019, est issue du rapprochement :

- Du système d'immatriculation des véhicules, administré par l'agence nationale des titres sécurisés (ANTS), contenant de nombreuses informations techniques sur chaque véhicule (émissions de CO_2 théoriques, puissance, carburant, etc.), son utilisateur (propriétaire et ou locataire longue durée). Ce système enregistre l'ensemble des opérations administratives liées au certificat d'immatriculation comme la mise en circulation, le changement d'adresse du propriétaire, la vente ou la destruction du véhicule.
- Des résultats des contrôles techniques, centralisés par l'Utac. Ces données contiennent la date de la visite le type de contrôle (véhicule particulier, poids-lourd, taxi, etc.), le kilométrage du véhicule et le résultat du contrôle (succès, passage d'une contre-visite, etc.), et ce pour chaque contrôle. Outre l'estimation de la distance parcourue entre deux contrôles techniques, cette deuxième source permet de s'assurer que les véhicules immatriculés circulent toujours, les opérations correspondant au retrait du véhicule de la circulation étant imparfaitement enregistrées par l'ANTS. Le contrôle technique étant obligatoire à intervalles réguliers (par exemple tous les deux ans à partir du quatrième anniversaire pour une voiture), on peut considérer qu'un véhicule n'est plus en circulation quand le propriétaire n'effectue plus cette démarche.
- Du répertoire Sirène des entreprises pour les véhicules utilisés par des entreprises.

Les informations du contrôle technique ne peuvent être directement utilisées pour établir les données annuelles sur le parc. Les visites interviennent en cours d'année et la dernière en date est souvent antérieure à la date à laquelle on souhaite arrêter les données. L'objectif de l'expérimentation est donc de tester plusieurs méthodes statistiques permettant d'estimer l'utilisation annuelle des véhicules et leur probabilité d'être toujours en utilisation au 1er janvier de chaque année. Les méthodes actuellement utilisées au Sdes, fondées sur des ratios et des moyennes de strates, fournissent des résultats agrégés fiables. En revanche, elles sont moins performantes pour reproduire l'hétérogénéité individuelle entre véhicules, notamment avant la première visite. L'expérimentation a donc pour objectif d'estimer ces deux variables pour les deux types de véhicules qui sont à la fois les plus répandus et pour lesquels les contrôles techniques obligatoires sont les plus espacés : les voitures et les véhicules utilitaires légers (VUL). La modélisation économétrique n'a en revanche pas été appliquée aux autres types de véhicules, le temps imparti pour cette expérimentation ne l'ayant pas permis. En outre, l'estimation des modèles sur les VUL est encore en cours et n'a pas été reportée dans cette version du document de travail, qui porte donc exclusivement sur les voitures.

Appartenance au parc des véhicules routiers. Il existe, par nature, une incertitude sur le fait qu'un véhicule passera ou non un contrôle technique après la dernière visite observée. Dans le souci produire des statistiques sur le parc des véhicules sans attendre que l'ensemble des véhicules ait passé une visite, tout en intégrant un retard raisonnable, il est nécessaire d'estimer la probabilité de cet événement.

On s'appuie pour cela sur une modélisation économétrique de la durée écoulée entre une visite et la suivante (voir Wooldridge (2001)). Celle-ci présente trois particularités, elle utilise tout d'abord sur une fonction de hasard de base non-paramétrique, à même de rendre compte

du fait que la survenue du contrôle technique est très regroupée autour de la date théorique de passage. Elle utilise également une méthode de sélection des variables explicatives (potentiellement nombreuses) de type elastic-net. Elle prend enfin en compte l'existence de « survivants de long terme » (Maller and Zhou, 1996 ; Zhao and Zhou, 2006), c'est-à-dire les véhicules qui ne passeront plus jamais de contrôle technique. Ce type de durées est aussi connu en économétrie sous le nom de modèle de *durée de vie défective* (Abbring, 2002). La durée entre l'ultime contrôle technique d'un véhicule et le suivant est par définition infinie, puisque, par définition, ce dernier ne survient jamais. Or les modèles de durée de vie standard supposent a priori une distribution de probabilité qui exclut l'existence d'une durée de vie infinie. Il est donc nécessaire d'adapter la modélisation pour prendre en compte ce phénomène.

Utilisation annuelle des véhicules routiers. La détermination de la distance annuelle parcourue par un véhicule est relativement simple lorsque l'on dispose de relevés kilométriques avant et après l'année considérée (hors période exceptionnelle de type Covid). Cette distance doit en revanche être estimée, en fonction de son utilisation passée, des caractéristiques du véhicule et de son conducteur, quand le dernier relevé kilométrique est antérieur à la fin de l'année considérée. C'est par exemple le cas d'un véhicule pour lequel souhaite déterminer la distance parcourue en 2021, dont la dernière visite de contrôle date de 2020. Sans cette estimation, il faudrait attendre que la totalité des véhicules aient passé une visite pour diffuser des statistiques sur l'utilisation du parc, c'est-à-dire cinq ans pour les voitures, en tenant compte des visites passées avec un retard limité.

Cette deuxième série de modèle vise à prédire la distance annuelle moyenne parcourue par un véhicule entre deux contrôles techniques. La méthodologie développée s'inspire largement des approches standard d'apprentissage statistique pour les problèmes de régression (Hastie et al., 2001) estimé à l'aide d'algorithmes classiques de machine learning : MCO, régression pénalisée de type elastic-net ou LASSO (voir L'Hour (2020) pour une introduction), régression quantile, gradient boosting machine.

L'approche finalement sélectionnée, développée après une phase initiale où des approches standard (mono-modèles) ne se sont montrées pas totalement satisfaisantes, repose sur une combinaison de modèles. En effet, la distance parcourue par le véhicule dans le passé, observée lors des visites antérieures, présente un pouvoir explicatif tellement fort qu'elle brouille le signal contenu dans les autres variables explicatives. La prédiction finale est donc la combinaison linéaire de deux sous-modèles : le premier repose sur l'utilisation des caractéristiques du véhicule et de l'utilisateur ; le second n'utilise que les distances parcourues avant la dernière visite (ou une valeur imputée quand la première visite n'a pas eu lieu). Les poids optimaux en terme de prédiction pour cette combinaison linéaire sont estimés via une régression quantile pour limiter l'influence des valeurs aberrantes.

La section 2 présente la structure des bases de données utilisées, ainsi que quelques statistiques descriptives. La section 3 présente l'estimation de la probabilité d'appartenance au parc, et la modélisation de la durée entre le passage de deux contrôles techniques. La section 4 présente la stratégie de prédiction des kilomètres annuellement parcourus.

2 Données

2.1 Données

Plusieurs tables du répertoire statistique des véhicules routiers ont été appariées pour obtenir les données nécessaires à l'expérimentation :

- la table `vehicule` contenant la date de mise en circulation ainsi que diverses caractéristiques techniques (carburation, poids, émissions de CO2, etc.) de chaque voiture existante ou détruite ;
- la table `ct` (pour contrôle technique), contenant la date, le résultat, la catégorie (ordinaire, taxi, auto-école, collection, etc.) et le relevé kilométrique de chaque visite ;
- la table `operation` contenant les différentes opérations affectant le certificat d'immatriculation, notamment les changements d'utilisateur du véhicule et les opérations de sortie de parc (destruction, vente à l'étranger, etc.) ;
- la table `utilisateur` contenant les informations sur les utilisateurs successifs du véhicule, notamment leur lieu de résidence ou leur type (professionnel ou particulier).

Les données finales présentent une ligne pour chaque contrôle technique. Elles contiennent les informations suivantes :

- les caractéristiques techniques du véhicule ;
- la date de la visite de contrôle ainsi que celle de la visite suivante (pour simplifier les traitements, la date de mise en circulation du véhicule est assimilée à une visite de contrôle « zéro », avec un relevé kilométrique nul) ;
- la distance journalière moyenne parcourue entre ces deux dates ;
- le cas échéant, la distance journalière moyenne passée, parcourue entre la visite considérée et la visite précédente ;
- le type et le succès de la visite ;
- les informations sur l'utilisateur du véhicule à la date de la visite (commune de résidence avec zonage Insee associé, professionnel ou particulier, changement de main du véhicule entre deux visites).

L'expérimentation est menée sur 2 % seulement des données, pour permettre des temps de calculs raisonnables. Les observations sélectionnées sont séparée entre :

- un échantillon d'estimation, constitué, sauf tests de robustesse, de visites passées en 2018 ;
- deux échantillons de test, le premier contemporain de l'échantillon d'estimation, le suivant composé de visites passées plus tard pour confirmer les capacités prédictives du modèle.

Les contrôles techniques des années 2020 et ultérieures ne sont pas utilisés pour la modélisation des distances kilométrique, la crise sanitaire ayant eu des conséquences sur la circulation automobile et la durée entre deux visites, qui nécessitent des traitements spécifiques hors du champ de ce document.

Les échantillons des types de modélisation – distance parcourue et durée entre deux visites – ne sont pas constitués de la même manière, chacun devant répondre à des contraintes spécifiques :

- pour l'estimation des distances, l'échantillon est constitué des visites passées en 2018, le modèle devant être utilisé pour estimer les distances annuelles de véhicules appelés à passer une visite dans un proche avenir, en utilisant les visites passées dans un passé récent ; il est expurgé d'observations manifestement aberrantes (durée entre deux visites dépassant la durée autorisée, augmenté d'un retard « raisonnable », distance annuelle moyenne excédant 200 000 km soit environ 550 km par jour) ;
- pour l'estimation des durées entre contrôles techniques, il est au contraire indispensable de conserver les véhicules présentant un fort retard ; les observations correspondent donc aux durées écoulées postérieurement à 2011, première année pour laquelle la remonté des

		Rang de la visite	
Année visite	Echantillon	Première	Suivante
2018	Apprentissage	18194	165007
2018	Test	18094	165292
2019	Test	19266	164109

TABLE 1 – Nombre d'observations des échantillons pour la modélisation des durées entre contrôles techniques

		Rang de la visite	
Statut visite	Echantillon	Première	Suivante
En attente (censuré)	Apprentissage	21903	82655
	Test	197844	746396
Passé	Apprentissage	25406	167989
	Test	230219	1519518

TABLE 2 – Nombre d'observations des échantillons utilisés pour la modélisation de la distance parcourue entre contrôles techniques

données des visites est à peu près complète. Les visites passées après le 31 décembre 2018 ne sont pas prises en compte pour l'estimation et les visites postérieures sont considérées comme des observations censurées. Les visites ayant eu lieu entre 2019 et 2021 sont en revanche utilisées pour tester les performances du modèle. Enfin, les durées *postérieures* aux visites ratées sont retirées de l'échantillon, une contre visite étant requise dans les deux mois dans ce cas de figure.

2.2 Statistiques descriptives

Les tableaux 1 et 1 présentent le nombre d'observations des échantillons d'apprentissage et de test, en distinguant celles qui concernent une première visite des autres. L'échantillon de test du modèle de durée a été volontairement augmenté pour contenir un nombre suffisant de véhicules en retard de plus de 3 mois mais de moins de 18 mois, vis-à-vis de leur contrôle technique. Ces observations sont centrales dans l'utilisation du modèle pour le calcul du nombre de véhicules en circulation (voir 3.6). Le tableau 3 présente les caractéristiques des deux échantillons utilisés pour les deux modélisations.

Variable	Durée	Distance
Nombre d'observations		
Nombre de visites	297953	183201
Nombre de véhicules	107872	157460
Âge à la visite précédente		
Moins de 5 ans	27.45	20.76
De 5 à 10 ans	29.13	31.47
De 10 à 15 ans	24.02	26.3
15 ans et plus	19.4	21.46
Carburation		
Autre ou inconnu	0.04	0.01
Diesel	63.16	65.25
Electrique	0.11	0.07
Essence	36.18	34.01
Gaz	0.4	0.57
Hybride rechargeable	0.11	0.09
Type d'utilisateur		
Inconnu	0.12	0.02
Particulier	90.16	92.09
Professionnel	9.72	7.9
Distance annuelle parcourue entre la visite et la précédente		
Moins de 5 000 km	16.23	21.81
De 5 000 à moins de 10 000 km	22.13	26.01
De 10 000 à moins de 15 000 km	20.26	22.87
De 15 000 à moins de 20 000 km	13.33	15.12
20 000 km et plus	11.43	14.19
Manquant ou Aberrant	16.62	0
Retard (statut au 31-12-2018 pour les durées censurées)		
Dans les temps	54.72	69.31
Un mois et moins	19.5	17.65
Un à trois mois	8.9	7.08
Trois mois à un an	8.31	5.97
Plus d'un an	8.57	0
Résultat de la visite		
En attente	12.38	0
Raté	15.6	15.87
Réussi	72.02	84.13

TABLE 3 – Caractéristiques des véhicules des échantillons d'entrainement des modèles (en %)

3 Appartenance au parc : contrôles techniques et sorties de parc

Cette seconde partie présente la modélisation de la sortie du parc d'un véhicule donné, en s'appuyant sur deux types d'événements :

— l'enregistrement par l'ANTS d'une opération entraînant l'arrêt de circulation du véhicule (mise à la casse, vol, sortie du territoire, véhicule non-réparable, etc.) ;
— l'absence de passage régulier de contrôles techniques réussis.

On notera que, si le premier type d'événement peut être observé, ce n'est pas le cas pour le second, par définition un non-événement. Une fois estimé, le modèle permet de calculer à une date donnée et pour chaque véhicule, la probabilité qu'un des deux événements signalant que le véhicule ne se déplace plus se produise dans le futur (sortie de parc ou non-passage d'un contrôle technique).

La prise en compte des contrôle technique pour la modélisation de la sortie du parc est rendue nécessaire par les lacunes du suivi de la sortie effective des véhicules du parc à travers les opérations enregistrées par l'ANTS. L'utilisation de cette source est complexe, du fait du retard avec lequel certains propriétaires présentent leur véhicule pour le contrôle périodique. La figure 1 présente la fonction de survie, calculée avec l'estimateur de Kaplan-Meier, de la durée entre deux contrôles techniques. Ces résultats confirment que :

— la plupart des voitures passent leur contrôle en temps et en heure, à savoir au 4^e anniversaire du véhicule puis tous les deux ans, mais les retardataires représentent une part non-négligeable ;
— la plupart des véhicules retardataires passent leur visite de contrôle dans les mois qui suivent la date de contrôle théorique ;
— certains véhicules ne passent plus de contrôle technique et leur proportion s'accroît avec l'âge du véhicule.

Notons que l'utilisation du modèle pour la production de statistiques sur le parc porte toujours, pour le moment, sur le passé proche. Il sera par exemple utilisé au 2e trimestre 2022 pour l'estimation du parc des véhicules en circulation au 1er janvier 2022. En conséquence, tous les événements de sortie de parc enregistrés par l'ANTS sont connus à la date d'intérêt, la remontée de ces informations dans le répertoire statistique des véhicules routiers ayant lieu en temps réel. Cela signifie que seule la survenue d'un contrôle technique dans le futur reste incertaine. La modélisation complète, intégrant les deux types d'événements, pourrait être utilisée pour des travaux complémentaires, par exemple le calcul de l'espérance de vie des véhicules, qui dépassent le cadre de l'expérimentation.

3.1 Modélisation de la survie : approche globale

Dans cette section, nous présentons l'approche complète pour modéliser l'appartenance au parc roulant, c'est-à-dire incluant la survenue d'un événement de sortie de parc, dans la mesure où cette présentation permet d'exposer précisément la modélisation de la durée entre deux contrôles techniques. Toutefois, seule cette dernière modélisation revêt un intérêt métier immédiat et sera développée dans la partie suivante, portant sur l'estimation du modèle.

Soit T la variable aléatoire positive qui mesure l'âge du véhicule (*i.e.* la durée écoulée depuis la mise en circulation du véhicule) à l'enregistrement d'un événement de sortie de parc (*i.e.* destruction du véhicule, sortie du territoire). Soit C_1 (resp. $C_2, \ldots$) la variable aléatoire positive mesurant l'âge auquel le véhicule passe son premier (resp. deuxième, ...) contrôle technique réussi. On note $C_{\ell(t)}$ l'âge au dernier contrôle technique passé lorsque l'on est à l'âge

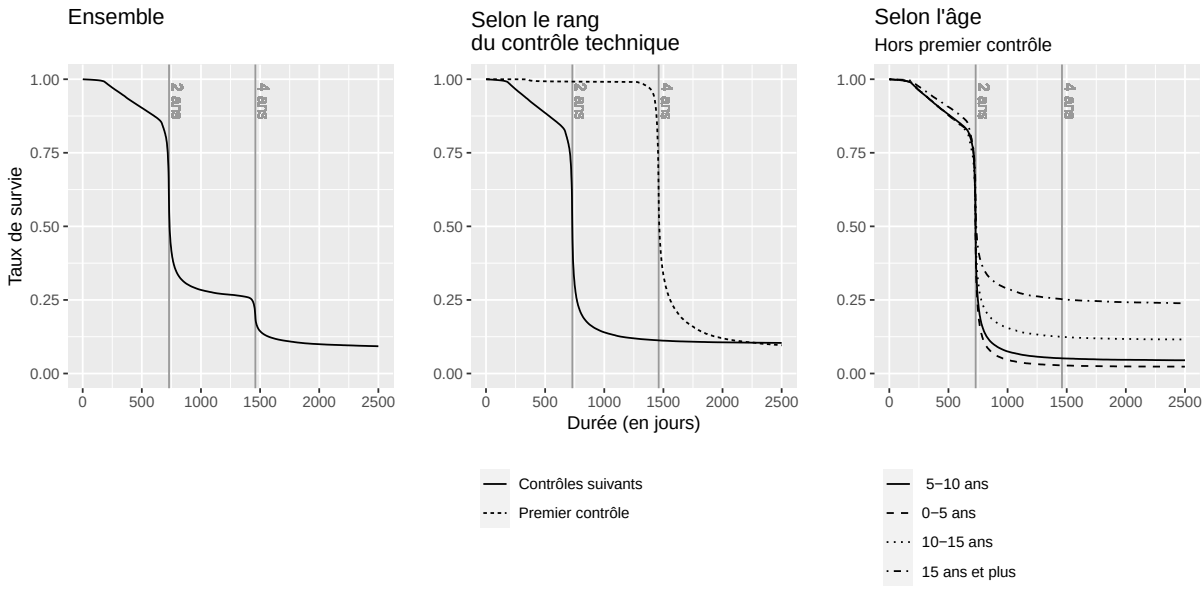

FIGURE 1 – Estimateurs de Kaplan-Meier de la durée entre deux contrôles techniques
Lecture : la probabilité qu'un véhicule passe son premier contrôle technique après 2000 jours (c'est-à-dire, sa probabilité de survie) s'élève à 12 %

t, c'est-à-dire $C_{\ell(t)} := \max\{c = C_1, C_2, \ldots : c \leq t\}$. Autrement dit : $\ell(t)$ est le nombre de contrôles techniques réussis à l'âge t. $C_{\ell(t)+1}$ désigne donc l'âge théorique du véhicule au contrôle technique à venir. Notons que l'événement $\{C_{\ell(t)+1} = \infty\}$ indique un contrôle technique qui ne se produit jamais et implique donc une sortie de parc. On utilise également la convention $C_0 = 0$.

Soit $R(t)$ la variable aléatoire binaire qui vaut 1 si le véhicule est dans le parc roulant à l'âge t et 0 sinon. L'événement $\{R(t) = 1\}$ se réécrit $\{R(t) = 1\} = \{T > t\} \cap \{C_{\ell(t)+1} < \infty\} = \{T > t\} \cap \{t < C_{\ell(t)+1} < \infty\}$. Autrement dit : aucune opération de sortie de parc n'a été enregistrée avant la date t et le prochain contrôle technique aura bien lieu à une date finie. L'événement contraire, celui de la sortie de parc, s'écrit $\{R(t) = 0\} = \{T < t\} \cup \{C_{\ell(t)+1} = \infty\}$.

Soit X un vecteur aléatoire correspondant les caractéristiques du véhicule [1]. On cherche à modéliser la probabilité d'être dans le parc roulant à la date t pour un véhicule sachant ses caractéristiques X et son historique de contrôles techniques, c'est à dire :

$$\begin{aligned} P\left[R(t) = 1 | X, C_1, \ldots, C_{\ell(t)}\right] &= P\left[T > t \cap t < C_{\ell(t)+1} < \infty | X, C_1, \ldots, C_{\ell(t)}\right] \\ &= P\left[T > t \cap t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty | X, C_1, \ldots, C_{\ell(t)}\right] \\ &= P\left[T > t | X, C_1, \ldots, C_{\ell(t)}, \left\{t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty\right\}\right] \\ &\times P\left[t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty | X, C_1, \ldots, C_{\ell(t)}\right]. \end{aligned}$$

On fait ici l'hypothèse que la durée avant l'enregistrement d'une opération de sortie de parc est indépendante de la durée écoulée avant le prochain contrôle technique réussi, conditionnellement à X ainsi qu'à l'historique des contrôles techniques réussis $C_1, \ldots, C_{\ell(t)}$, soit $TC_{\ell(t)+1} - C_{\ell(t)} X, C_1, \ldots, C_{\ell(t)}$. Cette hypothèse n'est pas forcément vérifié pour les véhicules

1. Dans les modèles plus complexes, X peut dépendre du temps, en incluant par exemple des informations sur l'utilisateur du véhicule.

abandonnés dans un hangar : ces véhicules n'enregistreront jamais de destruction ($T = \infty$), et ne passeront pas de contrôle technique ($C_{\ell(t)+1} = \infty$). Néanmoins, elle simplifie beaucoup l'estimation du modèle et importe peu pour la seule modélisation de la durée entre deux contrôles techniques. On suppose en outre que l'historique des contrôles techniques réussis n'apporte pas d'information sur la durée écoulée avant la destruction du véhicule ou sur le temps qui s'écoulera avant le prochain contrôle technique (sauf éventuellement quand cet historique est intégré aux variables aléatoires X). On a donc :

$$\begin{aligned} P\left[R(t) = 1|X, C_1, \ldots, C_{\ell(t)}\right] &= P\left[T > t|X, \left\{t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty\right\}\right] \\ &\times P\left[t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty|X, C_{\ell(t)}\right] \\ &= S_T(t|X, \left\{t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty\right\}) \\ &\times \left(S_C(t - C_{\ell(t)}|X, C_{\ell(t)}) - S_C(\infty|X, C_{\ell(t)})\right), \end{aligned}$$

où $S_T(.|X, \left\{t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty\right\})$ désigne la fonction de survie de T conditionnellement à Xet à $\left\{t - C_{\ell(t)} < C_{\ell(t)+1} - C_{\ell(t)} < \infty\right\}$ et $S_C(.|X, C_{\ell(t)})$ celle de la durée écoulée entre deux contrôles techniques. Nous allons estimer ces deux fonctions séparément, et donc modéliser séparément l'âge à la sortie de parc et la durée entre deux contrôles techniques.

Par simplification, nous adopterons la notation suivante : pour un véhicule donné et une date fixée t, $C_{-1} = C_{\ell(t)}$ désigne l'âge du véhicule au dernier contrôle technique, $C_{-2} = C_{\ell(t)-2}$ l'âge du véhicule à l'avant-dernier contrôle technique, et $C_{+1} = C_{\ell(t)+1}$ l'âge qu'aura le véhicule au prochain contrôle technique.

3.2 Censure

D'après la section précédente, estimer la probabilité d'appartenir au parc roulant à l'âge t, $P\left[R(t) = 1|X, C_{-1}, C_{-2}, \ldots\right]$ requiert deux quantités : la probabilité de n'avoir pas encore enregistré un événement de sortie de parc à l'âge t, et la probabilité de passer le prochain contrôle technique. La première est simple à estimer, pour la seconde il convient distinguer le premier contrôle technique des contrôles techniques suivants dans la mesure où, pour les voitures, la durée réglementaire entre les contrôles n'est pas identique.

Notons que l'on n'observe jamais T, l'âge du véhicule à la sortie de parc, pour les véhicules qui n'ont aucune opération de sortie de parc. On observe cependant, pour ces véhicules, une variable aléatoire O mesurant l'âge du véhicule à la date où l'on souhaite calculer le parc des véhicules en circulation (habituellement le 1er janvier). Par définition, O représente une borne inférieure de T. On notera la variable observée $Y = \min(T, O)$ et $D = \{T < O\}$. D indique que l'ANTS a enregistré une opération de sortie de parc concernant le véhicule. Il s'agit donc d'une observation qui n'est pas censurée. Pour chaque véhicule on observe le vecteur suivant : (Y, D, X). Les données sont constituées de n copies indépendantes de ce vecteur : $(Y_i, D_i, , X_i)_{i=1,\ldots,n}$.

Le problème de censure se présente également pour les contrôles techniques, puisque par définition, l'âge du véhicule au prochain contrôle technique n'est pas encore connu à une date donnée et constitue donc une durée censurée. Deux situations se présentent :

1. le véhicule n'a jamais passé de contrôle technique : dans ce cas on n'intègre, pour l'estimation du modèle, qu'une seule donnée (censurée, donc) : l'âge du véhicule à la date de clôture ;
2. les véhicules a déjà passé un contrôle technique, on intègre également au modèle l'historique des durées de passages entre les différentes visites de contrôles réussies.

Dans les deux cas, seuls les contrôles techniques réussis sont conservés, la durée pour passer une contre-visite en cas d'échec étant limitée (deux mois maximum) et inférieure à la durée écoulée

entre la date à la quelle on souhaite calculer le parc des véhicule en circulation (1er janvier) et le moment où ce calcul a lieu (2e trimestre). A noter que les observations non censurées sont nécessaires pour l'identification du modèle tandis que les observations censurées sont nécessaires pour définir la part des survivants à long terme, c'est-à-dire les véhicules qui ne passeront plus jamais de contrôle technique (cf. *infra.*).

3.3 Choix de modèle et estimation

Modèle à défaillance proportionnelle. Un modèle à défaillance proportionnelle (*Cox proportional hazard model*) est utilisé pour modéliser les trois durées décrites dans ce paragraphe (durée avant la sortie de parc, avant le premier contrôle technique ou entre deux contrôles techniques). Il consiste à décomposer le taux de défaillance instantané $h(t)$ [2], c'est-à-dire la probabilité infinitésimale d'occurrence de l'événement considéré (sortie de parc ou contrôle technique) à un instant t sachant que cet événement n'a pas encore eu lieu, en deux parties : une fonction de taux de défaillance de base $h_0(t)$, commune à tous les individus, modulée par un facteur multiplicatif $\exp(X_i'\beta_0)$, spécifique au véhicule i, selon ses caractéristiques observables X_i. Le taux de défaillance proportionnel s'exprime de la manière suivante :

$$h\left(t|X\right) = \exp\left(X'\beta_0\right) h_0(t),$$

Le modèle de Cox, largement répandu, a été ici retenu pour éviter, à la différence de modélisations reposant sur des lois exponentielles ou de Weibull, de faire des hypothèses paramétriques sur la fonction de survie. Il apparaît particulièrement adapté à la modélisation des contrôles techniques car, comme l'illustre la figure 1, on s'attend à ce que la fonction de survie de base décroisse très fortement autour de la date réglementaire de passage (alors que les modélisations plus simples la supposent constante ou régulièrement croissante dans le temps). En outre, le modèle de taux de défaillance proportionnel permet de décomposer l'estimation en plusieurs étapes puisqu'il n'est pas nécessaire d'estimer le taux de défaillance de base pour estimer le coefficient β_0.

Sortie de parc. On utilise le modèle de Cox avec prise en compte de la censure. On peut ensuite estimer la fonction de survie $S_T(.|X)$, qui s'exprime de la façon suivante :

$$S_T(t|X) = \exp\left(-e^{X'\beta_0}\int_0^t h_0(u)du\right) = S_0(t)^{\exp(X'\beta_0)}.$$

Contrôles techniques. On se propose d'estimer la distribution de la durée entre le dernier contrôle technique connu et le suivant $C_{+1} - C_{-1}|X$ en utilisant les durées écoulée entre les contrôles techniques précédents $(C_{n,i} - C_{n-1,i}, X_i)_{i=1,\ldots,n}$ avec $n \leq -1$ et les durées censurées écoulées depuis C_{-1}. A noter que sans la prise en compte des observations censurées, le modèle n'inclurait que des durées finies et cela exclurait nécessairement la possibilité de ne plus jamais passer de contrôle technique.

Cas particulier du premier contrôle technique. Pour les véhicules particuliers (hors taxis, véhicules de collection, auto-écoles, etc.), le premier contrôle technique intervient au quatrième anniversaire du véhicule, contrairement aux suivants qui doivent avoir lieu tous les deux ans. Il est donc incorrect de supposer que C_1 possède la même distribution que $C_2 - C_1, C_3 - C_2$, etc. (où l'on fixe $C_0 = 0$, l'âge à la mise en circulation, par convention), ce qui nécessite une modélisation séparée de C_1.

2. def : Pour une variable aléatoire à densité, on a $h(t) = f(t)/S(t)$ où $f(.)$ est la densité et $S(.)$ la fonction de survie. Pour plus de détails à un niveau introductif, voir par exemple Wooldridge (2001)

Pénalisation. Pour l'estimation de β_0, on utilise un estimateur de Cox avec une pénalisation de type *elastic-net* (Tibshirani, 1997) :

$$\min_{\beta} \sum_{r \in \mathcal{D}} X'_{i_r}\beta - \log\left(\sum_{i \in \mathcal{R}_r} \exp(X'_i\beta)\right) + \lambda\left[\alpha\left(\sum_{j}^{p} |\beta_j|\right) + (1-\alpha)\left(\sum_{j}^{p} \beta_j^2\right)\right],$$

pour $\alpha \in [0,1]$ et $\lambda > 0$. $\mathcal{D}$ est l'ensemble des durées distinctes entre deux événements observées dans les données et $\mathcal{R}_r$ l'ensemble des indices des véhicules pour lesquels aucun événement ne s'est encore produit à la durée $r \in \mathcal{D}$. La contrainte sur les coefficients permet d'éviter le sur-apprentissage, c'est-à-dire qu'elle permet d'avoir un modèle qui se généralisera mieux à de nouvelles données, non utilisées pour l'estimation du modèle. On fixe $\alpha = .5$ et λ est choisi par validation croisée. Il est à noter que la solution résultant de l'estimation sera parcimonieuse dans le sens où seuls un petit nombre d'éléments de $\widehat{\beta}$ seront différents de zéro. Pour une description en Français avec plus de détails de ces modèles pénalisés et de la validation croisée, le lecteur peut se référer à L'Hour (2020).

Fonction de défaillance de base. Étant donné que le but de ce travail est de prédire la probabilité qu'un événement survienne après une date donnée, il est nécessaire d'estimer la fonction de défaillance de base $h_0(t)$, une fois le coefficient β_0 estimé. On utilise pour cela l'estimateur de Breslow (1972), solution standard dans cette littérature, qui permet une modélisation totalement non-paramétrique de la fonction de survie.

Survivants de long-terme. Contrairement aux modèles de survie les plus simples, qui font l'hypothèse que l'événement modélisé finit toujours par se produire (par exemple, le décès), il est ici tout à fait possible qu'un véhicule ne passe jamais plus de contrôle technique, c'est-à-dire que l'événement $\{C_{+1} = \infty\}$ a une probabilité non nulle de se produire. La durée après le dernier contrôle technique du véhicule sera donc infinie, le suivant n'ayant jamais lieu, et le véhicule concerné sera considéré comme un *survivant de long-terme*. A noter que, du point de vue de la modélisation de la durée entre deux contrôles techniques, un survivant de long-terme peut être un véhicule qui a enregistré une sortie de parc. De fait, il ne passera plus de contrôle technique, et sera donc survivant de long-terme concernant le passage d'un contrôle technique.

La plupart des modèles paramétriques ne permettent pas d'avoir de survivants de long-terme, alors qu'il s'agit d'une caractéristique importante du phénomène. Notons que la probabilité qu'un véhicule passe un contrôle technique après un âge *fini* t est alors :

$$S(t|X) - S(\infty|X) = S_0(t)^{\exp(X'\beta_0)} - (1-p_0)^{\exp(X'\beta_0)},$$

avec $S_0(t) = \exp\left(-\int_0^t h_0(t)dt\right)$ la fonction de survie de base et $(1-p_0) = P[C_{+1} = \infty]$ la probabilité d'être un survivant de long-terme. On note donc qu'il faut soustraire la probabilité individuelle d'être un survivant de long-terme à la probabilité de survie standard pour obtenir la probabilité rigoureuse de passer un contrôle technique dans le futur. Dans le cas d'un modèle de défaillance proportionnelle, on peut estimer la probabilité d'être un survivant de long-terme directement via l'estimateur de Breslow. Cette méthode est justifié, par l'annexe A et par Zhao and Zhou (2006) pour les résultats théoriques.

3.4 Estimation : résultats

La figure 2 représente les deux fonctions de survie de base calculées via un estimateur de Breslow, distinguant le premier contrôle technique et les suivants. On remarque que la probabilité de ne jamais passer de premier contrôle technique (*i.e.* d'être un survivant de long-terme)

est de 7%, inférieure à celle observée pour les visites suivantes (8,7%).

La figure 3 représente le facteur multiplicatif de risque associé à chaque caractéristique du véhicule, comparé au niveau de base. Lorsque cette valeur est proche de 1, cela signifie que que la caractéristique n'a aucune influence en tant que facteur de risque [3].

3.5 Prédiction du passage d'un contrôle technique futur

Cette section présente les différentes métriques mobilisées pour évaluer la capacité du modèle à prédire correctement le passage d'un contrôle technique dans le futur. Pour rappel, les données sont constituées des durées écoulées après une visite réussie. Seules les visites comprises entre 2011 et 2018 sont prises en compte. Cela signifie que les visites survenues entre 2019 et 2021 n'ont pas été utilisées dans les estimations et peuvent donc servir à mesure les performances du modèle, selon deux critères :

- la classification, c'est-à-dire la capacité du modèle à répartir correctement les véhicules en deux catégories : ceux passant une visite dans le futur et ceux n'en passant pas ;
- la calibration, c'est-à-dire l'adéquation entre probabilité de passage d'une visite prédite par le modèle et la proportion empirique des véhicules passant effectivement un contrôle technique.

A noter que les visites postérieures à 2018 offrent seulement un recul de 3 années. On utilise donc le modèle pour prédire la probabilité de passage d'un contrôle technique dans les 3 ans et pas au delà.

Calibration La calibration consiste à estimer la probabilité empirique de passage d'un contrôle technique futur pour des véhicules appartenant à des intervalles fins calculés à partir d'une probabilité estimée par un modèle. Si le modèle est bien calibré, la proportion empirique des véhicules passant contrôle technique entre 2019 et 2021 sera autour de 1/8 pour des véhicules ayant une probabilité estimée proche de 1/8. Ainsi, plus la courbe correspondant à un modèle donné est proche de la première bissectrice, meilleure est la calibration du modèle.

La calibration est étudiée pour cinq types d'estimateurs de la probabilité de passer un contrôle technique dans le futur (voir figure 4) :

- La probabilité de passer un contrôle technique dans le futur telle qu'estimée par un modèle de durée (courbe verte pointillée),
- La probabilité de passer un contrôle technique dans le futur *conditionnellement au fait d'être encore dans le parc à la date de clôture,* telle qu'estimée par un modèle de durée (courbe verte pleine),
- Ces mêmes probabilités mais calculées sans variables explicatives (courbes mauves pointillées et vertes),
- La probabilité de passer un contrôle technique dans le futur telle qu'estimée par un modèle logit qui utilise la durée écoulée depuis le dernier contrôle technique (courbe pointillée orange).

Il est à noter que, contrairement aux quatre premiers, l'estimateur logit utilise des informations inconnues au 31/12/2018. Cet estimateur est donc incalculable en pratique, mais sert de comparaison pour souligner la pertinence d'une modélisation économétrique de la durée.

D'après la figure 4, on constate que la probabilité de survie conditionnelle offre la meilleure calibration, ce qui est tout à fait logique : pour deux véhicules en tous points identiques, la probabilité de passer une visite de contrôle entre 2019 et 2021 est nécessairement plus importante

3. Pour faciliter la lecture, on a directement retranché 1, afin qu'une valeur de zéro signifie l'absence d'effet détecté par le Lasso.

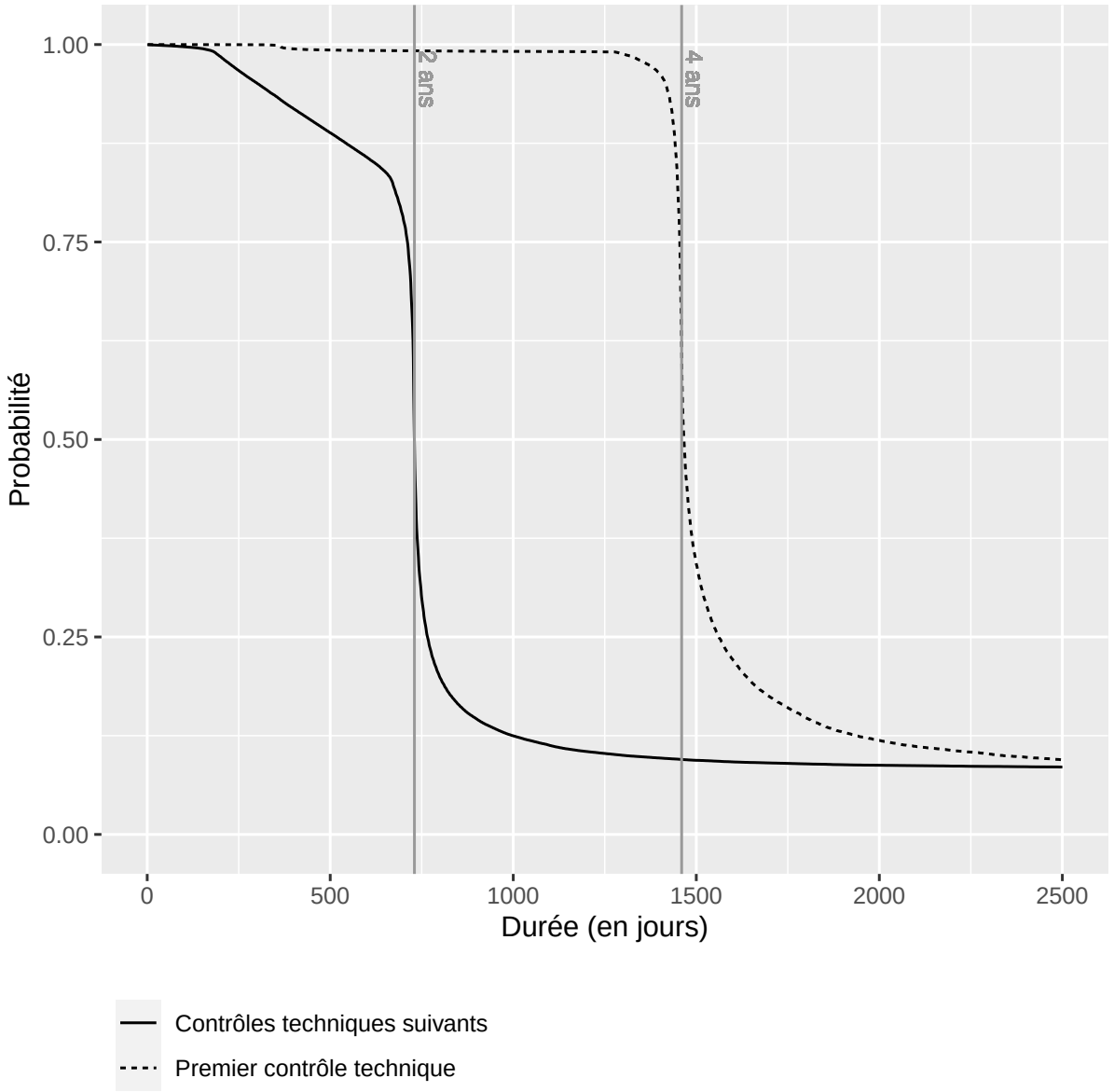

FIGURE 2 – Estimateur de Breslow pour la survie de base

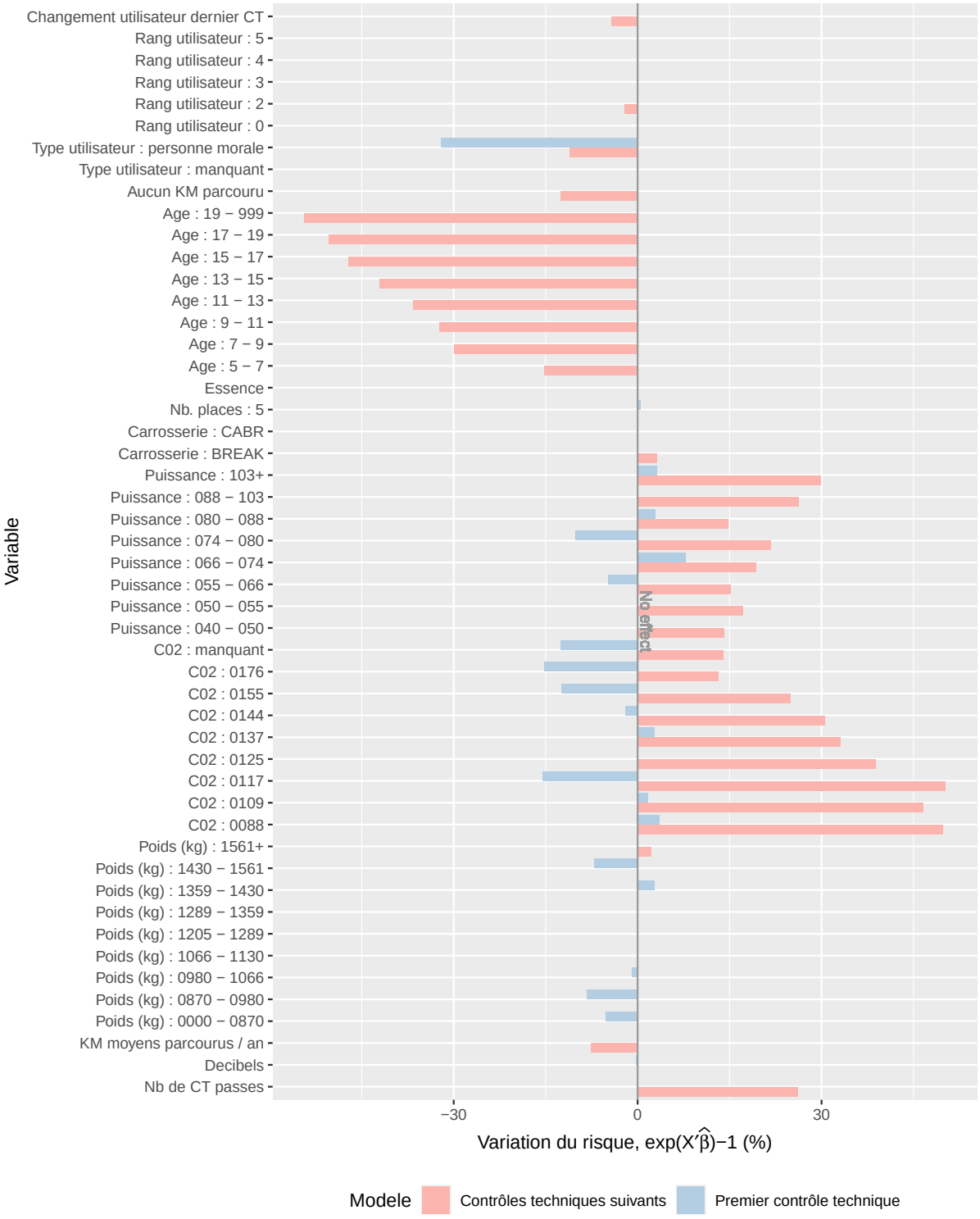

FIGURE 3 – Durée du contrôle technique : risques relatifs

Lecture : ce graphique représente le facteur multiplicatif de risque associé à chaque caractéristique du véhicule, comparé au niveau baseline en pourcentage i.e. la valeur de $\exp(\widehat{\beta}) - 1$. *Lorsque cette valeur est proche de 0, cela signifie que que la caractéristique n'a aucune influence en tant que facteur de risque. Par exemple, à partir du second contrôle technique, les véhicules âgés de 11 à 13 ans ont 35% de chance de moins de passer un prochain contrôle technique à chaque point du temps.*

pour un véhicule qui, 31/12/2018, vient de passer une visite, que pour un autre qui à la même date présente déjà deux années de retard. En revanche, l'utilisation des covariables ne semble pas substantiellement améliorer la prédiction. Enfin, le prédicteur logit se révèle inadapté.

Classification L'approche prédictive proposée ici se base sur des concepts standard utilisés dans des tâches de classification. Pour chaque individu, on souhaite prédire un événement de nature binaire. De façon générique, un algorithme simple consiste alors à construire une variable qui vaut 1 si l'on estime qu'un véhicule donné va passer un contrôle technique dans le futur et zéro sinon. On utilise pour cela un certain seuil s : si la probabilité estimée le dépasse on considère que le véhicule passera un contrôle technique dans le futur et inversement.

Pour évaluer la performance prédictive de cet algorithme, on considère généralement deux métriques : la précision et le rappel. La précision mesure la probabilité qu'un véhicule déclaré comme « survivant de long-terme » par l'algorithme ne passe pas de contrôle technique dans le futur (*i.e.* quand on prédit que l'événement ne se produira pas, quelle est la proportion de vrais négatifs, c'est-à-dire de véhicules pour lesquelles cette prédiction s'avère juste ?). Le rappel mesure la proportion de vrais positifs, c'est-à-dire de véhicules passant effectivement un contrôle technique dans le futur parmi ceux pour lequels une visite de contrôle a été prédite.

En amont du choix du seuil s, la courbe ROC (*Receiver Operating Characteristic*) permet d'évaluer l'arbitrage entre précision et rappel. Cette courbe ROC représente la performance du classifieur en comparant le vrais de faux positifs pour un taux de faux positifs donnés, calculé pour un seuil s donné. Plus s est grand, plus grande est la part des véhicules que l'on considère comme devant passer un contrôle technique dans le futur. La part des faux positifs s'accroît mécaniquement tandis que celle des faux positifs décroît. La figure 5 présente les courbe ROC des cinq prédicteurs utilisés pour la calibration. Contrairement à la courbe de calibration, plus la courbe ROC d'un classifieur est située au dessus de la première bissectrice, meilleur il est, car pour un taux de faux positifs donné, on obtient un taux de vrais positifs plus grand. Inversement, une courbe ROC qui proche de la première bissectrice indique qu'un classifieur est peu discriminant puisqu'une hausse de la part des vrais positifs s'accompagne d'une hausse équivalente du taux de faux positifs. La figure 5 confirme ainsi la supériorité des modélisations conditionnelles, qui seront utilisées en production.

3.6 Calcul du parc en circulation

Pour un véhicule donné, l'utilisation du modèle en production vise à déterminer s'il en circulation ou non au 1er janvier d'une année donnée. Cette utilisation porte uniquement sur la période postérieure au dernier contrôle technique connu puisque, pour les périodes antérieures, il n'existe par d'incertitude sur la durée entre visites de contrôle. Les prédictions découlant du modèle portent donc uniquement sur le passé proche, elles sont donc par nature provisoires et ont vocation à être replacées par des données définitives. Les données sur le parc doivent, dans la mesure du possible, remplir deux conditions :

— les données définitives doivent être disponibles dans un délai raisonnable ;
— l'écart entre données provisoires et définitives doit être limité.

La production de données définitives dans un délai raisonnable impose de définir un retard maximal, au delà duquel un véhicule n'est plus considéré comme appartenant au parc. En l'absence d'un tel critère, les données sont théoriquement révisable sans limite dans le temps. Par exemple, un véhicule n'ayant plus passé de contrôle technique après 2010 aura, à partir 2013 (*i.e.* au bout d'un an de retard), une probabilité faible d'être présenté à une visite dans le futur.

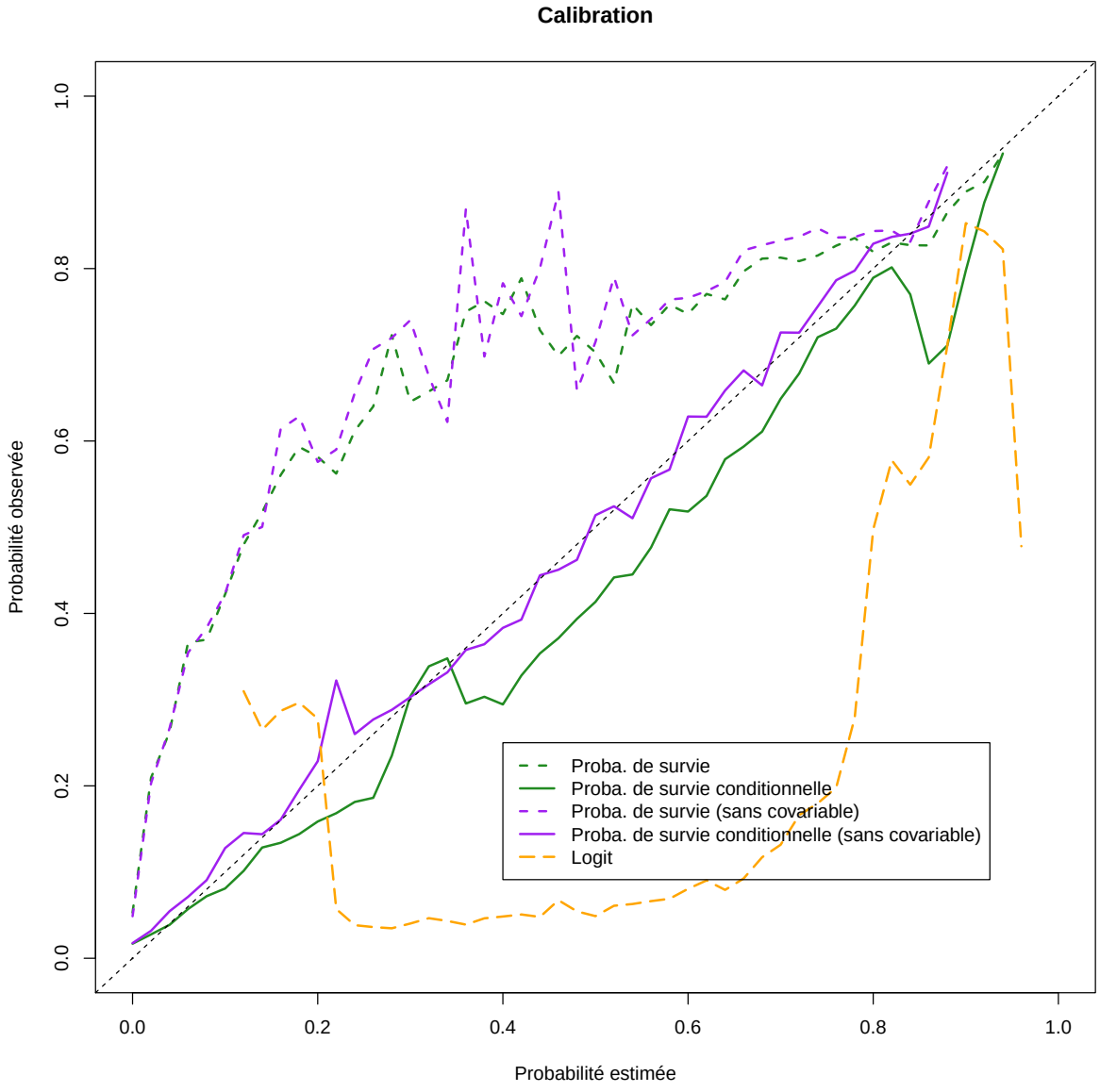

FIGURE 4 – Calibration

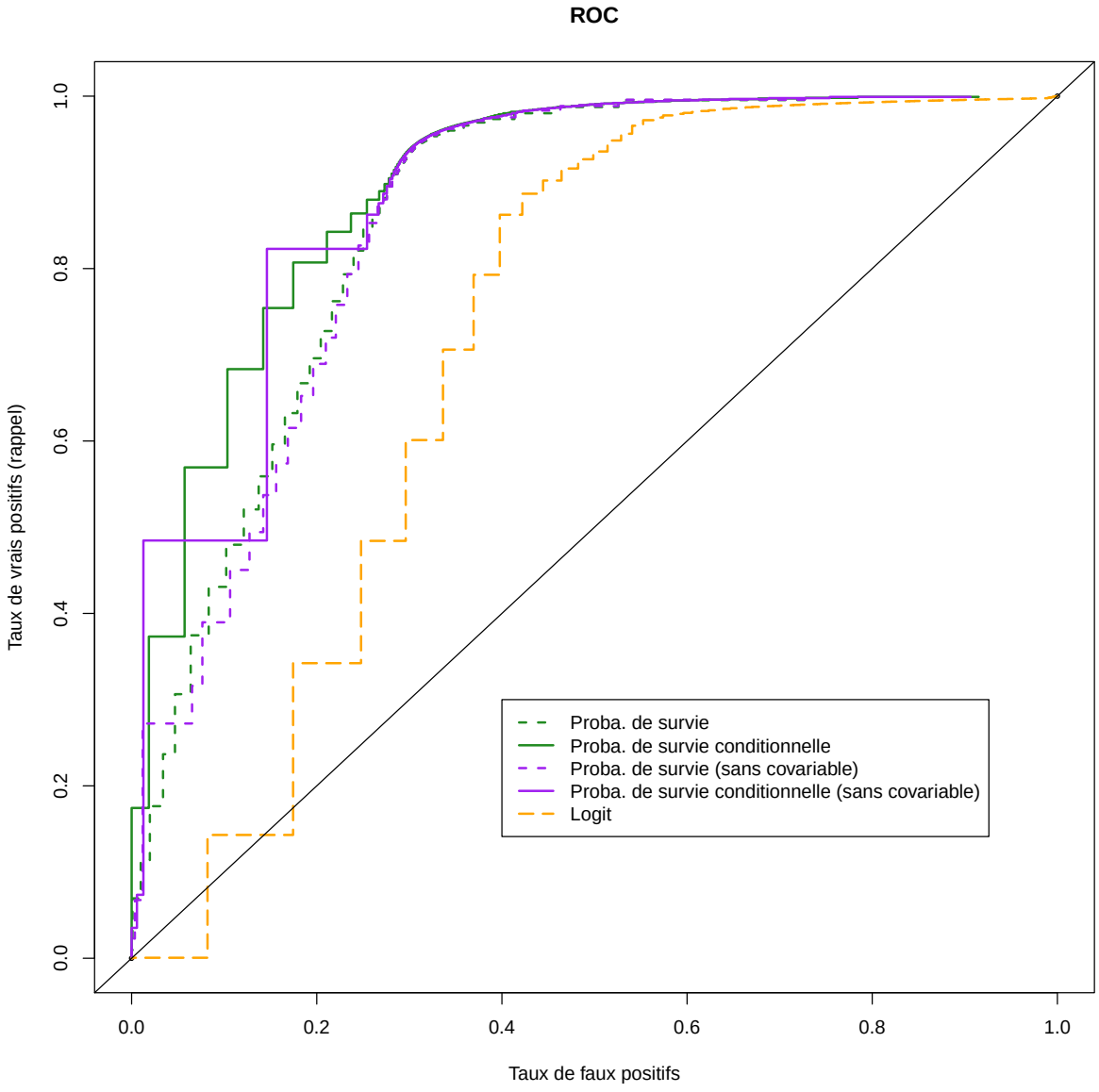

FIGURE 5 – Courbe ROC

Imaginons qu'il passe à nouveau un contrôle technique en 2019, son appartenance au parc devra être alors réévaluée des millésimes 2013 à 2018. Pour éviter cet écueil, le modèle devra être utilisé pour prédire la probabilité de passer un contrôle technique non pas à n'importe quel point du futur mais avec un retard raisonnable maximal. Autre argument en faveur d'un retard maximal, il est tout à fait possible qu'une période prolongée sans contrôle technique soit le signe de l'arrêt temporaire de l'utilisation du véhicule. Le retrait du parc durant la période concernée est donc, dans ce cas, justifié. Enfin, comme l'ensemble de l'approche fondée sur l'utilisation des contrôles techniques, il n'est pas possible de prendre en compte la fraude, qui peut expliquer l'absence temporaire de visites de contrôle.

Retard	2015	2016	2017	2018	2019
Dans les temps	51.50	51.00	53.30	54.50	47.40
Moins de 3 mois	34.60	35.00	32.80	32.10	36.50
3 à 6 mois	5.50	5.60	5.50	5.30	7.20
6 à 12 mois	4.70	4.80	4.70	4.60	5.30
12 à 18 mois	2.30	2.30	2.20	2.20	2.40
18 à 24 mois	1.40	1.30	1.40	1.40	1.30

TABLE 4 – Répartition des voitures selon l'année et la durée entre deux contrôles techniques (en %)

Le tableau 4 détaille, au 1er janvier de chaque année, la répartition des véhicules selon la durée entre deux visites de contrôle. La part des véhicules passant une visite avec un retard inférieur à 3 mois est très importante. Cela justifie, entre autres arguments, d'attendre le 31 mars pour établir le parc au 1er janvier de l'année correspondante. Il est ainsi possible de limiter l'incertitude sur le passage d'une visite dans le futur aux retards supérieurs à 3 mois, beaucoup moins fréquents. Le tableau 5 confirme qu'en suivant cette méthode, la part des véhicules dont l'appartenance au parc est incertaine s'avère très limitée. En 2015, le statut de seulement 39 188 véhicules était incertain sur un total de 803 329 véhicule potentiellement inclus dans le parc, c'est-à-dire présentant un retard de moins de 12 mois. Ceci permet de réduire l'écart entre données observées et estimées, quand on s'intéresse au nombre total de véhicules en circulation, qui n'est plus que de 0,24 % en 2015 contre 18,14 % sur les seuls véhicules au statut incertain.

Parc au 01/01	Appartenance	2015	2016	2017	2018	2019
Potentiel	Incertaine	39188	39783	40286	41139	45870
Observé		9398	9526	9861	9443	9242
Estimé		11103	11334	11444	11764	13183
Ecart (en %)		18.14	18.98	16.05	24.58	42.64
Potentiel	Ensemble	803329	814485	826956	840242	852431
Observé		775574	786128	798545	810674	817554
Estimé		777431	788135	800343	813051	822031
Ecart (en %)		0.24	0.26	0.23	0.29	0.55

TABLE 5 – Nombre de véhicules appartenant au parc selon le retard maximal admis avant retrait de la circulation

Note : Retard maximal autorisé fixé à 12 mois ; parc estimé avec les informations disponibles au 31 mars de l'année considérée, en pondérant les véhicules en retard selon leur probabilité de passage d'un contrôle technique futur ; véhicules retirés de la circulation comptés dans le parc jusqu'à la fin de validité de leur dernier contrôle technique.

On remarque également l'effet probable du Covid pour l'année 2019. En effet, une part importante des véhicules n'a pas passé le contrôle technique dans les temps à cause du très fort ralentissement de l'activité des centres de contrôle de mars à mai 2020.

Le tableau 6 détaille l'effet de la durée du retard maximal choisi sur le nombre de véhicules qu'on considère en circulation. L'impact de ce retard s'avère assez limité. Logiquement, son accroissement augmente l'écart entre données provisoires et définitives [4]. Compte tenu de ces éléments, un retard maximal d'un an semblant offrir un bon arbitrage entre délai de production et risque d'exclure du parc un nombre trop important de véhicules en circulation.

Parc au 01/01	Retard	2015	2016	2017	2018	2019
Observé	6 mois	763126	773478	785280	798188	805811
Estimé		763350	773687	785509	798142	806345
Ecart (en %)		0.03	0.03	0.03	-0.01	0.07
Observé	12 mois	770430	781088	793123	805830	812778
Estimé		771344	782079	794097	806754	815173
Ecart (en %)		0.12	0.13	0.12	0.11	0.29
Observé	18 mois	775574	786128	798545	810674	817554
Estimé		777431	788135	800343	813051	822031
Ecart (en %)		0.24	0.26	0.23	0.29	0.55

TABLE 6 – Nombre de véhicules appartenant au parc selon le retard maximal admis avant retrait de la circulation

Note : Parc estimé avec les informations disponibles au 31 mars de l'année considérée, en pondérant les véhicules en retard selon leur probabilité de passage d'un contrôle technique futur ; véhicules retirés de la circulation comptés dans le parc jusqu'à la fin de validité de leur dernier contrôle technique.

Deux approches sont envisageables pour l'utilisation du modèle en production :

— le calcul d'une variable dichotomique, prenant uniquement les valeurs 0 et 1, en fonction d'un certain seuil de probabilité s : les véhicules en deçà de ce seuil se voient affecter la valeur 0 et sont considérés hors du parc et inversement ;

— l'utilisation comme pondération de la probabilité prédite par le modèle : moins il est probable qu'un véhicule passe une visite de contrôle dans le futur, moins il compte dans les statistiques du parc.

Dans le cas d'une variable dichotomique, on choisit un seuil qui équilibre le nombre de faux positifs et de faux négatifs pour limiter les révisions à la hausse ou à la baisse une fois les données définitives. Le seuil est fixé à 0,323 en utilisant les données de 2015, il est maintenu ensuite. Les chiffrages basés sur l'utilisation d'un seuil présentent naturellement un écart plus faible avec les données définitives, le choix du seuil permettant une correction à posteriori d'un éventuel décalage systématique dans les prévisions. Cependant, la méthode fondée sur l'utilisation d'une pondération a été privilégiée pour la production. En effet, la méthode du seuil ne s'avère pas plus robuste au changement de comportement, comme celui observé en 2019, et elle entraîne, qui plus est, l'exclusion systématique des véhicules présentant une probabilité faible de passage d'un contrôle technique futur.

4. Sachant que le parc au 1er janvier est estimé en fonction des informations disponibles au 31 mars, l'incertitude porte uniquement sur les véhicules présentant un retard compris entre 3 et 6 mois au 1er janvier quand cette durée est fixée à 6 mois. Si ce retard maximal est porté à 12 mois, les véhicules présentant un retard compris entre 6 mois et 1 an sont, cette fois-ci, susceptibles d'appartenir au parc.

Parc au 01/01	Méthode	2015	2016	2017	2018	2019
Observé		770430	781088	793123	805830	812778
Estimé	Pondération	771344	782079	794097	806754	815173
Ecart (en %)		0.12	0.13	0.12	0.11	0.29
Estimé	Seuil	770453	781366	793122	806094	814631
Ecart (en %)		0	0.04	0	0.03	0.23

TABLE 7 – Nombre de véhicules appartenant au parc selon la méthode de calcul de l'appartenance au parc

Note : Retard maximal autorisé fixé à 12 mois ; parc estimé avec les informations disponibles au 31 mars de l'année considérée ; véhicules retirés de la circulation comptés dans le parc jusqu'à la fin de validité de leur dernier contrôle technique.

Autre difficulté de l'utilisation du modèle en production, alors qu'en cas de destruction d'un véhicule, la date de sortie du parc est précisément définie, il est impossible de dater, même rétrospectivement, la date où il cesse de circuler lorsqu'il n'est plus présenté aux visites de contrôle régulières. En l'absence de données complémentaires, deux hypothèses sont envisageables concernant la durée pendant laquelle un véhicule continue de circuler après son dernier contrôle technique :

— tant qu'il est autorisé en circulation, c'est-à-dire jusqu'à la date théorique du contrôle technique suivant ;
— immédiatement après la dernière visite.

La seconde hypothèse est assez peu crédible, le passage du contrôle technique étant inutile si le propriétaire du véhicule cesse de l'utiliser immédiatement après la visite, elle n'est donc pas retenue. Le tableau 8 présente l'impact très important du choix de la date de retrait de circulation sur le nombre de véhicules qu'on considère appartenir au parc. Cet écart est d'environ un dixième, ce qui correspond peu ou prou à la part des véhicules qui ne passent plus de contrôle technique après une visite donnée. En outre, la différence entre le parc observé, une fois toutes les visites enregistrées, et le parc estimé avec les informations disponibles au 31 mars d'une année donnée, est structurellement plus importante avec la deuxième méthode. Dans ce cas, l'appartenance au parc s'avère incertaine pour l'ensemble des véhicules puisque le passage d'un contrôle technique dans le futur n'est jamais certain. Enfin, on remarque l'effet important du Covid sur l'estimation du parc au 1er janvier 2019 avec la seconde méthode. Certains véhicules déjà en retard auraient passé leur visite pendant le première confinement, mais la fermeture temporaire des centres de contrôle a entraîné l'allongement de ce retard au delà d'un an. Les véhicules concernés se voient donc retirés du parc pour les 5 années antérieures s'il s'agit de leur première visite (4 années autorisées + 1 an de retard) et les 3 années antérieures pour les autres, en l'absence de modification des règles habituelles.

Parc au 01/01	Retrait	2015	2016	2017	2018	2019
Observé	Fin de contrôle valide	771344	782079	794097	806754	815173
Estimé		770430	781088	793123	805830	812778
Ecart (en %)		0.12	0.13	0.12	0.11	0.29
Observé	Immédiat	669448	678942	682565	678762	639113
Estimé		665375	673671	686102	697736	707179
Ecart (en %)		-0.61	-0.78	0.52	2.8	10.65

TABLE 8 – Nombre de véhicules appartenant au parc selon la date de retrait de circulation en cas de retard excessif

Note : Retard maximal autorisé fixé à 12 mois ; parc estimé avec les informations disponibles au 31 mars de l'année considérée, en pondérant les véhicules en retard selon leur probabilité de passage d'un contrôle technique futur.

4 Utilisation du parc roulant : distances annuelles parcourues

Cette partie décrit la modélisation de la distance parcourue annuellement par une voiture. Elle s'appuie sur les relevés kilométriques effectués à chaque contrôle technique, enregistrés dans le RSVERO. L'objectif est de prédire une distance la plus proche de la réalité à partir des caractéristiques du véhicule et de son utilisateur ainsi que de sont utilisation passée.

Pour une année donnée, cette distance n'est observée que si l'on dispose de relevés kilométriques antérieurs au 1er janvier et postérieurs au 31 décembre. En faisant l'hypothèse, raisonnable hors période de pandémie, que l'utilisation d'un véhicule est constante dans le temps, la distance parcourue au cours de l'année considérée correspond à la distance journalière moyenne entre les deux visites de contrôle, multipliée par 365 ou 366 jours. Cet exercice d'interpolation peut éventuellement être partagé en divers sous-périodes si une ou plusieurs visites supplémentaires ont eu lieu pendant l'année d'intérêt.

La modélisation développée ici permet d'obtenir une estimation de la distance annuelle parcourue quand cette information n'est pas observée. C'est notamment le cas pour les :

- **années postérieures au dernier contrôle technique** ; sans le recours à la modélisation, il faudrait attendre que la totalité des véhicules aient passé une visite pour diffuser des statistiques sur l'utilisation du parc, c'est-à-dire cinq ans pour les voitures (en tenant compte des visites passées en retard) ;
- **valeurs aberrantes ou douteuses** ; c'est le cas lorsque les relevés kilométriques indiquent une diminution de la valeur relevée au compteur entre deux visites, une distance annuelle moyenne supérieure à 200 000 km (soit plus de 500 km par jour) ou quand une durée excessive s'est écoulée entre deux visites (retard supérieur à un an). Le retrait des durées excessive correspond à la fois à une logique d'exclusion des valeurs douteuses (le véhicule a pu être utilisé à l'étranger ou de manière irrégulière entre les deux visites) mais également de production. En effet, la prise en compte de relevés kilométriques très espacés dans le temps conduirait à une révision des données au delà d'une durée raisonnable : par exemple, un véhicule qui passerait une visite en 2019 après n'en avoir pas passé pendant 8 années, verrait sa distance annuelle parcourue révisée des millésimes 2011 à 2018.

Contrairement aux données portant sur l'appartenance au parc, celles sur les distances annuelles ne sont donc jamais entières observées : c'est bien sûr le cas des des distances considérées comme aberrantes mais aussi de celles parcourues après la dernière visite d'un véhicule, qui pré-

cède son retrait du parc, qui restent par nature toujours inconnues.

De manière générale, les prédictions fournies par les différents modèles testés dans cette partie suivent deux objectifs :

— fournir une estimation non-biaisée de la somme des distances, afin de limiter au maximum les révisions entre les données provisoires, établies à partir des distances estimées, et les données définitives, calculées au bout de cinq ans, lorsque toutes les voitures présentes dans le parc ont passé une visite ;

— réduire l'hétégénéité individuelle inobservée en fournissant une prédiction non baisée et la plus précise possible de la distance annuelle parcourue par un véhicule donné.

Le premier objectif prime sur le second, et conduit à pondérer systématiquement les observations par la durée écoulée entre deux visites, quitte à obtenir une prédiction légèrement biaisées au niveau individuel.

4.1 Variable d'intérêt, mesure de la performance et évaluation des modèles

On souhaite idéalement modéliser la courbe kilométrique du véhicule à chaque point du temps. Notons que c'est une fonction nécessairement non-décroissante dans le temps puisque le compteur kilométrique est cumulatif.

Néanmoins, l'estimation directe de cette fonction est difficile la variance des kilomètres parcourus, importante au sein de la population, s'accroît avec la durée de vie du véhicule. D'autre part, l'objectif de cet exercice consiste seulement à déterminer les kilomètres parcourus pour une année donnée. On choisi donc de rapporter le nombre de kilomètres parcourus à la durée de circulation. Et, plutôt que de s'intéresser à la distance annuelle parcourue depuis la mise en circulation du véhicule, on se restreint à son utilisation récente, c'est-à-dire depuis le dernier contrôle technique. En effet, l'utilisation d'un véhicule est généralement plus intense au début de vie et décroît ensuite. L'indicateur finalement retenu pour cette partie est la différence entre les relevés kilométriques réalisés entre deux visites de contrôle successives, rapportée au nombre de jours écoulé entre ces deux dates.

Soit $K(t)$, le nombre total de kilomètres parcourus par un véhicule à l'âge t. On n'observe pas $K(t)$ à chaque point du temps mais uniquement aux dates de contrôles techniques $C_1, C_2, \ldots$ [5] Le fait d'observer le kilométrage uniquement à quelques points précis de la vie du véhicule ne pose pas de problème particulier pour l'estimation (*i.e.* pas d'effet de sélection) car on mesure l'information pour l'ensemble du stock. On peut donc employer des techniques standard d'apprentissage statistique.

On se place à une date de référence (par exemple le 31/03 d'une année donnée pour calculer l'utilisation du parc au cours des années précédentes) où le véhicule est âgé de t_i. Soit $K_{\ell(t_i)} := K(C_{\ell(t_i)})$ le dernier relevé kilométrique observé avant ou à cet âge. Dans un souci de simplification des notations, on notera $K_{-1,i}$ le relevé kilométrique au dernier contrôle technique pour le véhicule i, $K_{-2,i}$ le relevé kilométrique à l'avant-dernier contrôle technique, etc. par rapport à la date de référence. De même $K_{+1,i}$ est le relevé kilométrique au premier contrôle technique qui aura lieu dans le futur, qui, par définition, n'est pas encore observé à la date de référence. Pour chaque véhicule ayant passé c contrôles techniques, on observe donc le vecteur

5. Notons que pour les voitures n'ayant pas passé le premier contrôle technique, aucun kilométrage n'est observé, on est donc contraint de considérer que l'utilisation du véhicule est uniforme jusqu'au passage de la première visite.

suivant : $(K_{-1}, K_{-2}, \ldots, K_{-c}, X)$. Par convention, la mise en circulation est considérée comme le premier contrôle technique K_{-c} et le relevé kilométrique est égal à zéro. Le but de l'exercice est de prédire le nombre de kilomètres parcourus dans l'année précédent l'âge t_i. Comme on n'observe pas directement le nombre de kilomètres parcourus dans l'année entre le 01/01 et le 31/12 (sauf cas particulier), la qualité des modèles sera jugée sur la capacité à prédire les kilomètres moyens parcourus (en base journalière) entre le dernier contrôle technique et le premier contrôle technique dans le futur :

$$Y_i := \frac{K_{+1,i} - K_{-1,i}}{C_{+1,i} - C_{-1,i}}.$$

Remarquons que pour être au plus proche des conditions d'utilisation, le choix du modèle doit intégrer sa capacité à prédire une valeur future à partir d'observation passées. On ne peut pas se contenter d'une approche standard où les données en coupe sont partitionnées en échantillons d'entraînement / validation / test. Les performances du modèle sont donc évaluées de deux manières : d'une part, en comparant les distances prédites aux distances réelles sur un échantillon de test, contemporain mais différent de celui utilisé pour l'estimation, (ici les distances parcourues antérieurement aux visites ayant eu lieu en 2018) et d'autre part, en effectuant les mêmes comparaison sur un échantillon postérieur (ici les visites ayant eu lieu en 2019).

Les performances du modèle sont évaluées selon trois critères :

— l'erreur quadratique moyenne (RMSE) ;

— l'erreur absolue moyenne (MAE), moins sensible aux valeurs extrêmes de l'indicateur précédent, avec un modèle aboutissant à une prédiction $\widehat{Y}_i$, on évalue la perte à $|\widehat{Y}_i - Y_i|$, ce qui donne sur un échantillon de test de taille n :

$$\widehat{EAM} = \frac{1}{n}\sum_{i=1}^{n} |\widehat{Y}_i - Y_i|;$$

cette perte L_1 est plus robuste aux valeurs extrêmes que la perte L_2 (RMSE) utilisée plus habituellement ;

— le biais non pondéré, mesuré de la manière suivante :

$$\widehat{\text{Biais}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_i - Y_i;$$

— le biais total, portant non pas sur la distance journalière parcourue mais sur la distance totale entre deux visites de contrôle. Ce critère est un élément primordial pour l'évaluation du modèle, le niveau total de la circulation automobile sur année représentant une des statistiques les plus importantes issues du RSVERO :

$$\widehat{\text{Biais total}} = \frac{1}{n}\sum_{i=1}^{n} (\widehat{Y}_i - Y_i) \times D_i$$

Avec D_i la durée écoulée entre les deux contrôles techniques. Le biais non-pondéré, correspondant au niveau individuel, et le biais total diffèrent s'il existe une corrélation entre l'intensité d'utilisation du véhicule et l'espacement temporel entre visites.

Le RMSE et EAM sont normalisés par leur valeur pour une prédiction à la moyenne tandis que le deux biais sont normalisés respectivement par la moyenne non pondérée et la somme des distance.

4.2 Variables explicatives et séparation en sous-échantillons

Plusieurs spécifications sont testées avec un modèle linéaire simple pour mesurer l'effet du choix des variables et de la sélection de l'échantillon sur les performances du modèle .

	RMSE		*MAE*		*Biais normalisé*	
Variables explicatives	(brut)	(normalisé)	(brut)	(normalisé)	indiv.	total
	Echantillon test futur (visite passée en 2019)					
Âge du véhicule	24.75	0.96	16.45	0.911	-0.022	0.006
idem + carac. véhicule	23.68	0.919	15.43	0.855	0.005	0.021
idem + carac. utilisateur	23.48	0.911	15.18	0.841	0.004	0.02
idem + dist. passée	21.92	0.851	13.23	0.733	0.008	0.016
	Echantillon test en coupe (visite passée en 2018)					
Âge du véhicule	24.51	0.958	16.35	0.907	-0.026	0
idem + carac. véhicule	23.44	0.916	15.23	0.845	-0.016	-0.002
idem + carac. utilisateur	23.23	0.908	15	0.832	-0.015	-0.002
idem + dist. passée	21.64	0.846	13.03	0.723	-0.006	-0.002

TABLE 9 – Prédiction de la distance quotidienne moyenne parcourue entre deux contrôle techniques – Variables explicatives sélectionnées

Variables explicatives. Les variables incluses dans le modèle portent sur les caractéristiques du véhicule (type de carburant, âge au moment de la visite, puissance en kW, type de carrosserie, cylindrée, émissions théoriques de CO_2, norme Euro d'homologation du véhicule, nombre de décibels, nombre de places assises, poids à vide, marque) ou celles de son utilisateur (professionnel ou particulier, zonage en aire d'attraction des villes, degré de densité ou département du lieu de résidence), ainsi son utilisation passée (distance journalière parcourue entre avant et avant-avant-dernier contrôles techniques connus). Lorsqu'elles sont continues, ces variables sont discrétisées selon des seuils appropriés pour limiter l'influence des valeurs extrêmes et tenir compte de potentiels effets non-linéaires. Le tableau 9 détaille l'effet de l'inclusion de ces différentes variables sur les performances du modèle. Les résultats soulignent l'apport important de l'utilisation passée du véhicule. Ils indiquent également que le modèle est légèrement meilleur pour prédire les distances parcourues en 2018 que dans le futur, suggérant une variation dans le temps de la valeur des paramètres estimés.

Estimation sur sous échantillons. L'utilisation passé du véhicule s'avère être une variable fortement prédictive de son utilisation ultérieure. Lorsque cette variable est ajoutée au modèle, il s'avère que les coefficients associés aux autres variables diminuent nettement, indiquant qu'elles jouent le rôle de « proxy » du comportement idiosyncratique de l'utilisateur. Si l'ajout de cette variable améliore les performances prédictives du modèle pour l'ensemble des véhicules, il s'avère qu'il les *dégrade* sur certaines sous catégories : les véhicules neufs, dont l'utilisation passée est par nature inexistante, et les véhicules ayant changé de main, pour lesquels la variable existe mais s'avère naturellement nettement moins liée à l'utilisation future. Les estimations sont donc réalisées sur trois sous-populations différentes : (i) les véhicules n'ayant jamais passé de contrôle technique, (ii) les véhicules ayant déjà passé un contrôle technique et conservant le même utilisateur et (iii) les véhicules ayant déjà passé un contrôle technique mais ayant changé d'utilisateur juste avant le dernier contrôle technique. Le tableau 9 détaille l'amélioration des prédictions permise par ces estimations séparées.

Taille de l'échantillon Les estimations sont réalisées sur 1/100e des contrôles techniques ayant eu lieu au cours de l'année 2018. Le tableau 13(page 32 en annexe) indique qu'une réduction de la taille de l'échantillon augmente l'erreur des distances journalières prédites. Une analyse détaillée des ces prévisions indique la dégradation de la précision concerne surtout les premiers contrôles techniques, moins nombreux dans l'échantillon initial.

	RMSE		*MAE*		*Biais normalisé*	
Sous-échantillons	(brut)	(normalisé)	(brut)	(normalisé)	indiv.	total
Echantillon test futur (visite passée en 2019)						
Un	21.92	0.851	13.23	0.733	0.008	0.016
Deux	21.71	0.842	12.98	0.719	0.006	0.017
Trois	21.57	0.837	12.73	0.705	0.005	0.017
Echantillon test en coupe (visite passée en 2018)						
Un	21.64	0.846	13.03	0.723	-0.006	-0.002
Deux	21.44	0.838	12.78	0.709	-0.007	-0.003
Trois	21.31	0.833	12.55	0.696	-0.007	-0.003

TABLE 10 – Prédiction de la distance quotidienne moyenne parcourue entre deux contrôle techniques – Sous populations
Note : Deux sous-échantillons = première visite/visites suivantes ; trois sous-échantillons = première visite/visites suivantes sans changement d'utilisateur/visites suivantes avec changement d'utilisateur.

Ancienneté de l'échantillon Le tableau 14 (page 32 en annexe) présente les erreurs et biais des estimations selon l'année de passage des contrôles techniques utilisés pour estimer le modèle. Cinq échantillons d'estimation sont constitués sur les années 2014 à 2018 (le dernier correspond donc à celui utilisé jusqu'ici). Les distances sont ensuite prédites sur les échantillons de test 2018 et 2019. Il s'avère que la qualité des prédictions diminue un peu lorsque la durée écoulée entre les échantillons d'estimation et de test augmente. Ainsi, c'est l'échantillon 2018 qui fournit les meilleurs estimations pour les deux échantillons de tests. Ceci plaide pour une actualisation annuelle des coefficients de la modélisation.

4.3 Méthodes d'estimation

La partie précédente a permis d'identifier les variables utiles à l'estimation et la manière la plus efficace de sélectionner et segmenter l'échantillon d'estimations à l'aide méthodes linéaires simples. Plusieurs algorithmes plus complexes ont ensuite été testés pour améliorer les prédictions obtenus. Ces algorithmes sont quasiment tous encapsulés dans le package `R caret` afin d'avoir une approche générique et modularisée. Des procédures de partitionnement de l'échantillon entre entraînement et test sont mises en place lors de la sélection des hyper-paramètres, qui est réalisée par de la validation croisée :

1. Moindres carrés ordinaires (MCO) : cette méthode standard est très rapide puisqu'aucun paramètre de régularisation n'est à choisir, elle a été mise en œuvre dans la partie précédente et peut servir d'étalon.
2. *Elastic net* : il s'agit d'une régression linéaire pénalisée. La solution elastic net minimise la somme des carrés des résidus, plus un terme de pénalisation égal à $\alpha \left(\sum_j^p |\beta_j| \right) + (1 - \alpha) \left(\sum_j^p \beta_j^2 \right)$ avec $\alpha \in [0, 1]$ et $\lambda > 0$. Cette méthode vise à régulariser les MCO afin de limiter la propension du modèle au sur-apprentissage.
3. *Lasso* : il s'agit également d'une régression linéaire pénalisée mais le paramètre est fixé à 1, ce qui peut conduire l'algorithme à annuler certains paramètres superflus. Le modèle est ensuite ré-estimé est excluant ces paramètres nuls (régression post-lasso).
4. Régression quantile à la médiane. C'est un modèle linéaire dont les coefficients sont calculés de manière à minimiser la perte L_1 et non la perte quadratique. Cette approche doit permettre de limiter l'influence des observations extrêmes et minimise directement l'écart absolu moyen, qui est l'un des critères de choix de modèle.

5. *Gradient boosting machine* (GBM, Friedman, 2001) : le GBM est une technique qui permet d'estimer la fonction de régression de manière itérative, en entraînant à chaque étape une fonction de régression de faible qualité (ici, un arbre de décision avec une faible profondeur) de façon à minimiser l'erreur effectuée par la fonction de régression de l'étape précédente. Pour plus d'informations, voir le Chapitre 10 de Hastie et al. (2001).
6. Forêt aléatoire (*random forest*) : une forêt aléatoire est un algorithme qui prédit une variable de résultat en faisant la moyenne d'arbres aléatoires [6] construits de façon indépendante. On fait "pousser" chaque arbre à partir d'un échantillon bootstrappé des données d'entraînement, et en faisant pousser une branche (*i.e.* en séparant les données sur la valeur d'une caractéristique) de façon à maximiser la variance inter-classe, à partir d'un sous-échantillon des variables explicatives. Ajouter de l'aléa à chaque arbre / branche permet de rendre les erreurs non corrélées pour une meilleure performance. Pour plus d'informations, voir le Chapitre 15 de Hastie et al. (2001).
7. Réseau de neurones *feed forward* avec une seule couche cachée : un réseau de neurones est algorithme permettant d'obtenir une fonction de régression au moyen d'une composition de fonctions bien choisies (les "couches"). L'ouvrage de référence sur les réseaux de neurones et le *deep learning* est Goodfellow et al. (2016). Ici le réseau n'a qu'une seule couche cachée, de telle sorte que la prédiction est donnée par $\widehat{Y} = f(g(X'\beta_1)'\beta_2)$. La couche d'entrée est donnée par les caractéristiques $X \in \mathbb{R}^p$, qui passent dans une couche cachée pour donner $g(X'\beta_1) \in \mathbb{R}^m$, et finalement aboutir à la prédiction finale en couche de sortie $f(g(X'\beta_1)'\beta_2) \in \mathbb{R}$. Notons toutefois que d'une part la complexité d'un tel modèle le rend peut interprétable et que d'autre part `R` n'est pas le langage le plus approprié pour mettre en oeuvre des méthodes de *deep learning* de façon efficace et créative. L'implémention proposée par le package `nnet` de `R` est en outre très contrainte et très pauvre. Des recherches parallèles ont été effectuées sous `python` avec `torch` mais n'ont pas permis d'obtenir une performance dépassant le coût de mise en œuvre du modèle.

Notons que les résultats des deux derniers algorithmes ne sont pas reportés car ils n'ont pas été mis en œuvre sur l'ensemble de l'échantillon de test, à cause d'un temps de calcul rédhibitoire. Le tableau 11 ne met pas en évidence d'amélioration nette des prédictions par rapport aux moindres carrés ordinaires.

4.4 Combinaisons de modèles

Le processus de prédiction des kilomètres repose finalement sur une approche par combinaison de modèles, détaillée par la figure 6. Cette approche a été développée du fait des résultats décevants des approches standards. Elle s'appuie sur le fait, déjà mentionné, que la variable mesurant les kilomètres moyens parcourus annuellement entre les deux derniers contrôles techniques a un pouvoir explicatif tellement fort que son introduction dans les modèles brouille le signal contenu dans les autres variables. Ainsi la prédiction finale est la combinaison linéaire de deux sous-modèles :

1. Le premier modèle repose sur l'utilisation d'une multitude de variables explicatives mesurant les caractéristiques du véhicule indépendamment de son utilisation passée. Les paramètres associé à ces variables sont estimés séparément sur les trois sous-populations mentionnées précédemment à l'aide des algorithmes de machine learning standards, ainsi

6. def *arbre (algorithme de régression)* : un arbre est un algorithme qui partitionne le support $\mathcal{X}$ des variables explicatives X de manière à trouver une partition telle que les points de données appartenant à chacun des groupes soit les plus homogènes possibles en terme de variable de résultats. En dimension 2, un arbre découpe le support des données en rectangles et prédit comme variable de résultat pour une observation la moyenne de la variable de résultat des points tombant dans ce rectangle. Pour les lecteurs familiers de la régression linéaire, on peut conceptualiser un arbre comme une méthode permettant de construire de façon optimisée des croisements entre des indicatrices dépendants de variables explicatives.

Millésime	*RMSE*		*MAE*		*Biais normalisé*	
Algorithme	(brut)	(normalisé)	(brut)	(normalisé)	indiv.	total
Echantillon test futur (visite passée en 2019)						
MCO	21.57	0.837	12.73	0.705	0.005	0.017
Lasso	21.57	0.837	12.76	0.707	0.009	0.019
Elastic Net	21.57	0.837	12.73	0.705	0.006	0.015
Boosting	21.73	0.843	13.05	0.723	0.015	0.021
Régression médiane	21.73	0.843	12.35	0.684	-0.064	-0.059
MCO (log)	21.94	0.851	12.75	0.706	-0.004	0.009
Lasso (log)	21.92	0.851	12.76	0.707	-0.004	0.01
Elastic Net (log)	21.92	0.851	12.77	0.707	0	0.012
Boosting (log)	21.97	0.853	12.99	0.719	0.012	0.016
Régression médiane (log)	21.82	0.847	12.59	0.697	-0.001	0.007
Echantillon test en coupe (visite passée en 2018)						
MCO	21.31	0.833	12.55	0.696	-0.007	-0.003
Lasso	21.31	0.833	12.56	0.697	-0.006	-0.003
Elastic Net	21.31	0.833	12.54	0.696	-0.007	-0.002
Boosting	21.43	0.838	12.81	0.711	-0.004	-0.003
Régression médiane	21.45	0.839	12.22	0.678	-0.07	-0.067
MCO (log)	21.64	0.846	12.57	0.697	-0.012	-0.003
Lasso (log)	21.64	0.846	12.57	0.697	-0.012	-0.003
Elastic Net (log)	21.62	0.845	12.56	0.697	-0.011	-0.003
Boosting (log)	21.7	0.848	12.77	0.708	-0.007	-0.002
Régression médiane (log)	21.52	0.841	12.4	0.688	-0.008	-0.003

TABLE 11 – Prédiction de la distance quotidienne moyenne parcourue entre deux contrôle techniques – Algorithmes

Note : La mention (log) *indique que le modèle a été estimé en utilisant le logarithme de la distance. La moyenne de la valeur prédite est ensuite recalée sur celle la variable expliquée (calculée sur l'échantillon d'estimation) pour redresser le biais propre à ce type de modélisations.*

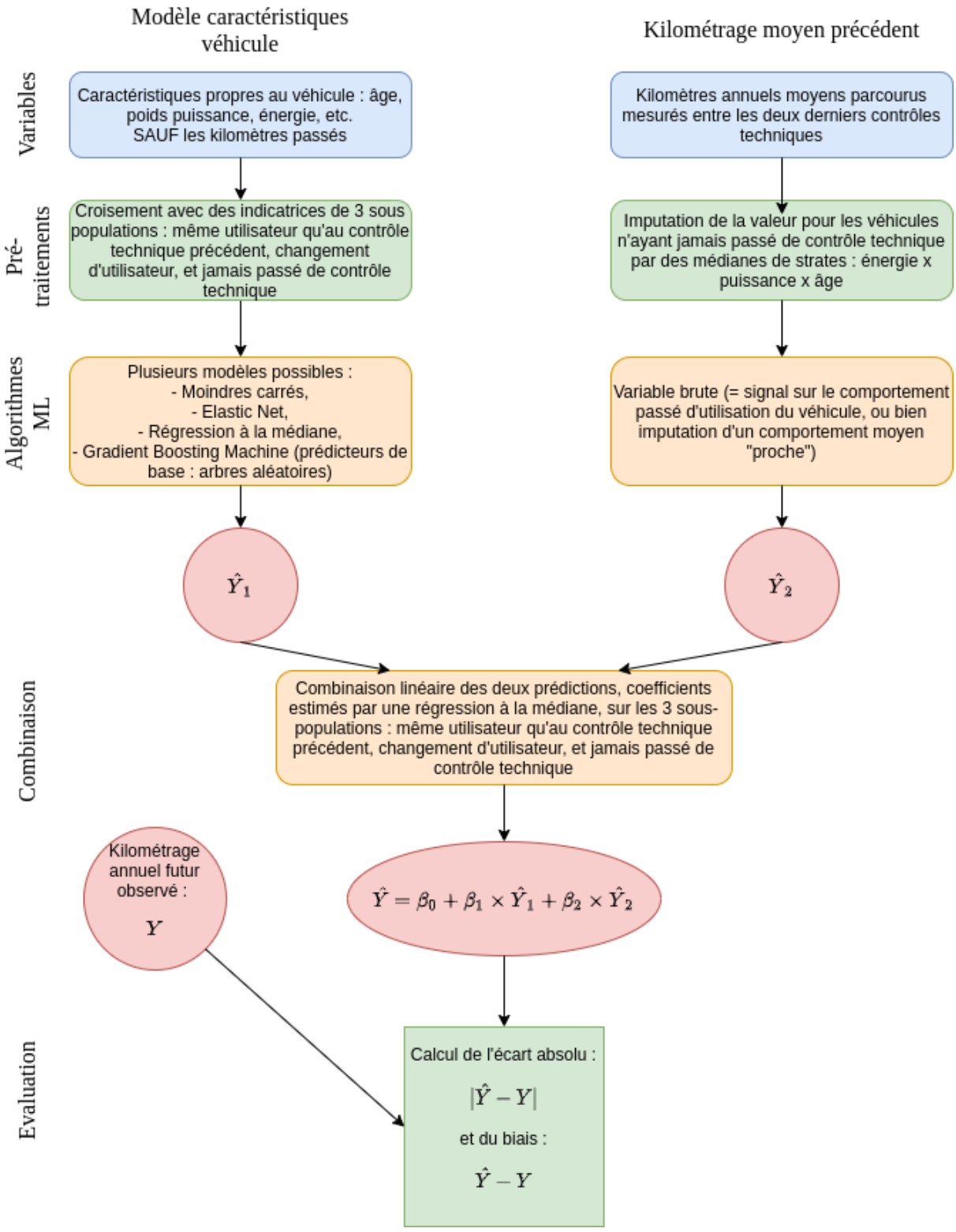

FIGURE 6 – Kilomètres annuels – processus de prédiction

	RMSE		*MAE*		*Biais normalisé*	
Algorithme	(brut)	(normalisé)	(brut)	(normalisé)	indiv.	total
Echantillon test futur (visite passée en 2019)						
MCO	21.57	0.837	12.69	0.703	0.006	0.017
Lasso	21.57	0.837	12.7	0.704	0.008	0.019
Elastic Net	21.57	0.837	12.69	0.703	0.006	0.017
Boosting	21.58	0.838	12.79	0.708	0.015	0.023
Régression médiane	21.6	0.838	12.64	0.7	0.001	0.008
MCO (log)	21.58	0.838	12.63	0.7	0.001	0.011
Lasso (log)	21.57	0.837	12.63	0.7	0.001	0.011
Elastic Net (log)	21.58	0.837	12.65	0.701	0.004	0.013
Boosting (log)	21.58	0.837	12.72	0.704	0.011	0.016
Régression médiane (log)	21.59	0.838	12.63	0.7	0.001	0.009
Echantillon test en coupe (visite passée en 2018)						
MCO	21.3	0.833	12.49	0.693	-0.007	-0.002
Lasso	21.3	0.833	12.5	0.693	-0.007	-0.002
Elastic Net	21.31	0.833	12.49	0.693	-0.007	-0.002
Boosting	21.3	0.833	12.53	0.695	-0.004	-0.002
Régression médiane	21.32	0.834	12.48	0.692	-0.008	-0.002
MCO (log)	21.3	0.833	12.46	0.691	-0.008	-0.003
Lasso (log)	21.3	0.833	12.46	0.691	-0.008	-0.002
Elastic Net (log)	21.3	0.833	12.46	0.691	-0.008	-0.002
Boosting (log)	21.33	0.834	12.5	0.694	-0.004	-0.002
Régression médiane (log)	21.31	0.833	12.46	0.691	-0.008	-0.003

TABLE 12 – Prédiction de la distance quotidienne moyenne parcourue entre deux contrôle techniques – Combinaison de modèles

que décrit à la section 4.3. Globalement, ce modèle cherche à capturer l'influence des caractéristiques propres au véhicule, indépendemment du comportement d'utilisation du conducteur. La prédiction qui en est issue peut s'interpréter comme le reflet du comportement moyen de la population qui utilise un véhicule possédant les mêmes caractéristiques.

2. Le second modèle n'utilise que les kilomètres moyens journaliers reportés entre les deux derniers contrôles techniques observés, étant donnée l'importance du signal contenu dans cette variable. Cette information n'est naturellement pas mesurée pour les véhicules n'ayant jamais passé de contrôle technique. Pour ces véhicules, nous mettons en place une méthode d'imputation par moyenne de strates de puissance et de carburant.

In fine, les prédictions issues de ces deux modèles sont combinées via un modèle linéaire dont les coefficients sont estimés par une régression linéaire pondérée, afin que les sommes observées et prédites des distances totales parcourues soient bien égales. La table 16 présente les coefficients utilisés pour combiner les modèles. On remarque, dans le premier tiers du tableau que pour les véhicules passant leur premier contrôle technique, la prédiction finale provient principalement de l'algorithme entraîné sur les caractéristiques du véhicule. Conformément à l'intuition, ce ratio augmente nettement pour les contrôles techniques suivants des véhicules ayant changé de main et devient dominant pour ceux ayant conservé le même utilisateur. Autre élément notable, le R^2 de cette régression finale est à peu près équivalent pour les deux premières catégories de véhicules, tandis qu'il est deux fois et demie plus élevé pour la troisième. Cela confirme la pertinence de l'utilisation passée d'un véhicule pour prédire son utilisation future mais surtout lorsqu'il s'agit du même conducteur.

Le tableau 12 détaille les résultats obtenus avec cette méthode combinée. Cette méthode permet une amélioration modeste mais réelle de la qualité des prédictions. La régression médiane combinée permet notamment de réduire le biais observé sur l'échantillon de test « futur » tout en améliorant un peu la précision de l'estimation. Le tableau 15 (page 33 en annexe) détaille les prédictions obtenues sur cet échantillon en distinguant les voitures passant leur première visite de celles plus anciennes. Ces résultats confirment, sur chacun des deux échantillons, la légère supériorité de cette méthode sur la plupart des critères.

Annexes

A Modélisation de la durée avec des survivants de long-terme

Pour prendre en compte le fait que les durées étudiées peuvent être infinies (*i.e.* le contrôle technique suivant n'est jamais passé, aucun événement de sortie de parc n'aura lieu, etc.), c'est-à-dire le fait que $P[T_i < \infty] < 1$, Zhao and Zhou (2006) utilisent une fonction de répartition impropre, de la forme $F_T(t) = pF_0(t)$ avec $F_0(.)$ une fonction de répartition telle que $\lim F_0(t) = 1$ lorsque $t \to +\infty$ et $p \in]0, 1]$. p peut être interprété comme la proportion des individus qui sont "à risque", c'est-à-dire qui ne sont pas des survivants de long-terme.

Le calcul de la probabilité $P[t < T < \infty|X]$ se fait de la manière suivante :

$$\begin{aligned} P[t < T < \infty|X] &= P[T < \infty|X] - P[T < t|X] \\ &= S_T(t|X) - S_T(\infty|X) \\ &= P[T > t|X] - P[T = \infty|X]. \end{aligned}$$

Par ailleurs, en supposant un taux de hasard proportionnel :

$$\begin{aligned} P[T = \infty|X] &= 1 - P[T < \infty|X] \\ &= \exp\left(-\int_0^\infty h(t, X)dt\right) \\ &= \exp\left(-\exp(X'\beta)\int_0^\infty \frac{pf_0(t)}{1 - pF_0(t)}dt\right) \\ &= \exp\left(-\exp(X'\beta)\left[-\log(1 - pF_0(t))\right]_0^\infty\right) \\ &= \exp\left(\exp(X'\beta)\log(1 - p)\right) \\ &= (1 - p)^{\exp(X'\beta)}. \end{aligned}$$

où $h(t, X) := \exp(X'\beta)h_0(t) := \exp(X'\beta)pf_0(t)/(1 - pF_0(t))$ est la fonction de hasard pour un individu de caractéristiques X. On a alors :

$$P[t < T < \infty|X] = S_0(t)^{\exp(X'\beta)} - (1 - p)^{\exp(X'\beta)}.$$

On sait que $S_0(.)$ peut être estimée au moyen de l'estimateur de Breslow. Par ailleurs, p peut être estimé de manière convergente par :

$$\widehat{p} = 1 - \widehat{S}_0(\infty),$$

voir Zhao and Zhou (2006).

Rentrons dans plus de détails pour mieux comprendre l'estimateur de la proportion de survivants de long-terme. Pour cela, on rappelle que l'on note $Y_i = \min(T_i, O_i)$ et $D_i = \{T_i < O_i\}$. Ainsi $D_i = 1$ correspond à une durée non censurée, tandis que $D_i = 0$ correspond à une durée censurée. En substituant l'estimateur de Breslow (1972), on peut écrire :

$$1 - \widehat{p} = \exp\left[-\sum_{t \in \mathcal{T}} \frac{\sum_{i=1}^n \{Y_i = t, D_i = 1\}}{\sum_{i=1}^n \{Y_i \geq t\} \exp(X_i'\widehat{\beta})}\right],$$

où $\mathcal{T}$ est l'ensemble des durées uniques trouvées dans l'échantillon. Ainsi la probabilité de survivants de long-terme estimée est une fonction du cumul, à chaque point du support, de la probabilité empirique de passer un contrôle technique, calculée à partir des véhicules n'ayant pas encore passé de contrôle technique.

B Estimations de l'utilisation du parc

	RMSE		*MAE*		*Biais normalisé*	
Sous-échantillons	(brut)	(normalisé)	(brut)	(normalisé)	indiv.	total
Echantillon test futur (visite passée en 2019)						
1/100e	21.57	0.837	12.73	0.705	0.005	0.017
1/200e	21.61	0.839	12.79	0.708	0.006	0.017
1/500e	21.75	0.844	12.91	0.715	0.008	0.016
1/1000e	21.85	0.848	13.04	0.723	0.004	0.013
Echantillon test en coupe (visite passée en 2018)						
1/100e	21.31	0.833	12.55	0.696	-0.007	-0.003
1/200e	21.35	0.835	12.6	0.699	-0.006	-0.002
1/500e	21.48	0.84	12.72	0.706	-0.006	-0.003
1/1000e	21.56	0.843	12.84	0.712	-0.01	-0.007

TABLE 13 – Prédiction de la distance quotidienne moyenne parcourue entre deux contrôle techniques – Taille de l'échantillon

Note : L'échantillon utilisé pour les autres estimations contient un 1/100e des données, il est réduit par un facteur 2, 5 et 10 pour déterminer la sensibilité des estimations à la taille de l'échantillon.

Millésime	*RMSE*		*MAE*		*Biais normalisé*	
de l'échantillon	(brut)	(normalisé)	(brut)	(normalisé)	indiv.	total
Echantillon test futur (visite passée en 2019)						
2014	21.48	0.84	12.8	0.71	-0.006	0.007
2015	21.33	0.834	12.65	0.702	0.001	0.006
2016	21.36	0.835	12.74	0.707	0.011	0.023
2017	21.3	0.833	12.65	0.702	0.005	0.012
2018	21.31	0.833	12.55	0.696	-0.007	-0.003
Echantillon test en coupe (visite passée en 2018)						
2014	21.75	0.844	12.98	0.719	0	0.017
2015	21.6	0.838	12.83	0.711	0.009	0.016
2016	22.06	0.856	13.43	0.744	0.048	0.095
2017	21.71	0.842	13.04	0.722	0.029	0.058
2018	21.57	0.837	12.73	0.705	0.005	0.017

TABLE 14 – Prédiction de la distance quotidienne moyenne parcourue entre deux contrôle techniques – Millésime de l'échantillon

	RMSE		*MAE*		*Biais normalisé*	
Algorithme	(brut)	(normalisé)	(brut)	(normalisé)	indiv.	total
Echantillon test futur – Première visite						
MCO	27.81	0.907	16.57	0.859	0.017	0.038
Lasso	27.81	0.907	16.62	0.862	0.019	0.041
Elastic Net	27.88	0.909	16.54	0.858	0.014	0.035
Boosting	27.83	0.908	16.88	0.875	0.022	0.041
Régression médiane	28.04	0.914	16.24	0.842	-0.015	0.007
MCO (log)	27.97	0.912	16.31	0.846	-0.002	0.02
Lasso (log)	27.83	0.908	16.33	0.847	-0.001	0.021
Elastic Net (log)	27.91	0.91	16.32	0.846	0	0.021
Boosting (log)	27.78	0.906	16.34	0.847	-0.004	0.017
Régression médiane (log)	28.02	0.913	16.25	0.843	-0.011	0.011
Echantillon test futur – Visites suivantes						
MCO	20.71	0.837	12.23	0.683	0.004	0.008
Lasso	20.71	0.837	12.24	0.684	0.006	0.01
Elastic Net	20.71	0.837	12.23	0.683	0.005	0.009
Boosting	20.73	0.838	12.31	0.687	0.014	0.015
Régression médiane	20.71	0.837	12.22	0.682	0.003	0.008
MCO (log)	20.71	0.837	12.2	0.681	0.001	0.007
Lasso (log)	20.71	0.837	12.2	0.681	0.001	0.007
Elastic Net (log)	20.71	0.837	12.22	0.682	0.005	0.01
Boosting (log)	20.73	0.838	12.29	0.686	0.014	0.015
Régression médiane (log)	20.71	0.837	12.21	0.682	0.003	0.008

TABLE 15 – Prédiction de la distance quotidienne moyenne parcourue entre deux contrôle techniques dans le futur – Combinaison de modèles

Algorithme	Constante β_0	Dist. passée β_1	Prédicteur ML β_2	R_2
	Première visite			
MCO	-0.00	-0.00	1.00	0.20
Lasso	-0.00	0.00	1.00	0.20
Elastic Net	-1.24	-0.09	1.12	0.19
Boosting	-3.52	-0.36	1.44	0.20
Régression médiane	0.99	-0.13	1.20	0.18
MCO (log)	3.68	-0.08	0.99	0.20
Lasso (log)	3.68	-0.07	0.99	0.20
Elastic Net (log)	3.64	-0.10	1.02	0.19
Boosting (log)	3.39	-0.30	1.22	0.19
Régression médiane (log)	1.47	-0.11	1.07	0.19
	Visites suivantes avec changement d'utilisateur			
MCO	-13.86	0.67	0.71	0.21
Lasso	-13.85	0.67	0.71	0.20
Elastic Net	-13.91	0.67	0.71	0.20
Boosting	-14.61	0.69	0.71	0.20
Régression médiane	-10.73	0.68	0.67	0.20
MCO (log)	-8.75	0.68	0.56	0.20
Lasso (log)	-8.75	0.68	0.56	0.20
Elastic Net (log)	-8.87	0.68	0.56	0.20
Boosting (log)	-9.92	0.69	0.58	0.20
Régression médiane (log)	-9.98	0.68	0.59	0.20
	Visites suivantes sans changement d'utilisateur			
MCO	-6.27	0.88	0.35	0.47
Lasso	-6.27	0.88	0.35	0.47
Elastic Net	-6.28	0.88	0.35	0.47
Boosting	-6.43	0.88	0.35	0.47
Régression médiane	-5.14	0.88	0.34	0.47
MCO (log)	-4.25	0.88	0.27	0.47
Lasso (log)	-4.25	0.88	0.27	0.47
Elastic Net (log)	-4.29	0.88	0.28	0.47
Boosting (log)	-4.57	0.88	0.28	0.47
Régression médiane (log)	-4.69	0.88	0.29	0.47

TABLE 16 – Combinaison de modèles – Coefficients estimés par la régression finale

Références

Abbring, J. H. (2002). Stayers versus defecting movers : a note on the identification of defective duration models. *Economics Letters*, 74(3) :327 – 331.

Breslow, N. (1972). Disussion of regression models and life-tables by cox, d. r. *J. Roy. Statist. Assoc., B*, 34 :216–217.

Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *Ann. Statist.*, 29(5) :1189–1232.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

L'Hour, J. (2020). L'économétrie en grande dimension. Technical Report M2020-01, Documents de Travail de l'Insee - INSEE Working Papers.

Maller, R. and Zhou, X. (1996). *Survival analysis with long-term survivors*. New-York : Wiley.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4) :385–395.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*. Number 0262232197 in MIT Press Books. The MIT Press.

Zhao, X. and Zhou, X. (2006). Proportional hazards models for survival data with long-term survivors. *Statistics & Probability Letters*, 76(15) :1685 – 1693.