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

Journées de méthodologie statistique

Jérémy L'Hour - Corentin Trevien

Insee, SSP Lab - Sdes

31 mars 2022

Plan

Le répertoire statistique des véhicules routiers

- Ensemble des véhicules (voitures, utilitaires, poids lourds, bus, etc.) immatriculés en France
  - Refonte en 2020-2021
  - Nombreux usages statistiques : indicateurs conjoncturels, base de sondage, études, etc.
- Première source : système d'immatriculation des véhicules (ANTS)
  - Informations techniques : poids, CO_2, puissance, carburant, etc.
  - Opérations administratives : mise en circulation, changement d'adresse, vente, destruction, etc.
- Deuxième source : résultats des contrôles techniques (Utac)
  - Nouveauté principale de la nouvelle version du répertoire
  - Pour chaque contrôle : la date, le type de contrôle, relevé kilométrique et le résultat
- Troisième source : répertoire Sirène (Insee)
  - Identifier la personne morale propriétaire du véhicule, le cas échéant

Etablir le parc des véhicules et son utilisation

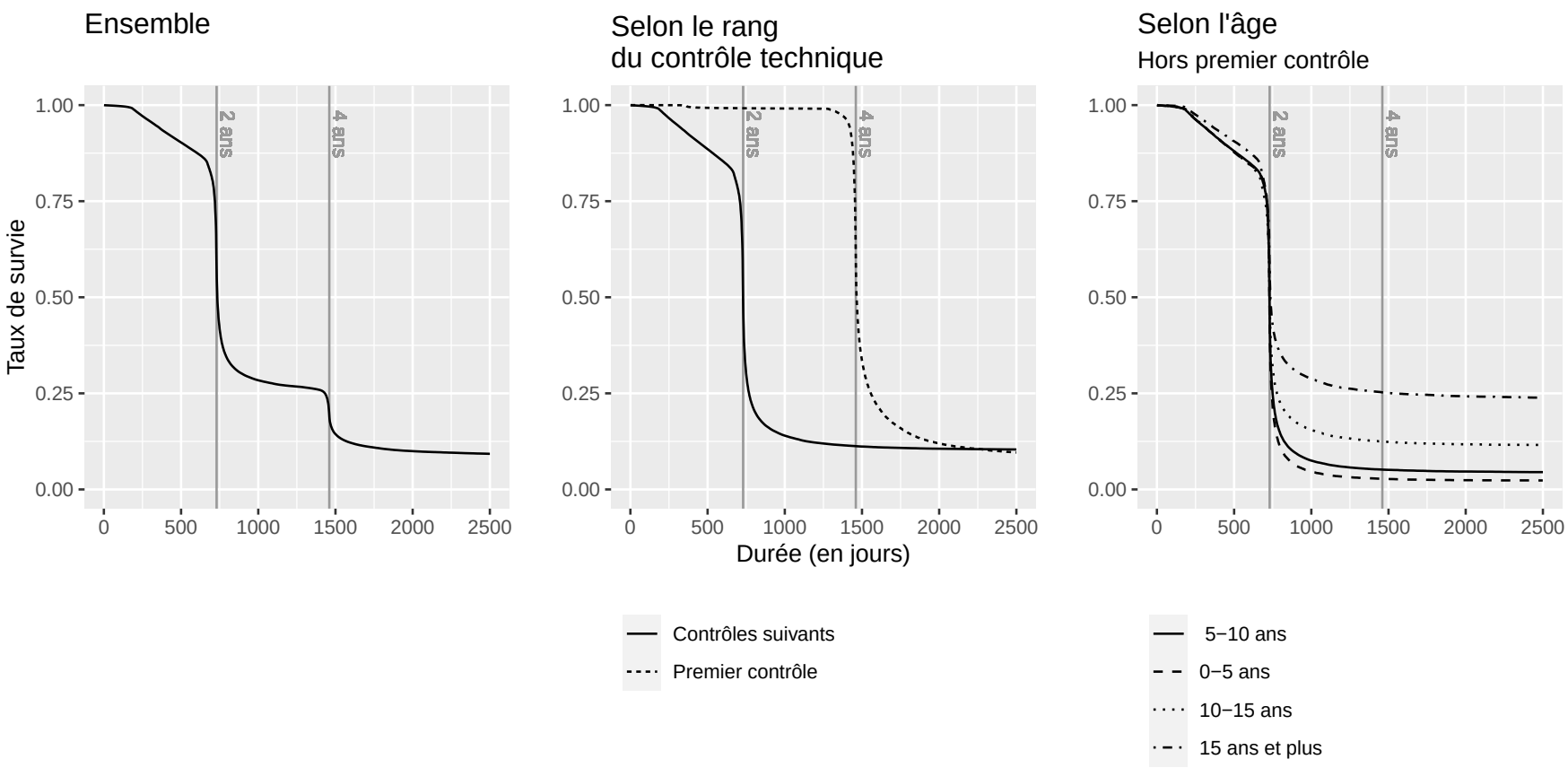
- Objectif de l'expérimentation SSP Lab-Sdes : affiner l'utilisation des données de contrôle technique
  - Basée sur les méthodes développées par le CPS de la refonte du répertoire
  - Centrée sur les voitures
- Appartenance au parc : véhicules en circulation au 1er janvier de chaque année
- Utilisation du parc : distance annuelle parcourue par chaque véhicule

Plan

Déterminer l'appartenance au parc avec les contrôles techniques

- Problème commun à de nombreuses sources administratives : les sorties de champ sont imparfaitement connues
  - Suivi partiel par les opérations administratives de sortie (mise à la casse, sortie de territoire, etc.)
- L'absence de passage régulier de contrôles techniques indique un arrêt de la circulation
  - Non-événement, par nature non-observé
  - En conséquence, pas de date précise de sortie de parc
- Données brutes nécessitant la modélisation de la durée écoulée entre deux visites
  - Pas de modélisation des événements de sortie de parc, la détermination des véhicules en circulation étant toujours rétrospective

Durée entre deux visites (Kaplan-Meier)



Censure

- Traitement séparé de la première visite et des suivantes
  - Contrôle obligatoire 4 ans après la mise en circulation puis tous les deux ans
- Observations censurées : durée écoulée depuis la dernière visite
  - Ou depuis la mise en circulation pour les véhicules n'en ayant jamais passé
  - Observations nécessaires pour déterminer la part des véhicules ne passant plus jamais de visites (*survivants de long terme*)
- Observations non-censurées : durée entre deux visites
  - Seuls les durées postérieures à un contrôle technique réussi sont conservées
  - Observations nécessaires pour l'identification de la fonction de survie

Modélisation de la durée entre deux visites

- Probabilité de passer un contrôle technique dans le futur :

$$P\left[t - C_{f(t)} < C_{f(t)+1} - C_{f(t)} < \infty | X\right]$$

  Avec :
  - $C_{f(t)}$ l'âge du véhicule au dernier contrôle technique
  - X des caractéristiques du véhicule
  - t est l'âge du véhicule à la date considérée

- Modèle de Cox à hasard proportionnel :

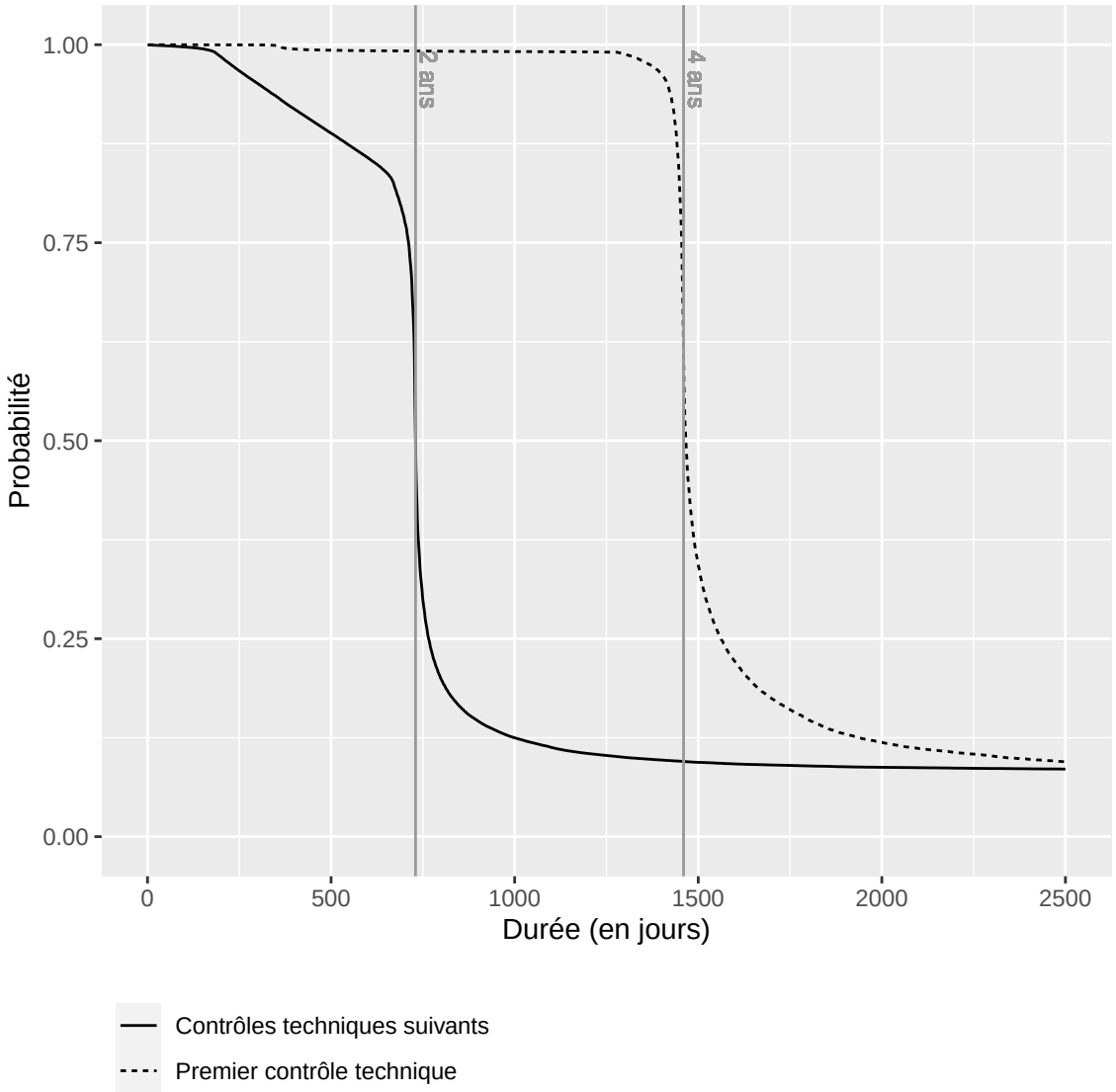
$$h\left(t|X\right) = \exp\left(X'\beta_0\right) h_0(t).$$

  Avec :
  - Taux de défaillance de base : $h_0(t)$
  - Modulée par un facteur multiplicatif : $\exp\left(X'\beta_0\right)$
  - Estimateur de Breslow pour la fonction de survie : absence d'hypothèse paramétrique

Modélisation de la durée entre deux visites

- Survivant de long terme :
  - Individu pour lequel la durée jusqu'à l'événement considéré (i.e. la visite suivante) est infinie
  - Observations nécessaires pour l'identification de la fonction de survie
  - Cas de figure pris en compte par l'estimateur de Breslow
    - Dans les modèles paramétriques plus simple, l'événement doit toujours finir par se produire
- Pénalisation : Modèle de Cox-Lasso (Tibshirani, 2007)
  - Beaucoup de variables disponibles sur le véhicule
  - Pénalisation sur les coefficients pour éviter le sur-apprentissage

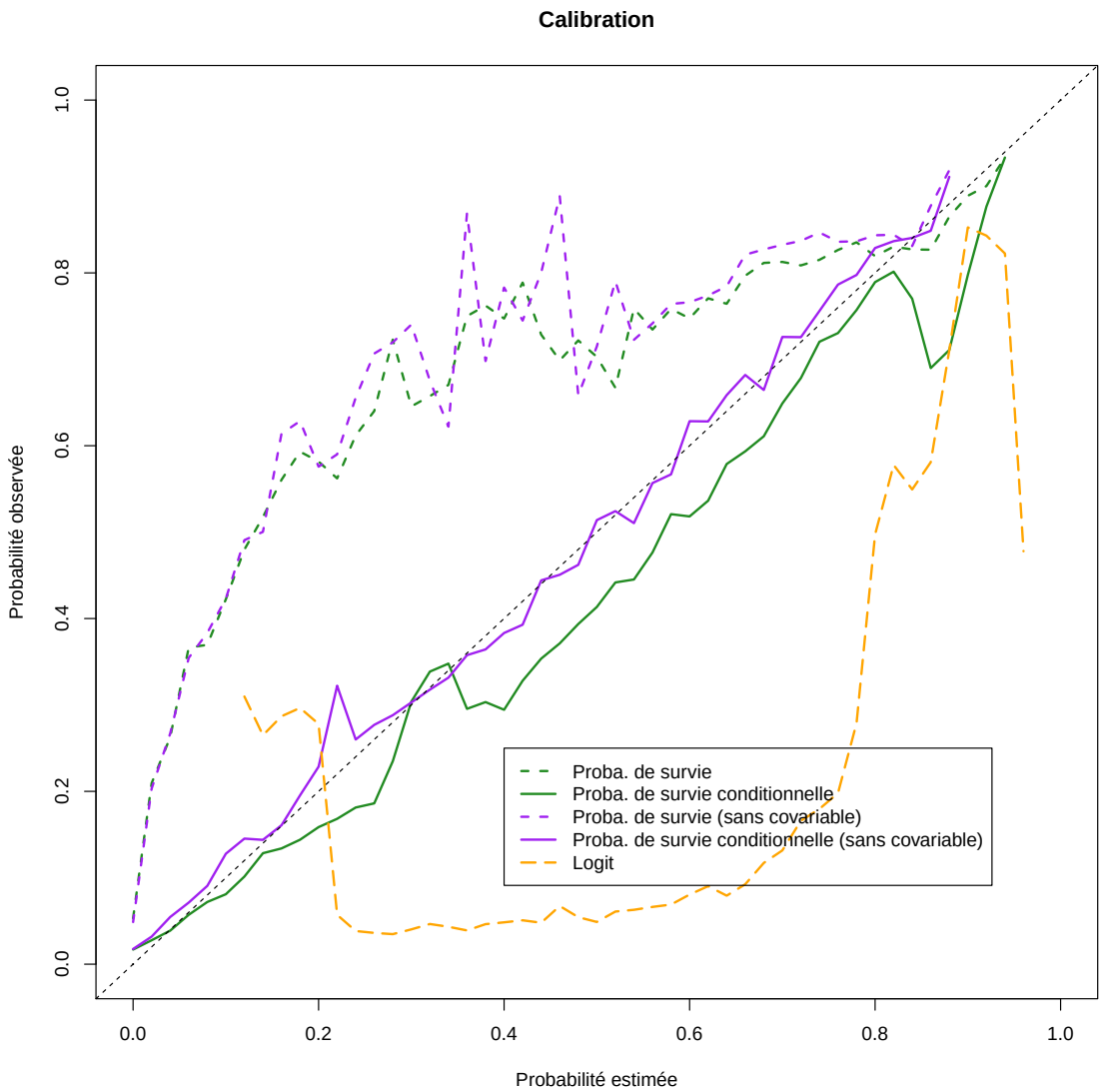
Estimateur de la fonction de survie de base



Utilisation du modèle pour le calcul du parc

- Déterminer l'appartenance au parc des véhicules en circulation au 1er janvier d'une année donnée
- Pas d'incertitude pour :
  - Les véhicules ayant passé un contrôle technique après cette date
  - ou ayant fait l'objet d'une opération administrative de sortie de parc
- Utilisation du modèle pour prédire la probabilité du passage d'un contrôle technique dans le futur
  - Pour les véhicules dont la dernière visite est antérieure à cette date
  - Probabilité conditionnelle ou non ?
  - Utilisation des caractéristiques du véhicule ?
- Calibration du modèle :
  - Pour un niveau de probabilité de passage p, quelle est la part des véhicules passant effectivement un visite ?
  - Estimation du modèle sur des données de 2018 et antérieures
  - Comparaison prédiction au 1er janvier 2019/réalisation ultérieure

Calibration du modèle



Hypothèses complémentaires

- Quel retard maximal retenir en cas d'absence de contrôle technique ?
  - Seuil indispensable au retrait des véhicules ne circulant plus
    - Probabilité de passer un contrôle technique dans le futur jamais nulle
  - Mise à l'écart d'un certain nombre de véhicules mais prise en compte d'éventuels arrêts temporaires d'utilisation
  - Test de différents seuils : 6, 12 et 18 mois
    - Impact modéré : écart respectif de 1 et 1,6 % en 2015

- A quelle date retirer un véhicule du parc quand son retrait est acté ?
  - Deux hypothèses polaires : immédiatement après la dernière visite ou à la date théorique de la visite suivante
  - Impact majeur sur le nombre de véhicule en circulation : écart de 15,2 % en 2015

Plan

Déterminer l'utilisation annuelle d'un véhicule

- Déterminer la distance annuelle parcourue par chaque véhicule
  - En utilisant le relevé kilométrique réalisé à chaque contrôle technique
- Calcul de la distance par interpolation si :
  - Relevés kilométriques disponibles avant et après l'année d'intérêt
  - Hypothèse d'une utilisation constante du véhicule (hors crise sanitaire)
- Imputation de la distance sur observables si :
  - Pas de visite postérieure à l'année d'intérêt
  - Valeur manquante ou aberrante (négative ou $>$ à 200 000 km/an)
- Pourquoi imputer ?
  - Diffuser des données sur l'utilisation du parc, sans attendre que la totalité des véhicules ait passé une visite
    - 4 ans + 1 an de retard

Objectifs et validation de la modélisation

- Variable d'intérêt : distance quotidienne moyenne entre deux visites

$$kmj_{f(t)+1} = \frac{K_{f(t)+1} - K_{f(t)}}{C_{f(t)+1} - C_{f(t)}}$$

- Priorité 1 : distance totale non biaisée
  - Pondération du modèle par $C_{(t)} - C_{(t+1)}$
- Priorité 2 : réduction de l'erreur individuelle
  - RMSE et MAE
- Découpage de l'échantillon : apprentissage/test
  - Estimation : 1 % des visites ayant eu lieu en 2018
    - 183 201 observations
  - Test : 1 % des visites ayant eu lieu en 2018 (coupe) et 2019 (futur)

Etape 1 : choix des variables et des échantillons

- Utilisation d'un modèle linéaire simple
- Variables disponibles
  - Sur le véhicule : âge, carburant, poids, norme euro, puissance, etc.
  - Sur son utilisateur : localisation, professionnel ou particulier, changement de main
  - Sur son utilisation passée : distances parcourues avant la dernière visite connue
- Distances parcourues dans le passé : un pouvoir explicatif fort
  - Quand la variable est disponible : postérieurement à la 1re visite
  - Quand l'utilisateur n'a pas changé
  - « Ecrase » la contribution des autres variables
- Séparation de l'échantillon en 3 parties
  - 1. Avant le 1re contrôle technique
  - 2. et 3. Visites suivantes, avec/sans changement d'utilisateur

Etape 2 : méthodes de machine learning

- Utilisation d'algorithmes de machine learning
  - Régressions pénalisées (elastic-net et LASSO)
  - Régression médiane
  - Gradient boosting machine
  - Réseau de neurone et random forest écartés du fait de temps de traitement rédhibitoires
  - Avec ou sans transformation en log
- Résultats assez décevants : peu d'amélioration par rapport au modèle linéaire

Etape 3 : modèle combiné

- Idée générale : prendre en compte de manière séparée l'information apportée par les caractéristiques du véhicule et son utilisation passée
- Modèle 1 : caractéristiques du véhicule
  - Utilisation des algorithmes de machine learning sans la distance parcourue dans le passé
- Modèle 2 : utilisation passée du véhicule
  - Moyenne de strates
- Combinaison linéaire des deux valeurs prédites par MCO
- Modèle retenu : régression médiane combinée
  - Réduction du biais sur la distance totale prédite en 2019
  - Meilleure précision

Conclusion

- Pour déterminer les véhicules en circulation :
  - Utilité d'une modélisation économétrique précise des durées entre contrôles techniques
  - Mais les règles « métier » nécessaires à l'application du modèle ont un fort impact sur les résultats
- Pour déterminer les distances annuelles parcourues :
  - Des méthodes de machine learning utiles quand elles intègrent bien la spécificité des données
- Mise en production de l'expérimentation pour le calcul du parc au 1er janvier 2022 et son utilisation en 2021
  - Échéance : T2 2022