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)

Résumé

Le 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 on ne peut connaître avec certitude le statut des véhicules dont la dernière visite précède la date à laquelle on souhaite déterminer le parc.

Cette communication 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.

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 tout du moins). Il est en revanche nécessaire d'estimer cette distance quand le dernier relevé est antérieur à la fin de l'année considérée. 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 en retard.

Ce premier modèle vise à prédire la distance annuelle moyenne parcourue par un véhicule entre deux contrôles techniques. L'approche proposée,développée après une phase initiale où des approches standard (mono-modèles) se sont montrées décevantes, 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, estimé à l'aide d'algorithmes classiques de machine learning (MCO, elastic-net, régression quantile, gradient boosting machine, forêt aléatoire) ; 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.

Appartenance au parc des véhicules routiers. La précédente version du RSVero s'appuyait uniquement sur les informations du système d'immatriculation des véhicules. Comme de nombreuses sources administratives, celui-ci enregistre seulement partiellement la cessation d'activité des unités suivies, ici le retrait des véhicules de la circulation. L'utilisation des données de contrôles techniques permet de résoudre ce problème, les véhicules retirés de la circulation n'étant plus présentés aux visites de contrôle régulières. Il existe, par nature, une incertitude sur le fait qu'un véhicule passera ou non un contrôle technique après la dernière visite observée. Toujours dans le souci produire des statistiques sur le parc des véhicules sans attendre que l'ensemble des véhicules ait passé une visite, dans un délai intégrant un retard raisonnable, il est nécessaire d'estimer la probabilité de cet événement.

On s'appuie pour cela sur une modélisant de la durée écoulée entre deux visites consécutives. Celui-ci présente trois particularités, il 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. Il est également apte à capturer l'existence de « survivants de long terme », c'est-à-dire les véhicules qui ne passeront plus jamais de contrôle technique. Il intègre enfin une méthode de sélection des variables explicatives, potentiellement nombreuses, dite elastic-net.

Bibliographie

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

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

[3] 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.

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

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

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

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

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