Détection d'outliers. Modélisation et prédiction. Application aux données de véhicules d'occasion.

Solohaja Faniaha Dimby

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Thèse de doctorat

pour obtenir le grade de :

Docteur ès-sciences

Spécialité : Mathématiques Appliquées

présentée par

Solohaja-Faniaha DIMBY

Detection d'outliers. Modélisation et prédiction. Application aux données de véhicules d'occasion

---

Soutenue publiquement le 21 décembre 2015 devant le jury composé de :

M. Jean-MarcBardet	Professeur, Université Paris 1	Directeur
M. Patrice Bertail	Professeur, Université Paris X	Rapporteur
M. Paul Doukhan	Professeur, Université Cergy Pontoise	Examinateur
M. Fabrice Gamboa	Professeur, Université Paul Sabatier	Rapporteur
M^{me} Cécile Hardouin	Maitre de Conférence, Université Paris X	Examinateur
M^{me} Dominique Haughton	Professeur, Bentley University	Examinateur
M. Joseph Rynkiewicz	Maitre de Conférence, Université Paris 1	Co-Directeur
M. Daniel Urbah	Directeur Informatique à Autobiz	Invité

Laboratoire SAMM - Université Paris 1
90, rue de Tolbiac
75 013 Paris

A Papa, Maman, Neny, Tonton Mi,
Harena.
Ceci est un modeste témoignage de mon amour
infini et inconditionnel...

Thèse : La loi ultime du monde est le hasard et tout déterminisme partiel qu'on peut y trouver est un effet de la loi des grands nombres.
Anthithèse : La loi ultime du monde est entièrement déterministe et tout phénomène aléatoire qu'on peut y observer est un effet du chaos déterministe.
Dans ***"Chaos et déterminisme"****, J.-L. Chabert, K. Chemla, A. Dahan Dalmedico, Paris, Edition Seuil, 1992.*

"Une cause très petite, qui nous échappe, détermine un effet considérable que nous ne pouvons pas ne pas voir, et alors nous disons que cet effet est dû au hasard. Si nous connaissions exactement les lois de la nature et la situation de l'univers à l'instant initial, nous pourrions prédire exactement la situation de ce même univers à un instant ultérieur. Mais, lors même que les lois naturelles n'auraient plus de secret pour nous, nous ne pourrions connaître la situation qu'approximativement. Si cela nous permet de prévoir la situation ultérieure avec la même approximation, c'est tout ce qu'il nous faut, nous disons que le phénomène a été prévu, qu'il est régi par des lois ; mais il n'en est pas toujours ainsi, il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux ; une petite erreur sur les premières produirait une erreur énorme sur les derniers. La prédiction devient impossible et nous avons le phénomène fortuit."
Extrait de ***"Calcul des probabilités"****, Henri Poincaré(1854-1912).*

Remerciements

Jean-Marc Bardet m'a accordé sa confiance en acceptant d'encadrer ce travail. Je lui voue un profond respect et une grande admiration. Je ne saurai comment lui exprimer ma gratitude autrement qu'en me faisant une promesse d'agir comme lui avec des étudiants qui seraient dans la même situation que moi, si un jour l'occasion me serait donnée.

Joseph Rynkiewicz a été mon co-directeur. Je lui suis reconnaissante pour ses enseignements et ses instructions.

Mes sincères remerciements envers la société Autobiz en la personne de Daniel Urbah, qui a été l'initiateur de ce projet de thèse. Je n'oublie pas Jean Gourdault-Montagne, grâce à qui, beaucoup de choses ont été plus faciles ainsi que toute l'équipe informatique avec qui j'ai travaillé étroitement.

Patrice Bertail et Fabrice Gamboa ont accepté d'être les rapporteurs de cette thèse et m'ont fait des commentaires très intéressants. Je souhaiterais leur exprimer toute ma gratitude.

Ce manuscrit a été grandement amélioré grâce aux remarques de relecture de Cécile Hardouin. Je lui suis très reconnaissante.

Paul Doukhan et Dominique Haughton m'honorent de leur présence en tant que membres du jury et en plus, ils m'ont fait des critiques très constructives. Je les remercie chaleureusement.

Durant ces années, j'ai pu travailler dans un cadre particulièrement agréable, grâce à l'ensemble de l'équipe du laboratoire SAMM. Je suis extrêmement redevable envers Marie Cottrell, qui m'a toujours été d'un soutien inestimable et de très bons conseils. J'adresse à chacun des membres du laboratoire un chaleureux remerciement. Je me souviendrai toujours des doctorants et des jeunes docteurs pour toutes ces discussions autour d'un café ou à la cantine, où, chacun, de par son opinion politique, ses convictions religieuses et philosophiques, pense qu'un autre monde est possible...

Mes dernières pensées iront vers ma famille, mes parents, $\mathscr{N}eny$, ma douce grand-mère, qui, selon les moments, conseille ou protège, mon frère, dont l'affection n'a jamais fait défaut. La satisfaction que je tire de ces années de thèse vient également beaucoup de la présence de $\mathscr{H}arena$, qui fait que chaque jour me soit unique. $\mathscr{V}aliko$, merci pour ton soutien et tes encouragements et je te souhaite beaucoup de courage pour ta soutenance prochaine.

Je souhaiterais apporter une dernière précision. Dans une société où tout est à l'image de la rapidité, de la concurrence, de la rentabilité, je revendique avoir fait au travers de ce travail de thèse et dans l'esprit du philosophe Pierre Sansot, "*l'éloge de la lenteur*" et "*de son bon usage*". En effet, il m'a fallu du temps pour dessiner les contours de mon sujet, comprendre ce que j'observais, traiter convenablement les données, avoir un certain recul sur l'applicabilité de la théorie,... faire de la recherche.

Table des matières

Introduction générale

« Vous me demandez de vous prédire les phénomènes qui vont se produire. Si, par malheur, je connaissais les lois de ces phénomènes, je ne pourrais y arriver que par des calculs inextricables et je devrais renoncer à vous répondre ; mais, comme j'ai la chance de les ignorer, je vais vous répondre tout de suite. Et, ce qu'il y a de plus extraordinaire, c'est que ma réponse sera juste. »

Dans *"Calcul des probabilités"*, Henri Poincaré, 1912.

1. Cadre de la thèse

1.1. Le contexte

Le souhait des entreprises d'avoir à leur disposition des outils permettant d'exploiter et de valoriser leurs données dans un cadre mathématique a suscité des liens entre le milieu académique et le monde industriel. Ces relations ont également étendu le champ de recherche des statisticiens qui se traduit par la construction de modèles simples à expliquer mais assez riches pour prendre en compte la complexité d'un phénomène étudié. C'est dans cette recherche de compromis entre adéquation aux données et relative simplicité du modèle que s'inscrit notre travail.

Cette thèse qui fait suite à un stage de Master 2 a été réalisée dans le cadre d'une CIFRE. Le partenaire industriel en est la société Autobiz Shakazoola qui est un acteur dans la valorisation des véhicules d'occasion. Le partenaire académique est le laboratoire SAMM de l'Université Paris 1 Panthéon-Sorbonne. Le travail se propose de contribuer à l'appréciation du marché des véhicules d'occasion, et plus précisément, en une élaboration de modèles de prédiction des prix et des délais de vente des véhicules d'occasion. Chaque partie de la thèse correspond au traitement d'un objectif autour de cette thématique. Selon le problème qui nous est posé, nous chercherons à appréhender les questions en faisant appel à des méthodes statistiques dans le but de proposer des solutions réalistes, exploitables et adaptées aux contraintes industrielles. Ces contraintes dont il est question sont des règles imposées par les experts du métier.

Le travail fait en grande partie appel à une démarche expérimentale. L'intérêt de cette démarche expérimentale est qu'elle offre un champ d'investigation très large : l'observation des phénomènes, l'identification des problèmes, la proposition des solutions et la compréhension des résultats. L'idée suivie dans la rédaction de ce rapport est de rendre ce travail accessible, tant au mathématicien qui, retrouvant les méthodes statistiques utilisées, découvrira les perspectives d'applications qu'ouvrent ces outils, qu'aux utilisateurs et praticiens en industrie, qui pourront

acquérir quelques notions mathématiques indispensables à la bonne compréhension de l'analyse des données réelles et de la modélisation statistique.

1.2. Les motivations du projet de recherche

La motivation découle des enjeux que nous décrivons ci-dessous.

La cote VO [1] et la VR [2] : La **cote VO** est une estimation correspondant au prix moyen auquel un professionnel propose un VO aux particuliers. La **VR** définie à un instant t_0 correspond à la valeur à une date future t_1 attribuée à un véhicule en fonction de son âge et du kilométrage qu'il aurait effectué. Si la cote VO représente un outil d'aide à la décision pour une vente ou un achat dans l'immédiat, la VR présente ainsi une motivation de plus lors d'une transaction puisqu' elle permet l'achat dans une intention de revente. Pour les professionnels de l'automobile, la gestion des cotes VO et VR est cruciale pour assurer une croissance profitable et durable.

Les marchés VN [3] et VO : En France, le secteur automobile fournit 300 000 emplois directs et 760 000 emplois indirects qui, en ajoutant les emplois induits, nous donne au total un chiffre de 2,5 millions d'emplois. Il constitue environ 9,5 % de l'ensemble de l'industrie et s'avère être la branche qui contribue le plus à la balance commerciale. *(Source : DGE [4], 2006).*

Les constructeurs français ont produit en 2007, près de 2,5 millions véhicules en France. Les ventes de VN s'effectuant sous plusieurs formes (ventes aux particuliers qui gardent leur voiture en moyenne cinq ans, ventes aux sociétés, ventes aux entreprises LCD [5], ventes aux LLD [6] qui mettent généralement en vente leurs véhicules au bout de 36 mois) induisent une offre abondante de VO sur le marché secondaire, plus de 5 millions de transactions de voitures particulières d'occasion ont été enregistrées en 2007.

La politique des constructeurs : Pour augmenter les ventes et les parts de marché, les constructeurs généralistes(*exemple* : Peugeot, Citroën) mettent en place des politiques basées sur la fluidité du marché VO. Les actions mises en place consistent en des renouvellements fréquents de la gamme, l'introduction de nouvelles phases des modèles existants, la stratégie des séries limitées et les incitations financières (remises, offres de financement promotionnelles, *etc.*). Les constructeurs spécialistes (*exemple* : Mercedes, BMW et Audi) misent sur une VR plus forte de leurs produits et pour une meilleure maîtrise de ces VR, leur rôle ne se limite pas à la vente des VN mais s'étend à travers une implication importante dans le marché VO.

La société Autobiz : La cote Argus domine le marché depuis fort longtemps. Ainsi, face à cette concurrence, Autobiz s'est investit dans un contrat de recherche pour produire une cote fiable et réaliste afin de conquérir le marché. Créée en 2002, la société Autobiz est aujourd'hui le leader de l'information de marché sur le secteur automobile. Cette information est mise à disposition sous forme d'études ciblées sur la distribution de VO, de panels, d'articles de presse, de bases de données recensant plus de 100 000 entreprises du secteur. La société apporte son

1. Véhicules d'occasion
2. Valeur Résiduelle
3. Véhicules Neufs
4. Direction Générale des Entreprises
5. Location de Courte Durée
6. Location de Longue Durée

expertise aux constructeurs, aux groupes de distribution, aux équipementiers, aux sociétés de financement, aux sites de petites annonces, aux spécialistes des enchères, et à d'autres professionnels travaillant dans le secteur. Parmi les outils d'aide à la décision proposés, on distingue **SystèmeVO** qui est un outil de benchmarking du marché et de la concurrence. SystèmeVO rassemble les informations nécessaires à une connaissance parfaite du marché.

1.3. La culture scientifique

Bien que dans cette thèse l'aspect pratique a été privilégié, les problématiques industrielles ont sollicité plusieurs thématiques de recherche. La détection d'outliers a nécessité une étude élargie et approfondie sur les méthodes de statistiques robustes, la minimisation des distances l_1 ainsi que les résultats théoriques sur les statistiques d'ordre. Le problème de modélisation et de prédiction a nécessité un investissement dans les différentes méthodes de régressions. Les méthodes de modélisations d'événements aléatoires ont été revus notamment pour la modélisation des délais de vente. Outre l'analyse des fondements théoriques, l'étude s'est surtout consacrée sur l'applicabilité de ces méthodes sur nos données. De plus, plusieurs animations scientifiques ont enrichi le parcours académique de cette thèse. Parmi cela, les séminaires qui sont organisés régulièrement par le laboratoire SAMM, les conférences nationales (MASHS 2011, 2012) et internationales (SFDS 2011, ESANN 2012, ICOR 2012) à travers lesquelles nous avons présenté des travaux sur des thématiques liées aux démarches méthodologiques adoptées dans le travail.

2. Le marché VO

2.1. Le mécanisme de la formation du prix

Pour identifier les facteurs quantifiables qui influencent les prix des VO et leur délai de vente ainsi que pour comprendre le mécanisme de leur formation, nous nous sommes entretenus avec des professionnels spécialisés dans la cote VO et dans la prévision des VR. Les informations ainsi obtenues nous ont amené à regrouper les facteurs cités en trois catégories :

- Les facteurs liés au véhicule et à sa perception par les consommateurs.
- Les facteurs liés à la dynamique du marché VN et à la politique de commercialisation.
- Les facteurs liés à la dynamique du marché VO et à la politique associée au marché secondaire mise en place par le constructeur.

En toute évidence, le prix des VO est lié à l'utilité future que le véhicule offre à son acheteur. Cette utilité future dépend de son **usure** qui est principalement traduite par les effets joints du **kilométrage** et de l'**âge**.

Un **effet d'obsolescence** est observé lorsque les constructeurs introduisent de nouvelles générations au sein d'une famille de véhicules alors que les anciennes continuent à être vendues sur le marché secondaire. Ces **changements incorporés** dans les nouvelles versions engendrent une dévaluation des prix des VO proportionnelle à l'amélioration de la nouvelle version.

La **réputation de la marque**, notamment sur les modèles jumeaux - des modèles faits sur la même plate-forme et qui ont essentiellement les mêmes attributs physiques mais des noms différents (ex : **Audi A3** et la **Peugeot 308**)- aurait un impact significatif sur la formation du prix des VO et donc sur leur dépréciation.

La **motorisation** a également une part importante dans la formation du prix. Soulignons qu'en France, environ 70 % du marché VN, et donc implicitement du marché VO est dominé par les véhicules à moteur diesel *(Source : CCFA*[7]*, 2007)*. Le fait que d'une part, les moteurs à diesel sont moins consommateurs de carburant par rapport aux moteurs à essence et que d'autre part, le prix du gazole demeure 10% moins cher que l'essence.*(Source : DGEMP*[8]*, 2007)* fait que le VO à motorisation diesel, aussi bien en neuf qu'en occasion peut se vendre plus cher qu'une voiture à essence et par conséquent, a une VR moins faible et reste donc un bon investissement pour une revente dans le futur.

La répartition géographique des annonces a également un impact sur l'évolution et la dispersion des prix VO. Cette affirmation a été relativisée dans [HB10] où, sur une étude comparative d'annonces sur papier et d'annonces sur internet, les auteurs ont associé le facteur régional aux coûts de recherche d'information sur le véhicule en vente, qui se sont pratiquement uniformisés entre les régions grâce à internet.

2.2. État de l'art

Les thèmes autour desquels s'articule le projet de recherche répondent à des besoins spécifiques dans un domaine très concurrentiel. Les travaux existants sur ce sujet ont été effectués en grande partie à des fins commerciales et par conséquent, il nous a été très difficile d'accéder à une bibliographie répondant à notre problématique.

La revue de l'existant révèle que le marché VO a été observé, analysé et modélisé depuis longtemps, mais surtout dans un objectif de compréhension de causes et d'effets. Il s'agit pour la plupart des explications qualitatives qui s'intéressent plutôt au comportement du prix par exemple qu'à sa valeur. Après quelques décennies de développement de modèles purement économiques, on a vu émerger de méthodes liées à des outils statistiques et mathématiques.

Dans le cadre de notre travail, les travaux qui présentent un intérêt sont ceux qui s'intéressent aux prédictions quantitatives faisant appel à des méthodes statistiques.

Nous exposerons un à un les articles que nous avons retenu. Chaque analyse d'article comprend deux parties. La première consiste à présenter une synthèse. La deuxième partie consiste à détailler notre point de vue sur la méthode présentée en soulevant des lacunes et les limitations qui n'ont pas permis leur application à notre problème, justifiant ainsi notre motivation à entreprendre de nouvelles pistes.

Dans cette approche critique, nous sommes amenés à introduire des résultats d'études empiriques présentées dans les chapitres suivants.

La cote Argus : Ayant été pour longtemps en France une valeur de référence et une base de négociation entre les acteurs du marché VO, il nous est légitime de commencer cet état de l'art par une description de la **cote Argus**[9]. Le calcul de la cote Argus consiste en une dépréciation annuelle régulière du prix VO à partir du prix à neuf en supposant une utilisation régulière du véhicule. Pour chaque type de motorisation, un kilométrage standard annuel est défini :

$$km_{standard} = \begin{cases} 15000 \text{ km/an si essence} \\ 20000 \text{ km/an si diesel} \end{cases}$$

et pour chaque version de VO, le prix P_{VO} est donné par l'équation suivante :

$$P_{VO} = \alpha_a \times P_{VN} + c \times \delta_{km}$$

7. Comité des Constructeurs Français d'Automobiles
8. Direction Générale de l'Énergie et des Matières Premières
9. Mode de Calcul connu en 2008

Segments	B1	B2		M1	M2	H1	H2
BERLINES							
Empattement (m)*	-	2,5		2,65	2,75	2,85	3
Longueur (m)	$< 3{,}7$	$> 3{,}9$		$> 4{,}2$	$> 4{,}6$	$> 4{,}9$	> 5
Largeur (m)	1,6	1,65		1,75	1,8	1,85	1,9
Puissance (ch)	60	80		110	150	200	300
Tarifs moyens (eur)	8 à 14000	11 à 18000		16 à 30000	22 à 40000	35 à 50000	> 50000
		B2/M1		M1/M2		H1	
MONOSPACES							
Empattement (m)		2,6		2,7		2,85	
Longueur (m)		4		4,3		4,7	
Largeur (m)		1,65		1,8		1,85	
Puissance (ch)		90		120		150	
Tarifs moyens (eur)		12 à 20000		20 à 30000		30 à 50000	
		B2	M0	M1	M2	H1	
BREAKS							
Empattement (m)		2,5	2,6	2,65	2,75	2,85	
Longueur (m)		4	4,1	4,4	4,7	4,9	
Largeur (m)		1,65	1,65	1,75	1,8	1,85	
Puissance (ch)		80	90	110	150	200	
Tarifs moyens (eur)		11 à 17000	13 à 19000	16 à 28000	22 à 40000	35 à 50000	

TABLE 1 – SEGMENTATION DU MARCHÉ VO EFFECTUÉ PAR L'ARGUS

où

- P_{VN} correspond au prix du véhicule à neuf,
- α_a est le coefficient de dépréciation du prix pour l'âge a (en année civile a) du véhicule,
- c constitue un ajustement monétaire associé à chaque segment

$$c = \begin{cases} c_1 & \text{si } \delta_{km} < 0 \\ c_2 & \text{si } \delta_{km} > 0 \end{cases}$$

avec $c_1 > c_2$, autrement dit, un kilométrage trop élevé est pénalisé.

On a $\delta_{km} = km_{observe} - km_{standard}$. Les coefficients α_a et c sont calculés au niveau du segment défini en TAB.1.

A ce prix P_{VO}, peut être effectué d'autres ajustements monétaires supplémentaires, correspondant à une valorisation, s'il y a lieu, des options disponibles sur le véhicule.

La règle de dépréciation des automobiles proposée par M. Boiteux, 1956 : L'auteur [Boi56] utilise les cours de l'Argus pour déduire une règle d'amortissement qui traduit de plus près les variations de la valeur des automobiles au fur et à mesure de leur vieillissement. Pour cela, il collecte des données sur les deux modèles les plus vendus à l'époque, qui sont la Primaquatre Renault et la Citroën 11 CV. Les informations exploitées sont la date de calcul de la cote(t), la date de sortie du véhicule(s) et la cote Argus(V). En supposant que $V = V(s,t)$, l'auteur propose comme fonction de dépréciation du prix $f(t-s_0) = \dfrac{V(t,s_0)}{V(s_0,s_0)}$. ($V(s_0,s_0)$ correspondant au prix neuf). L'idée est de caractériser la loi de cette dépréciation au cours du temps. Il obtient ainsi une estimation de la fonction de dépréciation qui est de la forme

$$f(a) = 1 - K \times S(a)$$

où $a = t - s$ correspond à l'âge du véhicule,
et $S(a) = \dfrac{1}{1} + \dfrac{1}{2} + \dfrac{1}{3} + \ldots + \dfrac{1}{a}$, avec $S(0) = 0$

$K \in]0;1[$ est une constante à définir et qui sera différent selon le modèle du véhicule considéré. la détermination statistique de la valeur de K doit être effectuée de manière à ajuster au mieux la fonction $f(a)$.

A partir des données que l'auteur a utilisées, il s'est avéré que K est sensiblement égale à $1/3$ pour la Renault Primaquatre et à $1/4$ pour la Citroën 11 CV . Dans sa conclusion, l'auteur affirme sans perte de généralité, que le rythme de dépréciation annuelle d'un véhicule n'est pas constant sur toute sa durée de vie et qu'elle peut être assez forte les deux premières années à partir de sa mise en circulation.

Étude effectuée par D. Purohit, 1992 : Une étude comparative du marché VN et VO dont l'objectif est d'observer dans quelles mesures le marché VN peut influencer le marché VO est présentée dans cet article [Pur92]. L'étude s'oriente surtout sur la perception du prix, de la dépréciation et du comportement de la vente lors de l'introduction des nouveaux modèles dans une même gamme de véhicule. L'auteur examine la question en évaluant l'impact des changements dans les nouvelles séries d'un même modèle de véhicules sur la dépréciation des prix VO. Parmi les analyses exposées dans l'article, celle qui présente un intérêt pour notre travail est la modélisation du taux de dépréciation du prix. Pour chaque segment de marché, le modèle proposé est sous la forme suivante

$$\log\left(\frac{P^t_{m,a}}{P^{t+1}_{m,a}}\right) = \alpha_0 + \sum_{j=1}^{k} \alpha^t_j Var^t_j + \varepsilon_i \tag{0.0.1}$$

où $\frac{P^t_{m,a}}{P^{t+1}_{m,a}}$ correspond au rapport du prix des véhicules d'un même modèle sur différentes années (t et $t+1$). Les Var^t_j sont des variables explicatives relatives à la cylindrée, la puissance, la consommation, la dimension du coffre, la longueur, la largeur, le poids et l'espace à l'arrière du véhicule. Une variable catégorielle à quatre niveaux caractérise le changement de style par rapport à la première série du même modèle. Une autre variable indique si des modèles sont jumeaux (du même constructeur ou non). Des variables exogènes, comme le prix du carburant et le coût d'entretien sont également utilisés. Il ressort de cette étude que les prix dans le marché VO réagissent aux changements introduits par le marché VN. Il a été observé aussi que les VO qui ont une forte valeur de revente, c'est-à-dire ceux dont le taux de dépréciation est le plus faible sont les modèles de véhicule relativement stables et pour lesquels il n'y a pas souvent de changement de style.

Engers, Hartmann et Stern, 2004 et 2007 : Les auteurs [EHS09] analysent les facteurs qui influent la disparité des kilométrages moyens annuels et examinent par la suite si ces mêmes facteurs peuvent expliquer la variabilité des prix des VO. Il s'agit d'une étude de la relation entre ces deux processus traduits par les deux modèles décrits ci-dessous.

Modèle 1 :

$$\log m_{ij} = \lambda_{00} + \lambda_{10} a_{ij} + \sum_{k=1}^{K} \lambda_{0k} b_{ijk} + \sum_{k=1}^{K} \lambda_{1k} b_{ijk} a_{ij} + \sum_{k=1}^{K} \lambda_{2k} b_{ijk} \min(a_{ij}, 5) + \sum_{h=1}^{H} \lambda_{3h} D_{ih} + e_i + \varepsilon_{ij} \tag{0.0.2}$$

où m_{ij} est le kilométrage annuel du véhicule j appartenant au ménage i, b_{ijk} est la variable muette pour désigner si le véhicule j est de marque k, a_{ijk} est l'age du véhicule j, D_i est un vecteur composé de H caractéristiques du ménages et ${}^t\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ est le vecteur des paramètres à estimer.

Type	cylindrée ($cc = cm^3$)
Compact	$<$ 1500 cc
Mid-Size	1500cc - 2000cc
Large-Size	2000cc - 2500cc
Luxury	$>$ 2500cc
SUV-RV(Recreational Vehicle, Camping Car)	-
Imported	-

TABLE 2 – SEGMENTATION PROPOSÉE POUR LES VO DE LOCATIONS DANS [CHO05].

Modèle 2 :

$$p_{ijt} = \mu_{it} + \delta_{ij} = \alpha_i + K_i t + \delta_{ij} \tag{0.0.3}$$

où p_{ijt} le prix moyen observé de la marque i à un age t , $\delta_{ij} \sim$ iid $\mathscr{N}(0, \sigma^2)$ (α_i, K_i) sont des paramètres à estimer et traduisant la dépréciation du prix.

Le lien entre les deux processus par la relation suivante :

$$\hat{K}_j = b_0 + b_1 \hat{\lambda}_j + u_j \tag{0.0.4}$$

Cette étude a mis en évidence que le kilométrage contient de façon intrinsèque les effets des autres facteurs qui peuvent contribuer à la dépréciation du prix. En effet, si nous pouvons observer sur le marché que certains modèles de véhicules se déprécient plus rapidement que d'autres, les résultats empiriques de l'article montrent que ces variabilités sont expliquées par le mode d'utilisation d'où le rôle fondamental du kilométrage dans la formation du prix des VO.

Sung Jin Cho, 2005 : L'article [Cho05] traite des facteurs qui déterminent les prix de vente des véhicules de location. Une segmentation des véhicules est effectuée telle qu'elle est décrite dans TAB.2. Deux modèles log-linéaire sont proposés. L'un modélise le taux de dépréciation des prix des véhicules et l'autre le prix de vente. Pour chaque segment, les modèles s'écrivent :

$$log(d_P) = b_0 + b_1 KM + b_2 KM^2 + b_3 age + b_4 age^2 + b_5 Nac + b_6 CTac + b_7 mdv + e \tag{0.0.5}$$

$$(P_{VO}) = b_0 + b_1 KM + b_2 KM^2 + b_3 age + b_4 age^2 + b_5 Nac + b_6 CTac + b_7 mdv + e \tag{0.0.6}$$

e étant un terme d'erreur aléatoire, $d_P = P_{VO}/P_{VN}$ est le taux de dépréciation correspondant au rapport du prix de vente VO au prix neuf P_{VN}, P_V : prix de vente, Nac et $CTac$ sont respectivement le nombre d'accident et le coût total de réparation lié aux accidents et mdv est le mois de vente. En conclusion, dans la dépréciation du prix, pour les petites voitures, la variable $CTac$ est significative tandis que Nac ne l'est pas. L'inverse est observé sur les grandes voitures (SUV, VO de luxe), qui d'après les résultats, donne de l'importance au nombre d'accident. La dépréciation du prix des grosses voitures et des SUV augmente avec leur âge. Il a été noté également que l'âge influe moins sur la dépréciation du prix lorsque le véhicule est récent.

Kooreman et Haan (2006) : Dans [KH06], pour un segment marque-modèle (*exemple* : Peugeot 206), le modèle de régression linéaire associé au prix du VO (P_{VO}) s'écrit :

$$ln(P_{VO}) = b_0 + b_1 \text{age}_{an} + b_2 \text{km} + b_3 \text{litre} + b_4 1_{(bv)} + b_5 1_{(km)} + b_6 1_{(Q1)} + b_7 1_{(Q2)} + b_4 1_{(Q3)} + \varepsilon \tag{0.0.7}$$

où km correspond au kilométrage effectué, litre est le volume du moteur en litre, bv correspond à la boite vitesse telle que $1_{(bv)} = \begin{cases} 1 & \text{si bv automatique} \\ 0 & \text{sinon} \end{cases}$

$1_{(km)}$ est une variable indicatrice telle que $1_{(km)} = \begin{cases} 1 & \text{si } km > 100000 \\ 0 & \text{sinon} \end{cases}$

Q_1, Q_2, Q_3 correspond respectivement au premier quadrimestre (janvier à avril), le second(mai à aout), le troisième(septembre à décembre) et $1_{(Q_1)} = \begin{cases} 1 & \text{si mois} \in \{ \text{ janvier, fevrier, mars, avril} \} \\ 0 & \text{sinon} \end{cases}$

Les informations supplémentaires que nous pouvons tirer de cet article sont que d'une part, le prix est fonction croissante du litrage et d'autre part, le fait que pour un véhicule, le fait d'avoir dépassé les 100 000 km a pour conséquence une réduction supplémentaire du prix. Notons que l'âge ici est mesuré en année civile.

Autres articles connexes : Beaucoup d'articles et d'ouvrage ont apporté d'explication sur le mécanisme du prix d' un VO. Pour plus de détails, nous pouvons nous réferer à [Gen91], [AHMftF93], [MT12], [SD70], [Mon67], [Kel88], [Kre59], [EC07] , [Ber85], [Pas13] ou [Jer08].

Ces articles montrent que le prix d'un VO est lié à l'utilité du véhicule sur la durée de vie restante et donc de sa valeur à la casse. L'idée des 100 000 km a été observée dans plusieurs articles. Cette valeur est perçue comme une valeur seuil dans la formation du prix du VO. De plus, nous avons pu voir que les prix des voitures d'occasion sont parmi les plus volatiles de l'IPC [10] . Par synthèse de ces articles, nous pouvons dire que l'industrie automobile est actuellement confrontée à une concurrence croissante sur les prix tandis que les marges de profit sont en diminution. De ce fait, l'enjeu repose fortement sur la stratégie de tarification des VO et de leur VR.

Commentaires sur les méthodes existantes : Remarquons que parmi les modèles statistiques utilisés, il est commun de voir le **logarithme du prix** comme variable dépendante et le **Km** et l'**âge** dans la liste des variables explicatives. D'ailleurs, comme il est affirmé dans [MS99] que "**quasiment toutes les études sur l'automobile ont suivi cette approche** ".

Après différents tests sur nos données, nous avons constaté que les méthodes proposées dans les articles ci-dessus avaient chacune leurs limitations notamment dans l'établissement des hypothèses.

La **kilométrage standard** proposée par **Argus** suppose une utilisation régulière de la VO tout au long de sa vie. Or, l'analyse des kilométrages que nous avons effectué dans le cadre de la détection d'outliers ainsi que les apports de certains articles contredisent cette affirmation. En effet, nous observons que deux ou trois pentes différentes régissent l'évolution du Km. L'utilisation d'un **coefficient d'ajustement du kilométrage** ne permet pas une prédiction pour des véhicules ayant un kilométrage trop éloigné des valeurs communément admises. L'application d'un **coefficient de dépréciation annuelle** nécessite une connaissance du prix VN de chaque version d'une part, et d'autre part, rend impossible la prédiction des prix pour des VO de moins de 12 mois.

L'affirmation dans [Boi56] selon laquelle le **rythme de dépréciation n'est pas constant** est vérifiée et est confirmée par un test effectué sur nos données, quoi que notre analyse ait été effectuée sur une agrégation plus fine, qui tient compte non seulement du modèle du véhicule mais aussi de la motorisation et de la carrosserie. Toutefois, l'utilisation d'une fonction de dépréciation pour prédire le prix, nous parait un peu trop approximative. En effet, considérer uniquement

10. Indice de Prix à la Consommation

l'âge du véhicule c'est négliger son usure. Or d'après nos résultats, le kilométrage est un facteur significatif dans la formation du prix. Néanmoins, cette méthode peut être utilisée pour fournir une indication sur l'ordre de prix.

Dans l'approche de [KH06], la formation du prix tient compte de la perception du produit (VO) par les consommateurs. En effet, l'évaluation de l'**âge en année civile**, aussi imprécise qu'elle puisse être, est plus perceptible pour les acheteurs que l'âge en mois. En ce qui concerne l'**effet des 100 000 km**, le fait pour un VO d'avoir franchi cette limite peut effectivement induire chez l'acheteur une certaine réticence pour l'achat et peut entraîner une baisse de prix supplémentaire pour le VO.

L'idée des **100 000 km** a été évoquée également dans [Cho05], ce qui nous a poussé a prendre en considération cet effet en l'intégrant dans notre modèle de base. Nous avons constaté que le résultat ne correspondait pas toujours à ce qui a été attendu. Nous avons abandonné cette piste puisque nous pensions qu'il est difficile de concevoir qu'un kilomètre supplémentaire puisse augmenter la probabilité d'une dégradation précoce (ex : différence entre VO ayant effectué **95 000 km** et un autre qui aurait effectué **103 000 km**).

L'étude dans [Cho05] fournit également des résultats pertinents sur **la particularité des VO de location**. Les résultats présentés peuvent bien être exploités pour la prédiction des prix VO que pour la prédiction des délais de vente. En outre, l'article confirme ce qui a été vu dans de nombreuses études sur la relation du prix avec le km et l'age.

Bien que, sur avis d'experts nous sommes contraints d'intégrer l'âge dans notre modèle, les conclusions dans [EHS09] peut nous permettre d'attribuer au kilométrage un rôle fondamental et suffisant pour la prédiction d'un prix VO. En effet, beaucoup de facteurs pouvant influer sur la prix du VO le sont à travers le km.

L'analyse dans [Pur92] a apporté beaucoup d'explication sur une problématique que nous avons rencontré lors de notre approche modélisatrice. En effet, en intégrant la variable indicatrice des options, il nous est arrivé d'obtenir lors de l'estimation, des coefficients dont le signe était négatif. Bien que cela soit inadmissible du point de vue métier, nous avons compris qu'il s'agit d'une conséquence de l'**effet d'obsolescence**. Cela a été confirmé par nos données, en testant le rapport de dépréciation comme l'auteur l'a défini.

L'analyse dans [AHMftF93] nous a permis de comprendre certains phénomènes observés sur le prix des VO ainsi que la rotation des stocks qui ont marqué l'année 2010. Il s'agit des phénomènes induits par la mise en place par le gouvernement en décembre 2008 d'une prime à la casse dont le montant a été fixée à 1 000 euros. Cette mesure est conditionnée principalement par l'acquisition d'un véhicule neuf, soit sous forme d'achat, soit sous forme d'une location longue durée avec ou sans option d'achat.

L'idée dans [EHS09] était donc d'incorporer dans le kilométrage tous les effets sociaux-économiques, influant sur le prix. Les coefficients associés au kilométrage sont ensuite utilisés dans la régression du prix VO. On y trouve également que la dépréciation du prix en fonction de l'âge due à l'obsolescence. La conclusion est que la disparité des taux de dépréciation de l'évolution du kilométrage entre les différentes marques explique la même disparité des taux de dépréciation du prix.

3. La base de données étudiée : la base VO

La base de données sur laquelle nous travaillons correspond à des annonces de véhicules d'occasion sur différents sites web en France et dans certains pays d'Europe. Cette base de données est alimentée régulièrement de manière automatique par l'équipe informatique d'autobiz.

vend skoda fabia 1.9 tdi break combi vert bouteille annee 12/2003 kilometrage reel 204000 vitres teintees av et ar , phares av leds (angel) reparations effectuees le 05/10 garage skoda amortisseur av et ar , distribution et courroie alternateur , silentblocs console , roulement av , biellette direction, disques plaquettes av et ar pneus av et ar ok , carter huile neuf . puce electronique gain de 20cv ou economie gazoil controle technique ok aucun travaux a prevoir prix : 4000 euros

FIGURE 1 – EXEMPLE1. *Lors du traitement des flux de caractère, la marque et le modèle sont bien reconnus dans FiG.1. En effet, il s'agit d'une SKODA FABIA et de motorisation DIESEL. Plusieurs identifiants que nous pouvons voir dans 3. peuvent alors correspondre à cette annonce. Pour cela, d'autres facteurs telle que l'année de sortie, l'ordre de prix peuvent être des indications supplémentaires pour la validation du référentiel.*

REN. / TWI. / 1.2L16V INIT. 4600EUROS
2001, 98 000 km, Ess 4cv, Boite Man 5 v,
3 ptes Argent Mét.Cuir
Ttes opt. Jtes all. - Toit ouv. élec Clim.
Dir ass. carnet d'entret., TBE
Élts réc : cour. de dis. - Elts à revoir : aucun

FIGURE 2 – EXEMPLE2. *Sur les informations disponibles, la marque transcrite sera RENAULT TWINGO.*

L'acquisition des données se fait en plusieurs étapes successives et complémentaires que nous citons ci-dessous par ordre chronologique.

1. Le téléchargement : un programme informatique effectue le téléchargement des petites annonces de vente de VO à partir de plusieurs sites web (*exemple* : ParuVendu, Autoscout....).
2. L'intégration dans la base VO : les petites annonces sont alors standardisées, formatées et intégrées dans une table qui constituera la table des annonces du mois.
3. Le parsing : les chaines de caractères sont analysées et traitées (suppression des informations superflues, décollage et / ou recollage des mots, réécriture des mots clés, *etc*) pour devenir des informations exploitables.
4. La validation des entrées : en fonction des mots reconnus et traités, l'annonce sera associée à un identifiant existant dans la base de données.

L'étape 4 conditionne la qualité des données. En effet, lors de ce processus d'acquisition de données, des difficultés pratiques sont rencontrées. L'insuffisance des informations recueillies rend difficile le parsing et réduit la fiabilité de la validation des entrées. Une identification incorrecte, que cela concerne la version, la carrosserie ou la boite vitesse, peut entraîner des erreurs d'interprétation importante lors de l'analyse des résultats issus de la base de données puisque qu'il s'agit d'une erreur sur la structure même des données. Dans la table ainsi constituée, nous disposons pour chaque annonce identifiée des informations dont quelques unes sont reportées en TAB.4.

Reference - ID	Version
40363	1.9 TDI 100 AMBIENTE
40365	1.9 TDI 100 CAP OUEST
40367	1.9 TDI 100 CONFORT
40369	1.9 TDI 100 ELEGANCE
40370	1.9 TDI 100 MORZINE
40371	1.9 TDI 100 PACK CLIM
40372	1.9 TDI 100 PEPS
40378	1.9 TDI 100 SPORT
40379	1.9 TDI 100 TECH DESIGN

TABLE 3 – EXTRAIT DE VERSIONS DISPONIBLES POUR LA MARQUE SKODA FABIA EN DIESEL. *Une erreur de parsing effectuée dans la validation des versions augmenterait les erreurs de prédiction des prix puisque les prix moyens associés à chacune de ces versions présentent des disparités. En exemple, pour des véhicules de PMEC en 2007 avec 180000 km, une Skoda FABIA II 1.9 TDI 105 AMBIENTE coute en moyenne 4900 euros alors que le prix moyen d'une Skoda FABIA II 1.9 TDI 105 SPORT est de 5600 euros.*

MARQUE	marque du véhicule
MODÈLE	modèle du véhicule
ÉNERGIE	type de carburant associé au véhicule
CARROSSERIE	type de carrosserie associé au véhicule
VERSION	version liée au modèle du véhicule
ORIGINE	site web de publication de l'annonce
TYPE	catégorie du vendeur ou de l'annonceur (particulier, professionnel)
DATE ANNONCE	date de parution de l'annonce
DÉPARTEMENT	département où se trouve le véhicule
PRIX	prix du véhicule
KM	kilométrage effectué
DATE PMEC	année de la première mise en circulation(PMEC)

TABLE 4 – EXTRAIT DE LISTE DES CHAMPS DISPONIBLES DANS LA BASE VO.

4. Organisation de la thèse

La thèse est présentée en trois (3) chapitres qui proposent de suivre un cheminement logique retraçant notre démarche.

Le chapitre 1 traite de la détection des outliers. Après une considération générale sur les outliers, nous présentons les outliers tels qu'ils sont présents dans la base VO. Des règles reposant sur des avis d'experts traitant surtout les outliers déterministes (erreurs de saisies et de transcription) ont été présentées. Un cadre méthodologique plus rigoureux a été développé pour établir des analyses empiriques permettant d'appréhender au mieux les phénomènes des outliers sur les données des kilométrages annuels. Des indications sur l'identification des outliers ont été proposées. Ensuite, nous proposons un cadre plus théorique par la proposition d'une méthode non paramétrique d'identification des outliers univariés. L'hypothèse faite dans ce travail, qui oriente fortement les solutions proposées, est que les lois considérées sont non bornées. L'abondante littérature et le succès croissant de la théorie des statistiques d'ordre et des valeurs extrêmes dans de nombreux domaines nous a incité à mieux examiner l'apport de cette théorie pour aborder notre problème. Pour un échantillon ordonné $(\xi_{1,N}, \xi_{2,N}, \ldots, \xi_{N,N})$, nous nous intéressons à l'identification d'au plus k outliers où $k = [\delta N]$ pour une proportion δ de l'échantillon total. En posant

$$\tau_j = \frac{\xi_{j+1,N}}{\xi_{j,N}}, \quad j = 1, \ldots, N-1$$

nous avons construit notre test sur l'ensemble $\{\tau_1, \tau_2, \ldots, \tau_{N-1}\}$. Nous pouvons espérer en absence d'outliers, que la valeur de τ_j soit faible pour tout j et qu'il y aurait K outliers si $\tau_K >> \tau_{K-1}$. Nous élaborons notre règle de décision sur la base de la statistique de test suivante :

$$\widehat{D}_{J_n} = \frac{\log(2)}{\widehat{L}_{J_n}} \max_{j=1,\cdots,J_n} j \log(\widehat{\tau}_{n-j}) \qquad \text{où} \qquad \widehat{L}_{J_n} = \text{median}\big\{(j \log(\widehat{\tau}_{n-j}))_{1 \le j \le J_n}\big\}.$$

$\widehat{D}_{J_n}$ est comparée à une valeur critique t que nous avons défini. Les propriétés asymptotiques de la statistique de test ont été établies. Une application sur des données réelles provenant de la base VO a été effectuée. Cette application concernent les résidus de régression sur les quantiles du prix des VO sur l'âge et le kilométrage ainsi que des données sur les kilométrages moyens mensuels. Les résultats obtenus sont pertinents et réalistes d'un point de vue pratique. Cette application s'est restreinte à une version particulière disponible dans la base de données et qui est la Peugeot 207 HDI Berline. Ce choix vient principalement du fait que cette version figure parmi les plus répandus sur le marché VO. Faute de temps, nous n'avions pas pu rapporter dans ce manuscrit des résultats des tests sur la totalité de la base de données.

Le chapitre 2 traite de la modélisation et de la prédiction des prix des véhicules d'occasion. L'état de l'art présenté dans l'introduction a fait apparaitre une étude quantitative assez limitée sur la prédiction des prix des VO. Ce peu d'études, l'absence de méthodologie bien définie ainsi que d'évidentes lacunes à un certain niveau d'analyse nous ont laissé entrevoir des nouvelles pistes pour aborder ce sujet.
Nous posons

$$y = g(X) + \varepsilon$$

où pour un VO, y désigne le prix et $X^T = (X_1, \ldots, X_k)$ l'ensemble de caractères invariantes dans le temps du véhicule (marque, modèle, type d'énergie, nombre de portes, ...), les caractéristiques mesurables(kilométrage, age, ...), les conditions de vente (lieu de vente, le type de marché ...). ε étant un terme aléatoire qui permet, contrairement à un VN, de prendre en compte les incertitudes dans la formation des prix d'un VO. Dans ce chapitre qui constitue la **coeur** du travail, nous exposons les démarches méthodologiques effectuées lors des différentes approches pour trouver une forme appropriée de g de telle sorte que la relation ainsi établie puisse satisfaire aussi bien les conditions d'optimalités perçu sous un angle mathématique que les contraintes imposées par la finalité commerciale du travail. Les contraintes définies par les experts sont exposées. Il s'agit d'une part, de contraintes réalistes liées à la nature de la relation entre les variables, et d'autre part, des critères de prédiction basés sur la distribution des erreurs relatives calculées à partir des résidus des estimations. Une formulation du critère de validation sur la base de ces contraintes est présentée. Les problèmes rencontrés, les avantages et les limites de chaque méthode sont détaillés. Quatre modèles ont fait l'objet de comparaison afin d'apprécier le réalisme et les apports du modèle retenu. De par leur spécificité, les VO de luxes sont exclus de notre étude.

Le chapitre 3 traite de la modélisation et de la prédiction des délais de vente des véhicules d'occasion. Les enjeux économiques, commerciaux du secteur automobile ainsi que l'activité de l'entreprise Autobiz qui ne peut se limiter à la prédiction des prix ont donc conduit à étudier les phénomènes liés à la vente des VO. Le délai de vente ainsi que la rotation de stock constitue un indicateur important dans la fluidité du marché de l'occasion. Nous pouvons supposer qu'il existe des incertitudes ou des variabilités de sources diverses qui peuvent être liées au nombre observé de VO vendus dans une certaine période. Ces variabilités peuvent en partie être le choix économique des consommateurs ou les politiques des vendeurs, nous ne prétendons pas les modéliser. Toutefois, ces effets aléatoires nous donne la possibilité de proposer des méthodes de modélisation statistique pour répondre à cette question. Prédire un temps de vente ou temps moyen de rotation revient à estimer la probabilité de la rotation moyenne de vente et, le cas échéant, à estimer les caractéristiques qui déterminent la rotation des VO : la marque, la motorisation, le volume du marché, le type d'énergie, la tranche d'âge, le kilométrage observé, la tranche de prix et, éventuellement la zone de vente. Les connaissances sur le mécanisme des prix nous ont permis d'effectuer la segmentation du marché VO afin de proposer une estimation des délais de vente des VO selon leur segment d'appartenance. Après une définition d'un profil de référence pour les délais de vente, plusieurs problématiques ont été traitées. Il a été question de modéliser et les délais de vente de VO et le nombre de VO vendus en une période T.

Dans chaque démarche proposée, qui est toujours associée à la connaissance du phénomène étudié, les résultats obtenus confirment la pertinence de notre raisonnement et l'adéquation de la démarche au problème posé.

Le rapport est conclu par une synthèse et un bilan des travaux présentés. Des perspectives d'extension et d'approfondissement sont proposées. Ces perspectives d'extension correspondent d'une part à des possibles améliorations des résultats exposés et d'autre part à une généralisation de la méthodologie utilisée pour des données sur les véhicules d'occasion provenant d'autres pays. De nouvelles pistes envisageables pour de future recherche sont également présentées.

Chapitre 1

Les outliers dans la base VO

La collecte automatique des données et le fait qu'elles soient de différentes origines créent une situation où la présence des outliers dans notre base de données devient incontournable.

1.1 Une considération générale sur les outliers

De nombreux auteurs ont cherché à décrire ce qu'est un outlier et les définitions fournies ont évolué au cours du temps [Gru69] [BL94] [Haw80]. La définition communément admise est telle que pour n réalisations $y_1, \ldots, y_n$ d'une variable aléatoire Y où $F_Y(y) = \mathbb{P}(Y \leq y)$ l'observation y_k, $k \in \{1, \ldots, n\}$ est un outlier si y_k n'a pas été généré par F_Y. Une définition plus objective est proposée dans [MGMRPA90] en précisant qu'un outlier est une observation qui dévie nettement du comportement général par rapport au critère sur lequel l'analyse est réalisée. C'est sur cette définition que nous conduisons notre étude sur les outliers.

Tout comme la manière de les définir, les explications relatives à l'apparition des outliers dans un ensemble de données ont évolué et ont fait apparaître des causes bien distinctes liées à leur nature [BC83]. Nous pouvons distinguer deux sources d'apparition pour les outliers. Une des origines est liée à une variabilité inhérente et une autre peut être liée à des circonstances bien déterminées. La variabilité inhérente correspond à l'expression par laquelle les observations varient de manière aléatoire à travers la population. Suivant leur nature, ces outliers peuvent être univariés ou multivariés.

La détection d'outliers univariés concerne principalement l'identification des valeurs extrêmes de l'échantillon. Les valeurs aberrantes multivariées ou relationnelles sont définies comme des observations non conformes aux relations qui existent entre les différents éléments [GK72]. L'outil de base le plus populaire pour la détection d'observations aberrantes dans un échantillon multivarié est la distance de Mahalanobis. Pour plus de détails, nous pouvons nous reporter à [MK85] [Pen96]. Notons que deux phénomènes peuvent surgir lors de l'application des méthodes de détection des outliers [HBK84] [TM72] :

(1) le *masking effect* où un outlier peut passer inaperçu,
(2) le *swamping effect* où une observation non discordante est identifiée comme un outlier.

Visualisation des outliers : La visualisation nous permet de comprendre le phénomène étudié. Pour la base de données qui nous préoccupe, les corrélations entre les différents éléments

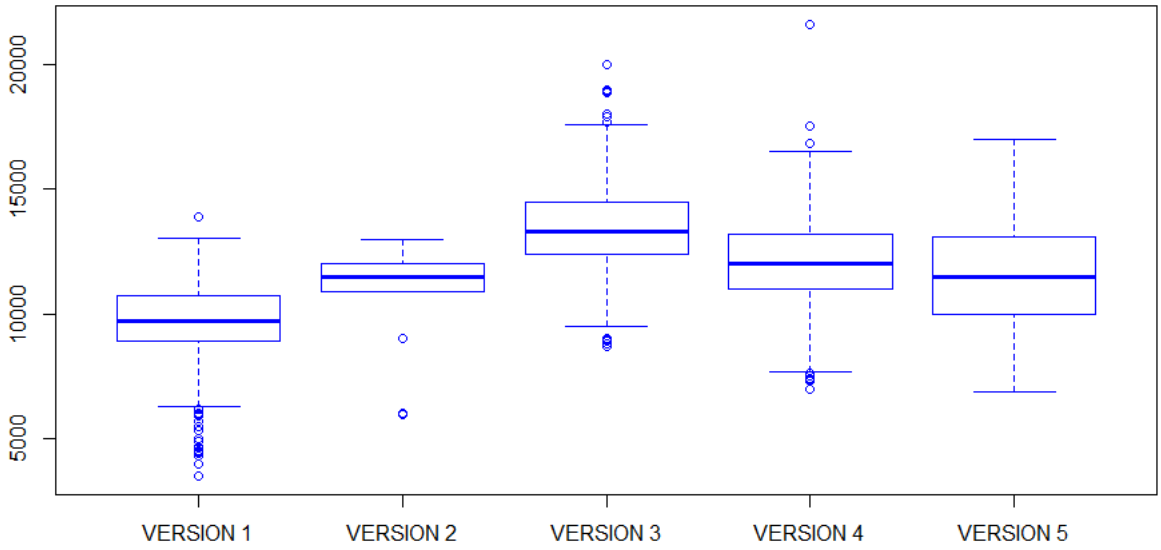

FIGURE 1.1 – BOXPLOT DES PRIX POUR CINQ VERSIONS DE LA MARQUE PEUGEOT 207. *Pour des véhicules appartenant à une même version et ayant tous la même année de première mise en circulation, nous constatons qu'il y a des observations qui se démarquent de façon évidente du reste des données. Une connaissance à priori de l'ordre des prix d'une Peugeot 207 nous permet de supposer que ces valeurs proviennent d'une erreur de transcription et/ ou d'un mauvais référencement. Le même raisonnement peut être suivi pour les observations égales à zéro.*

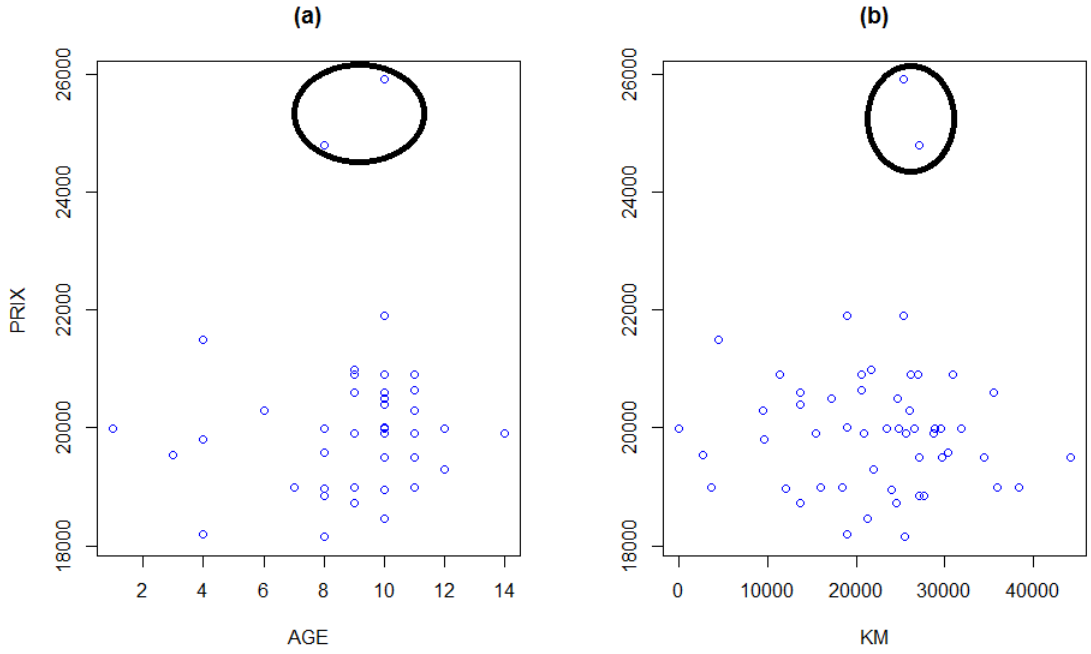

FIGURE 1.2 – LES OUTLIERS MULTIVARIÉS. *Nous observons des observations qui se démarquent significativement des autres en présentant une discordance dans la relation entre le prix du véhicule et son âge (a) ainsi que le prix et le kilométrage (b).*

peuvent fournir de bonnes indications sur la nature des relations entre ceux-ci. Une illustration des outliers est observée en FIG.1.1, FIG.1.2 et FIG.1.3.

1.2 Les traitements des données

1.2.1 Les premières restrictions

Sur l'avis des experts, des premières restrictions sont effectuées sur nos données afin de réduire les outliers d'amplitude qui sont principalement d'origine déterministe. Notons que le traitement de tels outliers n'est pas du domaine de l'analyse statistique mais nécessite tout simplement du bon sens ainsi qu'une connaissance *a priori* de la nature et de la structure des données.

Si $\mathscr{O}_{prix}$, $\mathscr{O}_{km}$ et $\mathscr{O}_{age}$ désignent respectivement l'ensemble des outliers supposés déterministes pour chacune des variables $prix$, km et $âge$, nous définissons

- $\mathscr{O}_{prix} = \{px_i, i = 1, \ldots, N_0 : (px \leq 900) \vee (px \geq 350000)\}$
- $\mathscr{O}_{Km} = \{km_i, i = 1, \ldots, N_0 : (km \leq 0) \vee (km \geq 500000)\}$
- $\mathscr{O}_{Age} = \{ag_i, i = 1, \ldots, N_0 : (ag \leq 1) \vee (Age \geq 120)\}$

Ces restrictions se justifient par le fait que les véhicules de plus de 10 ans ne présentent pas

Annonce Urgent - Affaire
9 000 euros
Annonce publiée le 27 juin 2011 par un Particulier > 34500 - Béziers
Couleur int : noir ext : bleu moyen - Excellent état 1ere main
- aucun frais à prévoir - 4 portes, toute option de série + ESBS ABS Clim auto, ordinateur de bord, vitres électriques, peinture metal -
Vend car départ à l'étranger - a débattre si paiement au comptant
Caractéristiques de l'annonce
- Marque : Peugeot
- Modèle : 407
- Kilométrage : -100000
- Année du modèle : 2007
- Carburant : Diesel

FIGURE 1.3 – EXEMPLE D'ANNONCE AVEC INSUFFISANCE D'INFORMATION. *Le genre d'aberration induite par cette annonce n'est pas une erreur de transcription ou d'exécution mais provient d'une insuffisance d'information de la part de l'annonceur qui au lieu de mentionner le kilométrage réel, a seulement précisé que le véhicule a fait moins de* $100000km$. *L'annonce dans 1.3 est un exemple qui donnerait une valeur du* Km *égale à* -100000.

d'intérêt pour notre étude puisqu'au delà de cet âge, une cote VO ne peut plus être fournie ; il en est de même pour les VO dont le Km va au-delà de 500000. Le minimum fixé pour le *prix* correspond à la prime à la casse qui a été instituée en France sur une certaine période.

Notons que le seul traitement que nous avons pu faire sur les outliers est de supprimer l'observation suspecte.

1.2.2 Règles empiriques à partir de l'analyse des courbes de quantile des kilométrages annuels

Le kilométrage enregistré pour un véhicule traduit le mode d'utilisation de ce véhicule comme il est affirmé dans [EHS09] [CM04]. Une grande partie de la variabilité au sein d'un ensemble de valeurs observées pour le kilométrage est expliquée par le type d'énergie du véhicule comme nous pouvons observer en FIG.1.4. En observant la répartition des valeurs associées au kilométrage selon l'âge du véhicule, l'identification d'un profil caractérisant l'évolution du kilométrage à travers les années d'utilisation est envisageable. Nous pouvons supposer ainsi que pour $i = 1, \ldots, N$ profils possibles, il existe une fonction X, continue et dérivable au moins jusqu'à l'ordre 2 telle que

$$\begin{aligned} X_i : \quad \mathscr{T} &\longrightarrow \mathscr{K} \\ t &\longmapsto X_i(t), i = 1, \ldots, N. \end{aligned}$$

où

- $\mathscr{K}$ est un intervalle fermé et borné de $\mathbb{R}$.
- $t \in \mathscr{T} = \{t_{min}, t_2, \ldots, t_{max}\}$ correspond au nombre d'années écoulées depuis l'année de la PMEC [1] du véhicule.

Chacune de ces X_i traduit le même phénomène et par conséquent, leur allure devra être relativement invariante, tout au moins par rapport à un profil de référence X_R. En supposant qu'en un instant t, pour n VO de même année de PMEC, chaque kilométrage observé $X_i(t), i = 1, \ldots, n$ est une réalisation indépendante d'une variable aléatoire et que, en absence d'outliers

1. Première Mise En Circulation

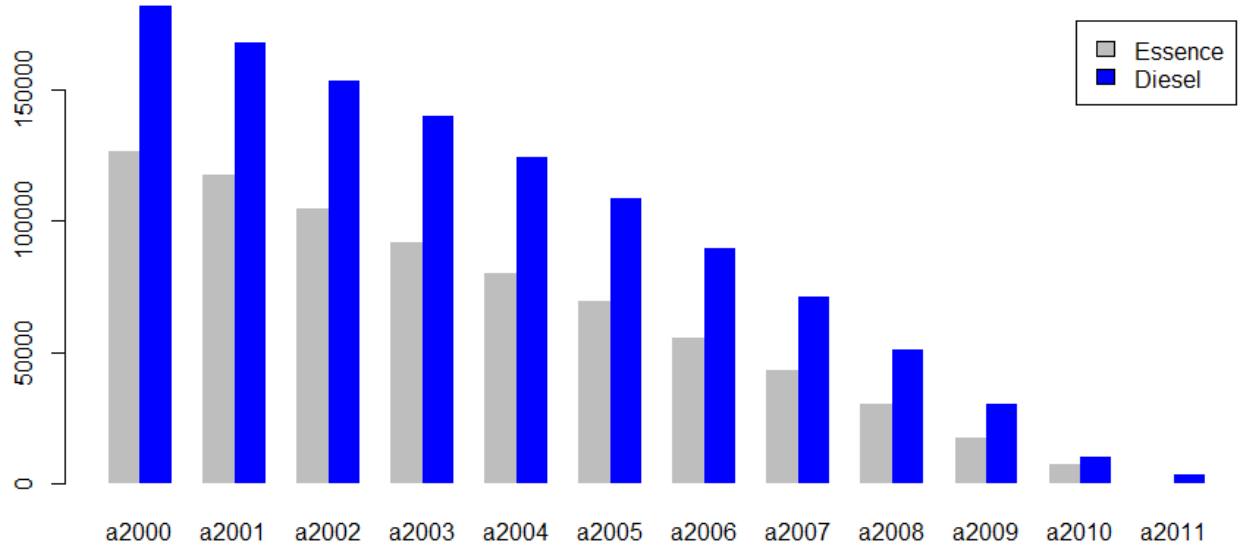

FIGURE 1.4 – KILOMÉTRAGE MOYEN PAR TYPE D'ÉNERGIE *Chaque couple de barre montre le kilométrage moyen calculées pour des véhicules dont l'année de PMEC correspondant par type d'énergie. Pour chaque couple de figure observées, nous voyons clairement que les véhicules diesel présente un kilométrage plus élevé que les essences.*

$$X_i(t) = \delta_i X_R(t) + v_i(t), \quad i = 1 \ldots, N.$$

où $\delta_i, i = 1, \ldots, N$ sont des réels vérifiant propriétés suivantes :

(P1) $\forall i = 1, \ldots, N : \delta_i \neq 0$.

(P2) Pour tout $k \leq j$, $\delta_k < \delta_j$.

Les termes d'erreurs v_i désignent des fluctuations aléatoires qui englobent les perturbations de diverses natures ainsi que l'influence de variables dont nous n'avons pas connaissance. Ils ne sont pas indépendants de t et sont tels que $\mathbb{E}[v_i(t)] = 0$.

Dans toute la suite, chaque profil $i, i = 1, \ldots, N$ correspond à un niveau de quantile $\alpha \in (0, 1)$.

Nous nous proposons de trouver la forme de la fonction y_R qui s'ajustera au mieux à la courbe de référence X_R. En absence d'outliers, chacune des courbes X_i associée à n'importe quel niveau de quantile i devrait respecter l'allure de y_R.

L'étude conduite ici concerne les kilométrages observés sur 856.731 véhicules à diesel dont l'année de PMEC est comprise entre 1998 et 2008.

Estimation de la courbe de référence $X_R(t)$. Nous reprenons l'idée exposée dans [FM01] permettant d'obtenir un estimateur robuste d'une courbe de référence. Pour chaque valeur associée à t, les quantiles sont estimés de façon usuelle et la courbe de référence $\widehat{X}_R(t)$ s'obtient par la formule ci-après

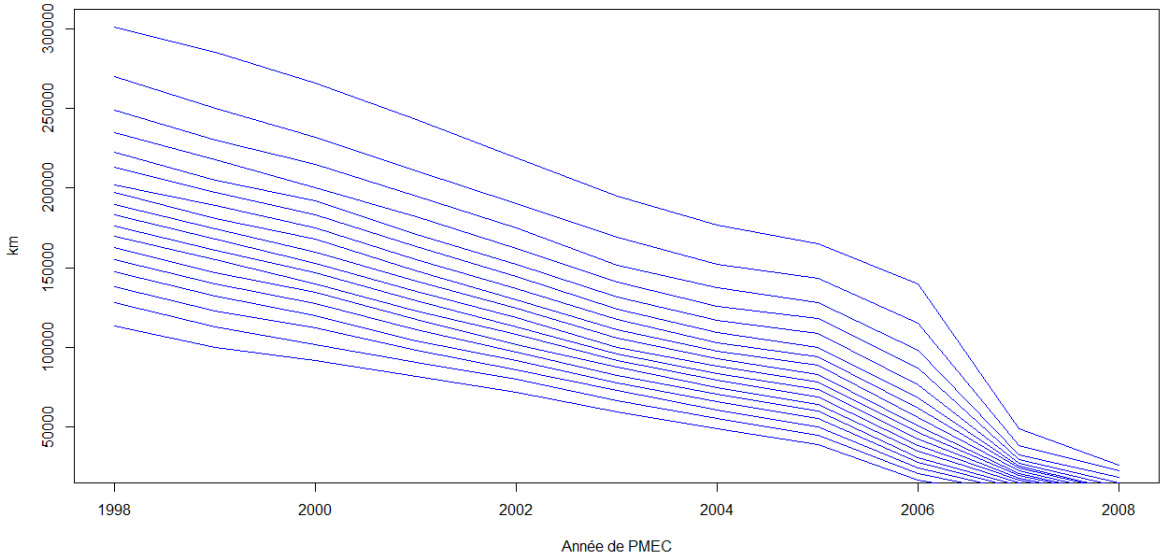

FIGURE 1.5 – COURBE DES QUANTILES DES KILOMÉTRAGES ANNUELS. *chacune des courbes représente* X_i *pour* $i = \alpha \in \{5, 10, 15, \ldots, 90, 95\}$. *Il ressort de l'analyse graphique que les courbes associées aux niveaux de quantiles compris dans l'intervalle* $[0.05, 0.95]$ *présentent la même allure, croissante selon* t, *avec des points d'inflexion très remarquables qui se situent à* $t = 1$ *et entre* $t = 2$ *et* $t = 5$ *(à la troisième année) pouvant s'expliquer par un changement de main. Nous pouvons déduire également une utilisation plus intensive du véhicule pendant les trois premières années de vie avec un légère modération entre 3 et 5 ans et une utilisation régulière jusqu'à la dernière année d'observation.*

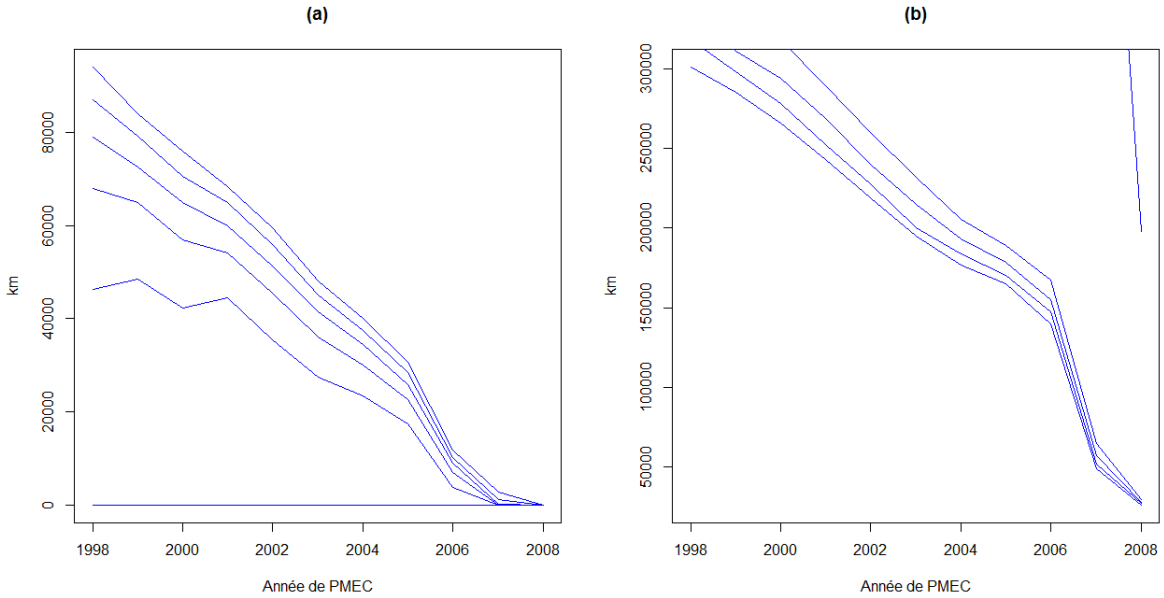

FIGURE 1.6 – ZOOM SUR LES COURBES ASSOCIÉES AUX QUANTILES EXTRÊMES. *Nous observons que pour la partie gauche des observations, certaines courbes sont clairement incontrôlables par rapport à la tendance générale. Il s'agit des courbes associées à* $\alpha = \{0, 1, 2\}$ *ainsi que les courbes associées à* $\alpha = 99$ *et* $\alpha = 100$.

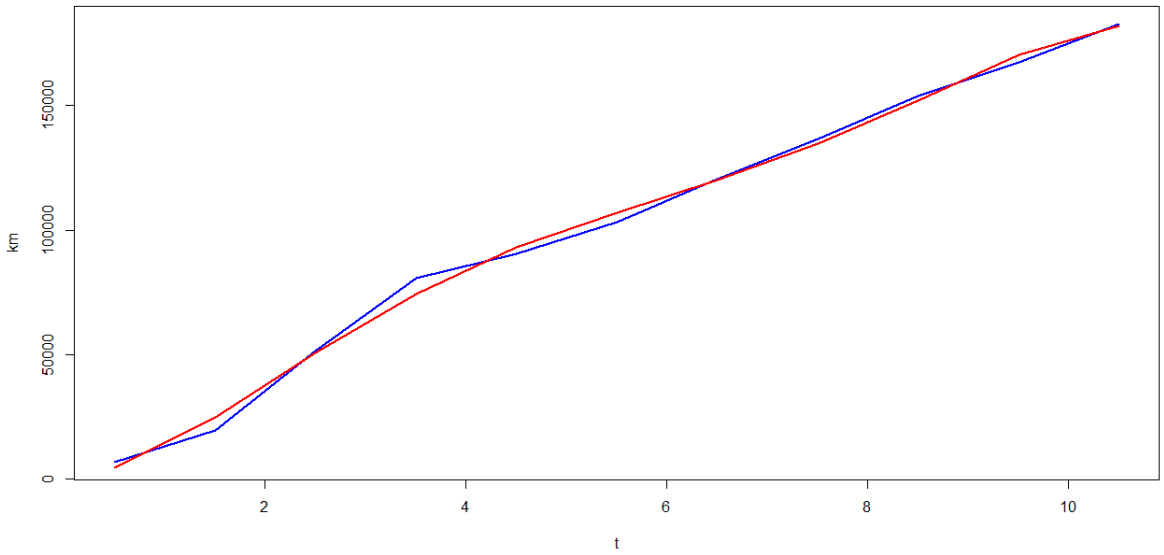

FIGURE 1.7 – AJUSTEMENT DE LA COURBE DE RÉFÉRENCE. — *indique la courbe* $\widehat{X}_R(t)$ *et* — *la fonction* y_R. *L'allure de la fonction qui s'ajuste au mieux à la courbe de référence est un polynôme de degré 5 définie telle que* $y_R = a_1 + a_2t^2 + a_3t^3 + a_4t^4 + a_5t^5$.

$$\widehat{X}_R(t) = \frac{1}{N - [N\alpha]} \sum_{i=1}^{N-[N\alpha]} X_i(t)$$

Dans notre cas, et en se basant sur les graphiques disponibles, nous posons $\alpha = 0.10$. Cette restriction nous assure une qualité de données avec une plus grande fiabilité.

Approximation de $\widehat{X}_R(t)$ par $y_R(t)$. $\widehat{X}_R(t)$ est alors lissé selon une fonction y_R qui déterminera l'allure de l'évolution du kilométrage annuel et telle que

C1 : y_R *est croissante sur l'ensemble des valeurs prises par* t.

C2 : y_R *admet des points d'inflexion sur l'ensemble des valeurs prises par* t.

Nos données étant discrètes, le lissage de $\widehat{X}_R(t)$ consiste en une approximation polynomiale sous les conditions $C1$ et $C2$ énoncées ci-dessus. Nous pouvons observer l'ajustement obtenu en FIG.1.7.

Identification des outliers. Pour identifier les outliers, nous nous focaliserons sur les données associées aux centiles compris entre 0 et 5 et entre 95 et 100. Toutes les courbes $\widehat{X}_i(t)$ associées à ces niveaux de quantiles sont alors ajustées selon une fonction y_i ayant la même structure que y_R. L'écart entre $\widehat{X}_i(t)$ et y_i évalué par le $RMSE = \sqrt{\frac{\sum_{t=1}^{K}(\widehat{X}_i(t)-y_i)^2}{K}}$ nous permettra de de statuer si l'allure générale est toujours respectée par la courbe observée. Les valeurs calculées du RMSE pour les données que nous utilisons pour cette étude sont représentées en FIG.1.8. Au vu des graphiques, les outliers se situeraient au delà du $99.85^{ème}$ centile et en deçà du $1.5^{ème}$ centile.

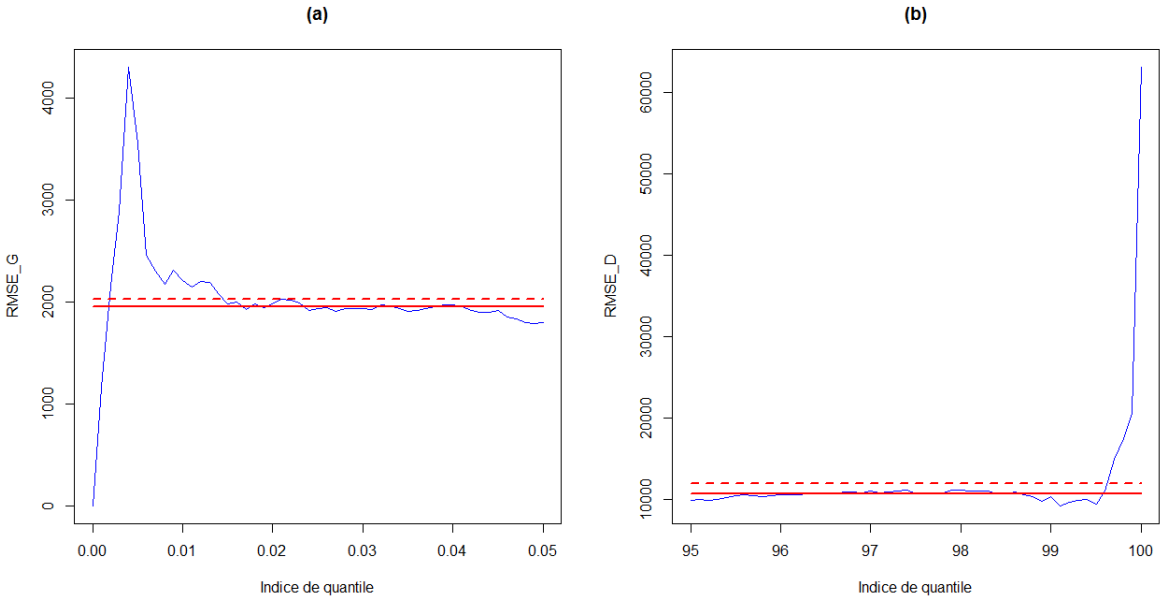

FIGURE 1.8 – VALEUR DU RMSE POUR L'AJUSTEMENT DES 5% PLUS PETITES (A) ET 5% PLUS GRANDES OBSERVATIONS (B). *Nous observons les valeurs prises par le RMSE qui nous indiquent les erreurs de prédiction. Les tirets indiquent respectivement le RMSE moyen (- -) et médian (-). Si la courbe prédite s'ajuste toujours aussi bien aux données, ce critère ne devrait pas trop dévier d'une valeur centrale. Cette déviation commence à se manifester à partir du centile* 99.85 *et en deçà du centile* 1.5.

1.3 Détection non-paramétrique des outliers univariés pour des lois de distribution à support positif et non borné.

Ces analyses empiriques, l'appréciation de la base de données ainsi que la conscience que la suite de notre étude dépendrait fortement de la qualité des données ont fait ressentir la nécessité d'une étude théorique et approfondie sur la manière d'identifier les outliers. Nous développons ainsi une statistique de test que nous présentons dans la section suivante. Cette étude a donnée lieu à la production d'une article écrit conjointement avec Jean-Marc Bardet. L'application du test a été effectuée sur des résidus de la régression du prix du véhicule sur les variables kilométrage et âge d'une part et sur les données du kilométrage moyen mensuel d'autre part. Cette application concerne uniquement la version "Peugeot 207 HDI Berline" qui constitue un échantillon extrait de la base VO. Le test a été pertinent pour ces données, cependant des résultats complémentaires sur la généralisation du test sur l'ensemble de la base de données n'ont pas pu être produits dans le cadre de ce travail et seraient profitables à de futurs développements.

A new non-parametric detector of univariate outliers for distributions with positive unbounded support

Jean-Marc Bardet
and
Solohaja-Faniaha Dimby*
S.A.M.M., Université de Paris 1 Panthéon-Sorbonne
90, rue de Tolbiac, 75634, Paris, France

Abstract

The purpose of this paper is the construction and the asymptotic property study of a new non-parametric detector of univariate outliers. This detector, based on a Hill's type statistics, is valid for a large set of probability distributions with positive unbounded support, for instance for the absolute value of Gaussian, Gamma, Weibull, Student or regular variations distributions. We illustrate our results by numerical simulations which show the accuracy of this detector with respect to other usual univariate outlier detectors (Tukey, MADE or Local Outlier Factor detectors). An application to real-life data allows to detect outliers in a database providing the prices of used cars.

Keywords: order statistics; Hill-type estimator; non-parametric test.

*The authors gratefully acknowledge the enterprise Autobiz

1 Introduction

Let $(X_1, \cdots, X_n)$ be a sample of positive independent identically distributed random variables with unbounded distribution. The aim of the article is to provide a non-parametric outlier detector among the "large" values of $(X_1, \cdots, X_n)$

Remark 1. *If we would like to detect outliers among the "small" values of* $(X_1, \cdots, X_n)$*, it could be possible to consider* $\max(X_1, \cdots, X_n) - X_i$ *instead of* X_i*, for* $i = 1, \cdots, n$*. Moreover, if* X_i*,* $i = 1, \cdots, n$*, are not positive random variables, such as in the case of regression residuals, we can consider* $|X_i|$ *instead of* X_i*.*

There exist numerous outlier detectors in such a framework. Generally, it consists on statistics directly applied to each observation which decides if this observation can be considered or not as an outlier (see for instance the books of Hawkins, 1980, Barnett and Lewis, 1994, Rousseeuw and Leroy, 2005, or the article of Beckman and Cook, 1983). The most used, especially in the case of regression residuals, is the Student-type detector (see a more precise definition in Section 3). However it is a parametric detector which is theoretically defined for a Gaussian distribution. Another famous other detector is the robust Tukey detector (see for example Rousseeuw and Leroy, 2005). Even it is frequently used for non-Gaussian distribution, its threshold is computed from quartiles of the Gaussian distribution. Finally, we can also cite the MAD_e detector which is based on the median of absolute value of Gaussian distribution (see also Rousseeuw and Leroy, 2005).
Hence all the most used outlier detectors are based on Gaussian distribution and they are not really accurate for less smooth distributions (for regression residuals, we can also cite the Grubbs-Type detectors introduced in Grubbs, 1969, extended in Tietjen and Moore, 1972). Such a drawback could be avoided by considering a non-parametric outlier detector. However there exist few non-parametric outlier detector. We could cite the Local Outlier Factor (LOF) introduced in Breunig *et al.* (2000) and also valid for mutlivariate outliers. Unfortunately a theoretical or numerical procedure for choosing the number k of cells and its associated threshold does still not exist. Other detectors exist essentially based on a classification methodology (see for instance Knorr *et al.*, 2000).
The order statistics provides an interesting starting point for defining a non-parametric detector of outlying observations. Hence, Tse and Balasooriya (1991) introduced a detector based on first differences of order statistics, but only for the exponential distribution.

Recently, a procedure based on the Hill's estimator was developed for detecting influential data point in Pareto-type distributions (see Hubert *et al.*, 2012). The Hill's estimator (see Hill, 1975) has been defined from the following property: first, define the order statistics from $(X_1, \cdots, X_n)$:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}. \tag{1.1}$$

Then for Pareto type distributions and more generally for distributions in the maximum attraction domain of extreme's theory, the family of r.v. $\big(\log(X_{(n-j+1)}) - \log(X_{(n-j)})\big)_{1 \leq j \leq k(n)}$ is asymptotically (when $\min\big(k(n)\ ,\ n-k(n)\big) \underset{n\to\infty}{\longrightarrow} \infty$) a sample of independent r.v. following exponential distributions. This induces the famous Pareto quantile plot (see Beirlant *et al.*, 1996 or Embrechts *et al.*, 1997), frequently used for exhibiting the behavior of the mean excess. The mean of this sample provides an estimator of the Parato power, but this requires an optimal choice of the tuning parameter $k(n)$.

Here we will use this previous property for detecting a finite number of outliers among the sample $(X_1, \cdots, X_n)$. Indeed, an intuitive idea is the following: the presence of outliers generates a jump in the family $\big(X_{(n-j+1)}/X_{(n-j)}\big)_j$ and therefore in the family $\big(\log(X_{(n-j+1)}) - \log(X_{(n-j)})\big)_j$. Hence an outlying data detector can be realized when the maximum of this family exceed a threshold (more details are notably given in (2.8) or (2.12)).

In the sequel we give some assumptions on probability distributions for applying this new test of outlier presence and providing an estimator of the number of outliers. It is relevant to say that this test is not only valid for Pareto-type distribution, but more generally to a class of regular variations distributions (for instance Pareto, Student or Burr probability distributions) and also to numerous probability distributions with an exponential decreasing (such as Gaussian, Gamma or Weibull distributions). Hence our new outlier detector is a non-parametric estimator defined from an explicit threshold, which does not require any tuning parameter and can be applied to a very large family of probability distributions.

Several Monte-Carlo experiments realized for several probability distributions attest of the good accuracy of this new detector. It is compared to other famous outlier detectors and the simulation results obtained by this new detector are extremely convincing especially for not detecting false outliers. Moreover, an application to real-life data (price, mileage and age of used cars) is realized, allowing to detect two different kinds of outliers.

We organized our paper as follows. Section 2 contains the definitions and main results. Section 3 is devoted to Monte-Carlo experiments, Section 4 presents the results of the numerical application on used car variables and the proofs of this paper are detailed in Section 5.

2 Definition and main results

For $(X_1, \cdots, X_n)$ a sample of positive i.i.d.r.v. with unbounded distribution, define:

$$G(x) = \mathrm{P}(X_1 > x) \qquad \text{for } x \in \mathbb{R}. \tag{2.1}$$

It is clear that G is a decreasing function and $G(x) \to 0$ when $x \to \infty$. Hence, define also the pseudo-inverse function of G by

$$G^{-1}(y) = \sup\{x \in \mathbb{R},\ G(x) \geq y\} \qquad y \geq 0. \tag{2.2}$$

G^{-1} is also a decreasing function. Moreover, if the support of the probability distribution of X_1 is unbounded then $G^{-1}(x) \to \infty$ when $x \to 0$.

Now, we consider both the following spaces of functions:

- $A_1 = \Big\{ f : [0,1] \to \mathbb{R}$, such as for any $\alpha > 0$, $f(\alpha x) = f_1(x)\Big(1 + \dfrac{f_2(\alpha)}{\log(x)} + O\big(\dfrac{1}{\log^2(x)}\big)\Big)$ when $x \to 0$ where $f_1 : [0,1] \to \mathbb{R}$ satisfies $\lim_{x\to 0} f_1(x) = \infty$ and f_2 is a $\mathcal{C}^1([0,\infty))$ diffeomorphism $\Big\}$.
- $A_2 = \Big\{ g : [0,1] \to \mathbb{R}$, there exist $a > 0$ and a function $g_1 : [0,1] \to \mathbb{R}$ satisfying $\lim_{x\to 0} g_1(x) = \infty$, and for all $\alpha > 0$, $g(\alpha x) = \alpha^{-a}\, g_1(x)\,\big(1 + O\big(\frac{1}{\log(x)}\big)\big)$ when $x \to 0 \Big\}$.

Exemple 2.1. *We will show below that numerous famous "smooth" probability distributions such as absolute values of Gaussian, Gamma or Weibull distributions satisfy $G^{-1} \in A_1$. Moreover, numerous heavy-tailed distributions such as Pareto, Student or Burr distributions are such as $G^{-1} \in A_2$.*

Using the order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, define the following ratios (τ_j) by:

$$\tau_j = \frac{X_{(j+1)}}{X_{(j)}} \quad \text{if } X_{(j)} > 0, \text{ and } \tau_j = 1 \text{ if not, for any } j = 1, \cdots, n-1 \tag{2.3}$$

$$\tau'_j = (\tau_j - 1)\log(n) \qquad \text{for any } j = 1, \cdots, n-1 \tag{2.4}$$

Proposition 1. *Assume $G^{-1} \in A_1$. Then, for any $J \in \mathbb{N}^*$, and with $(\Gamma_i)_{i\in\mathbb{N}^*}$ a sequence of r.v. satisfying $\Gamma_i = E_1 + \cdots + E_i$ for $i \in \mathbb{N}^*$ where $(E_i)_{j\in\mathbb{N}^*}$ is a sequence of i.i.d.r.v. with exponential distribution of parameter 1,*

$$\max_{j=n-J,\cdots,n-1}\{\tau'_j\} \xrightarrow[n\to\infty]{\mathcal{D}} \max_{k=1,\cdots,J}\big\{f_2(\Gamma_k) - f_2(\Gamma_{k+1})\big\}. \tag{2.5}$$

Now, we consider a particular case of functions belonging to A_1. Let A'_1 the following function space:

$$A'_1 = \big\{f \in A_1 \text{ and there exist } C_1, C_2 \in \mathbb{R} \text{ satisfying } f_2(\alpha) = C_1 + C_2 \log\alpha \text{ for all } \alpha > 0\big\}.$$

Exemple 2.2. *Here there are some examples of classical probability distributions satisfying $G^{-1} \in A'_1$:*

- **Exponential distribution $\mathcal{E}(\lambda)$:** *In this case, $G^{-1}(x) = -\log(x)$, and this implies $G^{-1} \in A'_1$ with $f_1(x) = -\frac{1}{\lambda}\log(x)$ and $f_2(\alpha) = \log\alpha$ ($C_1 = 0$ and $C_2 = 1$).*
- **Gamma distributions $\Gamma(a)$** *In this case, $G(x) = \frac{1}{\Gamma(a)}\int_x^\infty t^{a-1}e^{-t}dt$ for $a \geq 1$ and we obtain, using an asymptotic expansion of the Gamma incomplete function (see Abramowitz and Stegun, 1964):*

  $$G^{-1}(x) = \frac{1}{\Gamma(a)}\big(-\log x + (a-1)\log(-\log x)\big) + O(|(\ln x)^{-1}|) \qquad x \to 0.$$

  As a consequence, we deduce $G^{-1} \in A'_1$ with

  $$f_1(x) = \frac{1}{\Gamma(a)}\big(-\log x + (a-1)\log(-\log x)\big) \quad \text{and} \quad f_2(\alpha) = \log\alpha \ \ (C_1 = 0 \text{ and } C_2 = 1).$$

- **Absolute value of standardized Gaussian distribution $|\mathcal{N}(0,1)|$:** *In this case, we can write $G(x) = \frac{2}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt = erfc(x/\sqrt{2})$, where $erfc$ is the complementary Gauss error function. But we know (see for instance Blair et al., 1976) that for $x \to 0$, then $erfc^{-1}(x) = \frac{1}{\sqrt{2}}\Big(-\log(\pi x^2) - \log(-\log x)\Big)^{1/2} + O(|(\ln x)^{-1}|)$. As a consequence, for any $\alpha > 0$,*

  $$erfc^{-1}(\alpha\, x) = erfc^{-1}(\alpha\, x)\Big(1 + \frac{\log\alpha}{2\log x} + O(|(\ln x)^{-2}|)\Big) \tag{2.6}$$

  Then, we obtain

  $$G^{-1}(\alpha\, x) = G^{-1}(x)\Big(1 + \frac{\log\alpha}{2\log x} + O(|(\ln x)^{-2}|)\Big) \qquad x \to 0.$$

Consequently $G^{-1} \in A_1'$ *with*

$$f_1(x) = \sqrt{-2\log x - \log(-\log x) - 2\log \pi} \qquad \textit{and} \qquad f_2(\alpha) = \frac{1}{2}\log \alpha,$$

implying $C_1 = 0$ *and* $C_2 = \frac{1}{2}$.

- **Weibull distributions:** *In this case, with* $a \geq 0$ *and* $0 < b \leq 1$, $G(x) = e^{-(x/\lambda)^k}$ *with* $\lambda > 0$ *and* $k \in \mathbb{N}^*$, *for* $x \geq 0$. *Then it is obvious that* $G^{-1}(x) = \lambda\big(-\log x\big)^{1/k}$ *and therefore* $G^{-1} \in A_1'$ *with* $f_1(x) = \lambda\big(-\log x\big)^{1/k}$ *and* $f_2(\alpha) = \frac{1}{k}\log \alpha$ *(implying* $C_1 = 0$ *and* $C_2 = 1/k$*).*

When $G^{-1} \in A_1'$, it is possible to specify the limit distribution of (2.5). Hence, we show the following result:

Proposition 2. *Assume that* $G^{-1} \in A_1'$. *Then*

$$\mathrm{P}\Big(\max_{j=n-J,\cdots,n-1}\{\tau_j'\} \leq x\Big) \underset{n\to\infty}{\longrightarrow} \prod_{j=1}^{J}\big(1 - e^{-jx/C_2}\big). \tag{2.7}$$

Such a result is interesting since it provides the asymptotic behavior of normalized and centered ratios τ_i which are a vector of independent exponential r.v. However the parameters of these exponential distributions are different. Thus, if we consider the "natural" outlier detector

$$\widehat{T} = \max_{j=n-J,\cdots,n-1}\{\tau_j'\}, \tag{2.8}$$

the computation of a threshold allowing to detect an outlier requires to consider the function $y \in [0,\infty[\mapsto P(y) = \prod_{j=1}^{J}\big(1 - e^{-jy}\big)$. This function fast converges to 1 when J increases. Hence we numerically obtain that for $J \geq 3$, $P(3.042) \simeq 0.95$. This implies that for instance that for $J \geq 3$,

- $\mathrm{P}\Big(\widehat{T} \leq 1 + \dfrac{3.042}{\log n}\Big) \simeq 0.95$ when X follows a Gamma distribution
- $\mathrm{P}\Big(\widehat{T} \leq 1 + \dfrac{1.521}{\log n}\Big) \simeq 0.95$ when $|X| = |\mathcal{N}(0,1)|$.

We remark that the ratio τ_{n-1}' is the main contributor to the statistic $\widehat{T}$ and it contains almost all the information. For giving equivalent weights to the other ratios τ_k', $k \leq n-1$ and not be trouble by the nuisance parameter C_2, it is necessary to modify the test statistic. Then we consider:

$$\widetilde{T}_n = \max_{j=n-J,\cdots,n-1}\big\{(n-j)\,\tau_j'\big\} \times \frac{1}{\overline{s}_J} \qquad \text{where} \qquad \overline{s}_J = \frac{1}{J}\sum_{j=n-J}^{n-1}(n-j)\tau_j'. \tag{2.9}$$

The following proposition can be established:

Proposition 3. *Assume that* $G^{-1} \in A_1'$. *Then, for a sequence* $(J_n)_n$ *satisfying* $J_n \underset{n\to\infty}{\longrightarrow} \infty$ *and* $J_n/\log n \underset{n\to\infty}{\longrightarrow} 0$,

$$\Pr\left(\widetilde{T}_n \leq x\right) \underset{n\to\infty}{\sim} \left(1 - e^{-x}\right)^{J_n}. \tag{2.10}$$

In the case where $G^{-1} \in A_2$, similar results can be also established.

EXEMPLE 2.3. *Here there are some examples of classical distributions such as* $G^{-1} \in A_2$*:*

- **Pareto distribution** $\mathcal{P}(\alpha)$**:** *In this case, with* $c > 0$ *and* $C > 0$, $G^{-1}(x) = C\,x^{-c}$ *for* $x \to 0$, *and this implies* $G^{-1} \in A_2$ *with* $a = c$.

- **Burr distributions** $\mathcal{B}(\alpha)$**:** *In this case,* $G(x) = (1 + x^c)^{-k}$ *for* c *and* k *positive real numbers. Thus* $G^{-1}(x) = (x^{-1/k} - 1)^{1/c}$ *for* $x \in [0, 1]$, *implying* $G^{-1} \in A_2$ *with* $a = (ck)^{-1}$.

- **Absolute value of Student distribution** $|t(\nu)|$ **with** ν **degrees of freedom:** *In the case of a Student distribution with* ν *degrees of freedom, the cumulative distribution function is* $F_{t(\nu)}(x) = \frac{1}{2}(1 + I(y, \nu/2, 1/2))$ *with* $y = \nu(\nu + x^2)^{-1}$ *and therefore* $G_{|t(\nu)|}(x) = I(y, \nu/2, 1/2)$, *where* I *is the normalized beta incomplete function. Using the handbook of Abramowitz and Stegun (1964), we have the following expansion* $G_{|t(\nu)|}(x) = \frac{2\nu^{\nu/2-1}}{B(\nu/2,1/2)}\,x^{-\nu} + O(x^{-\nu+1})$ *for* $x \to 0$, *where* B *is the usual Beta function. Therefore,*

$$G^{-1}_{|t(\nu)|}(x) = \frac{B(\nu/2, 1/2)}{2\nu^{\nu/2-1}}\,x^{-1/\nu} + O(x^{-1/\nu-1}) \qquad x \to \infty.$$

Consequently $G^{-1}_{|t(\nu)|} \in A_2$ *with* $a = 1/\nu$.

Remark 2. *The case of standardized log-normal distribution is singular. Indeed, the probability distribution of* X *is the same than the one of* $\exp(Z)$ *where* $Z \sim \mathcal{N}(0, 1)$. *Therefore,* $G(x) = \frac{1}{2}\,erfc\big(\frac{\log x}{\sqrt{2}}\big)$ *implying* $G^{-1}(x) = \exp\left(\sqrt{2}\,erfc^{-1}(2x)\right)$. *Using the previous expansion (2.6), we obtain for any* $\alpha > 0$*:*

$$\begin{aligned} G^{-1}(\alpha\,x) &= \exp\left(\sqrt{2}\,erfc^{-1}(2x\,\alpha)\right) \\ &= \exp\left(\sqrt{2}\,erfc^{-1}(\,2x)\big(1 + \frac{\log\alpha}{2\log x} + O(|(\ln x)^{-2}|)\big)\right) \\ &= G^{-1}(x)\big(1 + O(|(\ln x)^{-1/2}|). \end{aligned}$$

Therefore, the standardized log-normal distribution is such that $G^{-1} \notin A_1 \cup A_2$.

For probability distributions such as $G^{-1} \in A_2$ we obtain the following classical result (see also Embrechts *et al.*, 1997):

Proposition 4. *Assume that $G^{-1} \in A_2$. Then,*

$$\mathrm{P}\Big(\max_{j=n-J,\cdots,n-1} \{\log(\tau_j)\} \leq x \Big) \underset{n\to\infty}{\longrightarrow} \prod_{j=1}^{J} \big(1 - e^{-jx/a}\big). \tag{2.11}$$

Finally, it is possible to consider an outlier detector with asymptotic distribution satisfied as well when G^{-1} belongs in A_1' and A_2. Hence, define:

$$\widehat{D}_{J_n} = \frac{\log 2}{\widehat{L}_{J_n}} \max_{j=1,\cdots,J_n} j \log(\widehat{\tau}_{n-j}) \qquad \text{where} \qquad \widehat{L}_{J_n} = \text{median}\Big\{ \big(j \log(\widehat{\tau}_{n-j}) \big)_{1\leq j\leq J_n} \Big\}. \tag{2.12}$$

Then, we obtain the following theorem:

Theorem 2.1. *Assume that $G^{-1} \in A_1' \cup A_2$. Then, for a sequence $(J_n)_n$ satisfying $J_n \underset{n\to\infty}{\longrightarrow} \infty$ and $J_n/\log n \underset{n\to\infty}{\longrightarrow} 0$,*

$$\Pr\big(\widehat{D}_{J_n} \leq x\big) \underset{n\to\infty}{\sim} \big(1 - e^{-x}\big)^{J_n}. \tag{2.13}$$

Remark 3. *In the definition of $\widehat{D}_{J_n}$ we prefer an estimation of the parameter of the exponential distribution with a robust estimator (median) instead of the usual efficient estimator (empirical mean) since several outliers could corrupt this estimation.*

The main advantage of Theorem 2.1 is the possibility to apply it as well for distributions such as G^{-1} belongs to A_1' and A_2, *i.e.* as well for Gaussian, Gamma or Pareto distributions. Hence, for detecting outliers, for a type I error $\alpha \in (0,1)$, a $1-\alpha$ threshold of the detector $\widehat{D}_{J_n}$ is computed as follows, and with $t = -\log\big(1-(1-\alpha)^{1/J_n}\big)$,

- If $\widehat{D}_{J_n} \leq t$ then we consider that there is no outlier in the sample.
- If $\widehat{D}_{J_n} > t$ then the largest index $\widehat{k}_0$ such as $\widehat{k}_0 \log(\tau_{n-\widehat{k}_0})/\widehat{L}_{J_n} \geq t$ induces that $(X_{(i)})_{n-\widehat{k}_0+1\leq i\leq n}$ are considered to be outliers $\Longrightarrow$ there are $\widehat{k}_0$ detected outliers.

3 Monte-Carlo experiments

We are going to compare the new outlier detector defined in (2.12) with usual univariate outlier detectors. After giving some practical details of the application of $\widehat{D}_{J_n}$, we present the results of Monte-Carlo experiments under several probability distributions.

Practical procedures of outlier detections

The definition of $\widehat{D}_{J_n}$ is simple and it just practically requires the specification of 2 parameters:

- The type I error α is the risk to detect outliers in the sample while there is no outlier. Hence, a natural choice could be the "canonical" $\alpha = 0.05$. However, the construction and perhaps a drawback of this detector is that a detection induces as well 1 or J_n possible outliers. Hence, we chose to be strict concerning the risk of false detection, *i.e.* we chose $\alpha = 0.01$ which implies that we prefer not to detect "small" outliers and hence we avoid to detect a large number of outliers while there is no outlier.

- The number J_n of considered ratios. In the one hand, it is clear that the smaller J_n, the smaller the detection threshold, therefore more sensible is the detector to the presence of outliers. In the other hand, the larger J_n, the more precise is the estimation of the parameter of asymptotic exponential distribution (the convergence rate of $\widehat{L}_{J_n}$ is $\sqrt{n}$) and larger is the possible number of detected outliers. After numerous numerical simulations not reported here, we chose $J_n = [4 * \log^{3/4}(n)]$ (which is negligible with respect to $\log(n)$), *i.e.* for $n = 100$, $J_n = 12$ and for $n = 1000$, $J_n = 17$.

We have compared the new detector $\widehat{D}_{J_n}$ to 4 usual and famous other univariate outlier detectors computed from the sample $(X_1, \cdots, X_n)$.

1. The Student's detector: an observation from the sample $(X_1, \cdots, X_n)$ will be consider as an outlier when $\mathrm{P}(X_k > \overline{X}_n + s_s \times \overline{\sigma}_n)$ where $\overline{X}_n$ and $\overline{\sigma}_n^2$ are respectively the usual empirical mean and variance computed from $(X_1, \cdots, X_n)$, and s_s is a threshold. This threshold is usually computed from the assumption that $(X_1, \cdots, X_n)$ is a Gaussian sample and therefore $s_s = q_{t(n-1)}\big((1 - \alpha/2)\big)$, where $q_{t(n-1)}(p)$ denotes the quantile of the student distribution with $(n-1)$ freedom degree for a probability p.

2. The Tukey's detector: X_k is considered as an outlier from $(X_1, \cdots, X_n)$ if $|X_k - m| > 3 \times IQ$, where $m = \mathrm{median}(X_1, \cdots, X_n)$ and $IQ = Q3 - Q1$, with Q_3 and Q_1 the third and first empirical quartiles of $(X_1, \cdots, X_n)$. Note that the coefficient 3 is obtained from the Gaussian case.

3. The MAD_e detector: X_k is considered as an outlier from $(X_1, \cdots, X_n)$ if when $|X_k - m| > 3 * 1.483 * \mathrm{median}(|X_1 - m|, \cdots, |X_n - m|)$. Once again the coefficient $3 * 1.483$

is obtained from the Gaussian case.

4. The Local Outlier Factor (LOF), which is a non-parametric detector (see for instance Breunig *et al.*, 2000). This procedure is based on this principle: an outlier can be distinguished when its normalized density (see its definition in Breunig *et al.*) is larger than 1 or than a threshold larger than 1. However, the computation of this density requires to fix a parameter k and a procedure or a theory for choosing a priori k does not still exist. Moreover, there does not exist a theory allowing its computation and the computation of the threshold. After numerous simulations not reported here, we tried to optimize the choices of k and the threshold. This leads to fix $k = J_n$, where J_n is used for the computation of $\widehat{D}_J$, and an observation X_i is considered to be an outlier when $LOF(X_i) > 8$.

The three first detectors, that are Student, Tukey and MAD_e detectors are parametric detectors based on Gaussian computations. We will not be surprized if they do not well detect outliers when the distribution of X is "far" from the Gaussian distribution (but these usual detections of outliers, for instance the Student detection realized on studentized residuals from a least squares regression, are realized even if the Gaussian distribution is not attested). Moreover, the computations of these detectors' thresholds are based on an individual detection of outlier, *i.e.* a test deciding if a fixed observation X_{i_0} is an outlier or not. Hence, if we apply them to each observation of the sample, the probability to detect an outlier increases with n. This is not exactly the same test than to decide if there are or not outliers in a sample. Then, to compare these detectors to $\widehat{D}_{J_n}$, it is appropriated to change the thresholds of these detectors as follows: if assumption H_0 is "no outlier in the sample" and H_1 is "there is at least one outlier in the sample", the threshold $s > 0$ is defined from the relation $\mathrm{P}(\exists k = 1, \cdots, n,\ X_k > s) = \alpha$, and therefore, from the independence property $\mathrm{P}(X_k < s) = (1 - (1 - \alpha)^{1/n})$. Then, we define:

1. The Student detector 2: we consider that X_k from $(X_1, \cdots, X_n)$ is an outlier when $X_k > \overline{X}_n + s_s \times \overline{\sigma}_n$ avec $s_s = q_{t(n-1)}\big((1 - \alpha/2)^{1/n}\big)$.

2. The Tukey detector 2: we consider that X_k from $(X_1, \cdots, X_n)$ is an outlier when $X_k - m > s_T \times IQ$. For computing s_T and since the random variables X_j are positive variables, we prefer to consider as a reference the exponential distribution for computing the threshold s_T, which implies $s_T = -\log(4 * (1 - (1 - \alpha)^{1/n}))/\log(3)$.

3. The MAD_e detector 2: we consider that X_k from $(X_1, \cdots, X_n)$ is an outlier when $X_k - m > s_M \times \text{median}(|X_1 - m|, \cdots, |X_n - m|)$. Using an exponential distribution similarly as in the case of Tukey detector 2, after computations we show that $s_M = \log\left(2(1-(1-\alpha)^{1/n})\right)/\log(2/(1+\sqrt{5}))$.

Results of Monte-Carlo experiments

We apply the differents detectors in different frames and for several probability distributions which are:

- The absolute value of Gaussian distribution with expectation 0 and variance 1, denoted $\left|\mathcal{N}(0,1)\right|$ (*case* A_1');
- The exponential distribution with parameter 1, denoted $\mathcal{E}(1)$ (*case* A_1');
- The Gamma distribution with parameter 3, denoted $\Gamma(3)$ (*case* A_1');
- The Weibull distribution with parameters $(3,4)$, denoted $W(3,4)$ (*case* A_1');
- The standard log-normal distribution, denoted $\log -\mathcal{N}(0,1)$ (*not case* A_1' *or* A_2);
- The absolute value of a Student distribution with 2 freedom degrees, denoted $|t(2)|$ (*case* A_2);
- The absolute value of a Cauchy distribution, denoted $|\mathcal{C}|$ (*case* A_2).

In the sequel, we will consider samples $(X_1, \cdots, X_n)$ following these probability distributions, for $n = 100$ and $n = 1000$, and for several numbers of outliers.

Samples without outlier

We begin by generating independent replications of samples without outlier and applying the outlier detectors. The results are reported in Table 1.

Samples with outliers

Now, we consider the case where there is a few number of outliers in the samples $(X_1, \cdots, X_n)$. Denote K the number of outliers, and $\ell > 0$ a real number which represents a shift parameter. We generated $(X_1 + \ell, \cdots, X_K + \ell, X_{K+1}, \cdots, X_n)$ instead of $(X_1, \cdots, X_n)$. We only

Table 1: Frequencies of outlier detection of the different outlier detectors, for the different probability distributions, $n = 100$ and $n = 1000$, while there is no generated outlier in samples.

$n = 100$	$\lvert\mathcal{N}(0,1)\rvert$	$\mathcal{E}(1)$	$\Gamma(3)$	$W(3,4)$	$\log -\mathcal{N}(0,1)$	$\lvert t(2)\rvert$	Cauchy
Prob. $\widetilde{D}_{J_n}$	0.009	0.009	0.011	0.010	0.012	0.011	0.019
Prob. LOF	0.001	0.029	0.013	0	0.643	0.259	0.970
Prob. student	0.637	0.957	0.770	0.117	0.998	0.997	1
Prob. Tukey	0.057	0.625	0.209	0.001	0.972	0.965	1
Prob. MAD_e	0.752	0.995	0.878	0.164	0.998	1	1
Prob. student 2	0.007	0.585	0.254	0.002	0.865	0.911	0.999
Prob. Tukey 2	0	0.019	0	0	0.612	0.472	0.984
Prob. MAD_e 2	0	0.019	0	0	0.614	0.522	0.992
$n = 1000$	$\lvert\mathcal{N}(0,1)\rvert$	$\mathcal{E}(1)$	$\Gamma(3)$	$W(3,4)$	$\log -\mathcal{N}(0,1)$	$\lvert t(2)\rvert$	Cauchy
Prob. $\widetilde{D}_{J_n}$	0.009	0.009	0.009	0.010	0.015	0.011	0.016
Prob. LOF	0.005	0.023	0.019	0.001	0.843	0.281	0.998
Prob. student	1	1	1	0.785	1	1	1
Prob. Tukey	0.255	1	0.839	0	1	1	1
Prob. MAD_e	1	1	1	0.656	1	1	1
Prob. student 2	0.009	0.996	0.826	1	1	1	1
Prob. Tukey 2	0	0.010	0	0	0.995	0.962	1
Prob. MAD_e 2	0	0.010	0	0	0.997	0.978	1

considered the second versions of Student, Tukey et MAD_e detectors, because the original versions of these detectors are not adapted to our framework. Moreover, we computed the mean of detected outliers by each detector. The results are reported in Table 2 and 3.

Conclusions of simulations

It appears that log-ratio detector $\widehat{D}_{J_n}$ provide the best results for not detecting outlier when there is no outlier in samples. Clearly, Student, Tukey or MAD_e detectors are parametric estimators associated to a probability distribution P_0 and therefore could be not at all appropriated for detecting outliers in samples generated with probability distributions "far" from P_0. The LOF detector provides reasonable results except for log-normal, Student or Cauchy distributions.

When outliers are added to samples, we could be a little disappointed in certain cases from the results obtained by the log-ratio detector $\widehat{D}_{J_n}$, notably with respect to the Student detector. Results of classical parametric detectors are accurate for distributions in A_1', and if $\widehat{D}_{J_n}$ provides reasonable results, there are not as convincing. But for log-normal, Student or Cauchy ditributions, these classical detectors often consider as outliers observations which could as well be considered not as outlier. For instance, let be the absolute values of Cauchy r.v., $n = 1000$, $K = 5$ and $\ell = 100$. Figure 1 exhibits the boxplot graph of these r.v. All the detectors accept the presence of outliers except the log-ratio detector $\widehat{D}_{J_n}$, while there are 9 variables with absolute values larger than 100. It could as well be legitimate to conclude that there is no outlier because there are "regular" observations which are larger than outliers.

4 Application to real data

We apply the theoretical results to real datasets of detailed data on individual transactions in the used car market. The purpose of the experiment was to detect as many outliers as possible. The original dataset contains information about $n = 6079$ transactions on the car *Peugeot 207 1.4 HDI 70 Trendy Berline* including year and month which is the date of

Table 2: Frequencies of outlier detection of the different outlier detectors, for the different probability distributions, $n = 100$ and $n = 1000$, while there are $K = 5$ generated outliers with a shift $\ell = 10$ in each replication of sample.

$n = 100$	$\lvert\mathcal{N}(0,1)\rvert$	$\mathcal{E}(1)$	$\Gamma(3)$	$W(3,4)$	$\log -\mathcal{N}(0,1)$	$\lvert t(2)\rvert$	Cauchy
Prob. $\widetilde{D}_{J_n}$	0.955	0.304	0.094	1	0.078	0.082	0.026
Nb. moy. outliers	5.07	5.54	6.39	5.07	9.07	9.08	11.46
Prob. LOF	0.964	0.296	0.034	1	0.529	0.070	0.934
Nb. moy. outliers	4.67	3.21	1.49	5	1.81	1.21	3.06
Prob. student 2	1	0.990	0.735	1	0.707	0.754	0.980
Nb. moy. outliers	2.81	2.47	1.28	4.23	1.18	1.26	1.45
Prob. Tukey 2	0.999	0.840	0.024	1	0.726	0.578	0.967
Nb. moy. outliers	4.806	3.82	1.07	4.97	2.44	2.04	3.33
Prob. MAD_e 2	0.991	0.885	0.008	0.981	0.788	0.732	0.990
Nb. moy. outliers	4.48	4.02	1.01	4.54	2.65	2.61	4.40
$n = 1000$	$\lvert\mathcal{N}(0,1)\rvert$	$\mathcal{E}(1)$	$\Gamma(3)$	$W(3,4)$	$\log -\mathcal{N}(0,1)$	$\lvert t(2)\rvert$	Cauchy
Prob. $\widetilde{D}_{J_n}$	1	0.307	0.041	1	0.015	0.015	0.023
Nb. moy. outliers	5.12	5.88	9.35	5.16	15.96	18.04	11.65
Prob. LOF	1	0.212	0.026	1	0.762	0.799	1
Nb. moy. outliers	5.00	1.95	1.16	5.00	1.93	2.05	16.75
Prob. student 2	1	1	1	1	1	1	1
Nb. moy. outliers	5.00	6.47	5.20	4.23	6.66	9.35	4.54
Prob. Tukey 2	1	0.666	0.001	1	0.997	0.965	1
Nb. moy. outliers	4.267	1.67	1	4.72	5.93	3.44	30.05
Prob. MAD_e 2	0.979	0.678	0	0.981	0.997	0.986	1
Nb. moy. outliers	3.09	1.69	1	2.15	6.03	4.10	36.18

Table 3: Frequencies of outlier detection of the different outlier detectors, for the different probability distributions, $n = 100$ and $n = 1000$, while there are $K = 5$ generated outliers with a shift $\ell = 100$ in each replication of sample.

$n = 100$	$\lvert\mathcal{N}(0,1)\rvert$	$\mathcal{E}(1)$	$\Gamma(3)$	$W(3,4)$	$\log -\mathcal{N}(0,1)$	$\lvert t(2)\rvert$	Cauchy
Prob. $\widetilde{D}_{J_n}$	1	1	1	1	0.904	0.936	0.250
Nb. moy. outliers	5.12	5.13	5.17	5.16	5.50	5.33	7.61
Prob. LOF	1	1	1	1	1	1	0.999
Nb. moy. outliers	5.01	5.03	5.01	5	6.03	5.33	8.58
Prob. student 2	1	1	1	1	1	1	0.971
Nb. moy. outliers	5	5	5	5	4.94	5	2.81
Prob. Tukey 2	1	1	1	1	1	1	1
Nb. moy. outliers	5	5.01	5	5	5.68	5.40	7.95
Prob. MAD_e 2	1	1	1	1	1	1	1
Nb. moy. outliers	5	5.01	5	5	5.75	5.54	8.84
$n = 1000$	$\lvert\mathcal{N}(0,1)\rvert$	$\mathcal{E}(1)$	$\Gamma(3)$	$W(3,4)$	$\log -\mathcal{N}(0,1)$	$\lvert t(2)\rvert$	Cauchy
Prob. $\widetilde{D}_{J_n}$	1	1	1	1	0.691	0.939	0.054
Nb. moy. outliers	5.33	5.25	5.35	5.29	6.20	5.48	15.79
Prob. LOF	1	1	1	1	1	1	0.979
Nb. moy. outliers	5.01	6.38	6.25	5	13.69	7.82	4.79
Prob. student 2	1	1	1	1	1	1	1
Nb. moy. outliers	5	5	5	5	5.79	5.19	34.88
Prob. Tukey 2	1	1	1	1	1	1	1
Nb. moy. outliers	5	5.01	5	5	10.64	8.05	40.97
Prob. MAD_e 2	1	1	1	1	1	1	1
Nb. moy. outliers	5	5.01	5	5	10.74	8.72	36.18

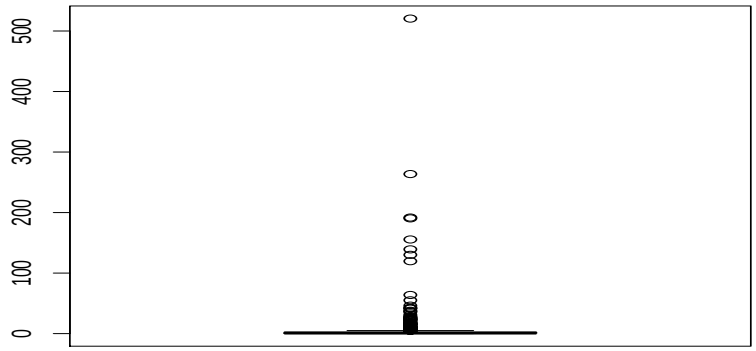

Figure 1: *Sample of* 1000 *Cauchy i.i.d.r.v., where* $K = 5$ *observations have been shifted of* $\ell = 100$.

"car birth", the price, and the number of kilometres driven. We choose these cars because they were advertised often enough to permit us to create a relatively homogeneous sample. Figure 2 depicts the relationship between the price and some variables: Price with Mileage, Price with Age. Such data were collected by Autobiz society, and be used for forecasting the price of a car following its age and mileage. Hence it is crucial to construct a model for the price from a reliable data set including the smallest number of outliers.

We now apply our test procedure to identify eventual outlying observations or atypical combination between variables. After preliminary studies, we chose two significant characteristics for each car of the sample. The first one is the number of kilometres per month. The second one is the residual obtained, after an application of the exponential function, from a linear quantile regression between the logarithm of the price as the dependent variable and the age of the car (in months) and the number of driven kilometres as exogenous variables (an alternative procedure for detecting outliers in robust regression has been developed in Gnanadesikan and Kettenring, 1972). The assumption of independence is plausible for both these variables the residuals. Figure 3 exhibits the boxplots of the distributions of those two variables.

The outlier test $\widehat{D}_{J_n}$ is carried out on those two variables with $J_n = 20$ (given by the

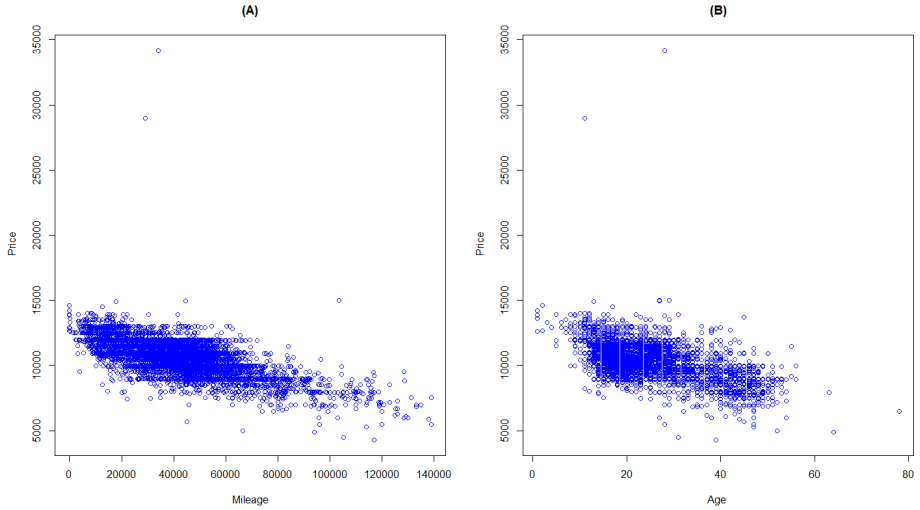

Figure 2: *Relationship between the dependant variables and the regressors: Price with Mileage (left), Price with Age (right).*

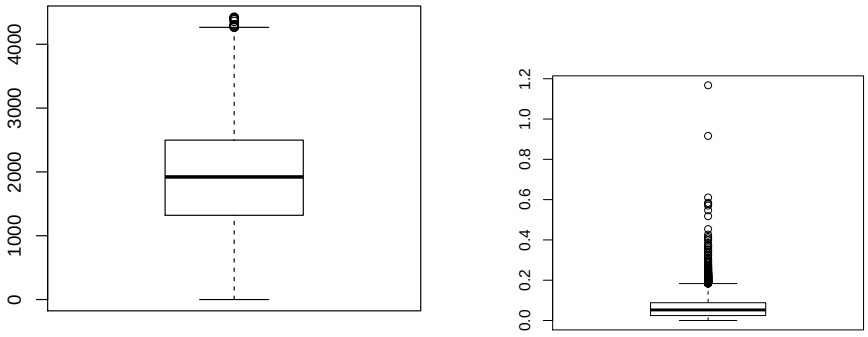

Figure 3: *Boxplots of kilometres per month (left) and of absolute values of linear quantile regression residuals (right).*

Table 4: The outlier test $\widehat{D}_{J_n}$ applied to 3 samples: the number of kilometres per month (km/m), $\max(km/m) - km/m$ and the residuals obtained from a quantile regression of the log-prices onto the age and the mileage.

Sample	J_n	$\widehat{D}_{J_n}$	t	Outliers
km/m (Sup)	20	6.7232	5.96721	$n = 6$
km/m (Inf)	20	5.1200	5.96721	$n = 0$
Res	20	6.3322	5.96721	$n = 2$

empirical choice obtained in Section 3 with $n = 6079$). As the sample size is large, we can accept to eliminate data detected as outliers while there are not really outliers and we chose $\alpha = 0.05$. The results are presented in Tables 4, 5 and 6. Note that, concerning the study of kilometres per month (km/m), we directly applied the test to this variable for detecting eventual "too" large values, but also to $\max(km/m) - (km/m)$ for detecting eventual "too" small values.

Conclusions of the application

We first remark that we did not get the same outliers from the different analysis. It could be expected because the test on residuals worked as a multivariate test and identify atypical association between the three variables Age, Mileage and Price while the tests done on kilometres/minute identifies outlying values in a bivariate case *i.e.* a typical association between the two variables Age and Mileage. From a practitioner's point of view it may be advisable to apply the test for the two cases together one by one to be sure to detect the largest number of outliers. A second remark concerns the "type" of the detected outliers. We can state that concerning kilometres/minute, outliers are simply the largest values (the test did not identify outliers for "too" small values). But for the regression residuals, the detected outliers probably correspond on transcription errors on the prices. Thus, two kinds of outliers have been detected.

Table 5: Detailed analysis of the detected outliers obtained from the sample of kilometers per month (large values).

Detected Outliers	Price	Mileage	Age	Kilometers per Month	Predicted Price
outlier(1)	9590	70249	16	4391	9909
outlier(2)	11690	61484	14	4392	10286
outlier(3)	10490	61655	14	4404	10280
outlier(4)	9390	61891	14	4421	10272
outlier(5)	11500	39826	9	4425	11285
outlier(6)	11900	65411	15	4361	10111

Table 6: Detailed analysis of outliers detected from the residual's sample.

Detected Outliers	Price	Mileage	Age	Predicted Price
Outlier(2)	34158	34158	28	10626
Outlier(3)	29000	29000	11	11600

5 Proofs

Proof of Proposition 1. We begin by using the classical following result (see for example Embrechts *et al.* 1997):

$$\Big(X_{(n-J)}, X_{(n-J+1)}, \cdots, X_{(n)}\Big) \stackrel{d}{=} \Big(G^{-1}\big(\Gamma_{J+1}/\Gamma_{n+1}\big), G^{-1}\big(\Gamma_J/\Gamma_{n+1}\big), \cdots, G^{-1}\big(\Gamma_1/\Gamma_{n+1}\big)\Big), \quad (5.1)$$

where $(\Gamma_i)_{i\in\mathbb{N}^*}$ is a sequence of random variables such as $\Gamma_i = E_1 + \cdots + E_i$ for $i \in \mathbb{N}^*$ and $(E_i)_{j\in\mathbb{N}^*}$ is a sequence of i.i.d.r.v. with distribution $\mathcal{E}(1)$. Consequently, we have

$$\Big(\tau_{(n-J)}, \tau_{(n-J+1)}, \cdots, \tau_{(n-1)}\Big) \stackrel{d}{=} \Big(\frac{G^{-1}\big(\Gamma_J/\Gamma_{n+1}\big)}{G^{-1}\big(\Gamma_{J+1}/\Gamma_{n+1}\big)}, \frac{G^{-1}\big(\Gamma_{J-1}/\Gamma_{n+1}\big)}{G^{-1}\big(\Gamma_J/\Gamma_{n+1}\big)}, \cdots, \frac{G^{-1}\big(\Gamma_1/\Gamma_{n+1}\big)}{G^{-1}\big(\Gamma_2/\Gamma_{n+1}\big)})\Big).$$

But for $j \in \mathbb{N}^*$, $G^{-1}\big(\Gamma_j/\Gamma_{n+1}\big) = G^{-1}\Big(\dfrac{1}{\Gamma_{n+1}} \times \Gamma_j\Big)$. From the strong law of large numbers, $\Gamma_{n+1} \xrightarrow[n\to\infty]{a.s.} \infty$, therefore since $G^{-1} \in A_1$, we almost surely obtain:

$$G^{-1}\big(\Gamma_j/\Gamma_{n+1}\big) = f_1\big(\frac{1}{\Gamma_{n+1}}\big) \times \Big(1 + \frac{f_2(\Gamma_j)}{\log(\Gamma_{n+1})} + O\big(\frac{1}{\log^2(\Gamma_{n+1})}\big)\Big).$$

Using once again the strong law of large numbers, we have $\Gamma_{n+1} \sim n$ almost surely. Hence, we can write for all $j = 1, \cdots, J$,

$$\begin{aligned}\frac{G^{-1}\big(\Gamma_j/\Gamma_{n+1}\big)}{G^{-1}\big(\Gamma_{j+1}/\Gamma_{n+1}\big)} &= \frac{1 + \frac{f_2(\Gamma_j)}{\log(n)} + O\big(\frac{1}{\log^2(n)}\big)}{1 + \frac{f_2(\Gamma_{j+1})}{\log(n)} + O\big(\frac{1}{\log^2(n)}\big)} \\ &= 1 + \frac{f_2(\Gamma_j) - f_2(\Gamma_{j+1})}{\log(n)} + O\big(\frac{1}{\log^2(n)}\big). \quad (5.2)\end{aligned}$$

By considering now the family $(\tau'_j)_j$ and the limit of the previous expansion, we obtain

$$\big(\tau'_{n-J}, \tau'_{n-J+1}, \cdots, \tau'_{n-1}\big) \xrightarrow[n\to\infty]{\mathcal{D}} \Big(f_2(\Gamma_J) - f_2(\Gamma_{J+1})\,,\, f_2(\Gamma_{J-1}) - f_2(\Gamma_J)\,, \cdots,\, f_2(\Gamma_1) - f_2(\Gamma_2)\Big).$$

The function $(x_1, \cdots, x_J) \mapsto \max(x_1, \cdots, x_J)$ is a continuous function on $\mathbb{R}^J$, therefore we obtain (2.5). □

Proof of Proposition 2. We use the asymptotic relation (2.5). Since $G^{-1} \in A'_1$, for $k = 1, \cdots, J-1$,

$$f_2(\Gamma_k) - f_2(\Gamma_{k+1}) = -C_2 \log\big(\Gamma_k/\Gamma_{k+1}\big) = -C_2 \log\big(\Gamma_k/\Gamma_{J+1}\big) + C_2 \log\big(\Gamma_{k+1}/\Gamma_{J+1}\big),$$

$f_2(\Gamma_k) - f_2(\Gamma_k + E_{k+1})$ is absolutely continuous with respect to Lebesgue measure since Γ_k and E_{k+1} are independent random variables. For $k = J$, $f_2(\Gamma_J) - f_2(\Gamma_{J+1}) = -C_2 \log\big(\Gamma_J/\Gamma_{J+1}\big)$. Using once again the property (5.1), and since for an exponential distribution $\mathcal{E}(1)$, $G^{-1}(x) = -\log(x)$, then

$$\Big(f_2(\Gamma_J) - f_2(\Gamma_{J+1}),\, f_2(\Gamma_{J-1}) - f_2(\Gamma_J), \cdots,\, f_2(\Gamma_1) - f_2(\Gamma_2)\Big) \stackrel{d}{=} C_2 \left(E'_{(1)}, E'_{(2)} - E'_{(1)}, \cdots, E'_{(J)} - E'_{(J-1)}\right)$$

where $(E'_j)_j$ is a sequence of i.i.d.r.v. following a $\mathcal{E}(1)$ distribution and $E'_{(1)} \leq E'_{(2)} \leq \cdots \leq E'_{(J)}$ is the order statistic from $(E'_1, \cdots, E'_J)$. This implies with $y = x/C_2$

$$\begin{aligned} \mathrm{P}\big(\max_{j=n-J,\cdots,n-1}\{\tau'_j\} \leq x\big) &\underset{n\to\infty}{\longrightarrow} \mathrm{P}\big(E'_{(1)} \leq y,\, E'_{(2)} \leq y + E'_{(1)}, \cdots, E'_{(J)} \leq y + E'_{(J-1)}\big) \\ &\underset{n\to\infty}{\longrightarrow} J!\, \mathrm{P}\big(E'_1 \leq y,\, E'_1 \leq E'_2 \leq y + E'_1, \cdots, E'_{J-1} \leq E'_J \leq y + E'_{J-1}\big). \end{aligned}$$

The explicit computation of this probability is possible. Indeed:

$$\begin{aligned} &\mathrm{P}\big(E'_1 \leq y,\, E'_1 \leq E'_2 \leq y + E'_1, \cdots, E'_{J-1} \leq E'_J \leq y + E'_{J-1}\big) \\ &\quad= \int_0^y e^{-e_1} de_1 \int_{e_1}^{y+e_1} e^{-e_2} de_2 \int_{e_2}^{y+e_2} e^{-e_3} de_3 \cdots \int_{e_{J-2}}^{y+e_{J-2}} e^{-e_{J-1}} de_{J-1} \int_{e_{J-1}}^{y+e_{J-1}} e^{-e_J} de_J \\ &\quad= \big(1 - e^{-y}\big) \int_0^y e^{-e_1} de_1 \int_{e_1}^{y+e_1} e^{-e_2} de_2 \int_{e_2}^{y+e_2} e^{-e_3} de_3 \cdots \int_{e_{J-2}}^{y+e_{J-2}} de_{J-1} e^{-2e_{J-1}} \\ &\quad= \frac{1}{2}\big(1 - e^{-y}\big)\big(1 - e^{-2y}\big) \int_0^y e^{-e_1} de_1 \int_{e_1}^{y+e_1} e^{-e_2} de_2 \int_{e_2}^{y+e_2} e^{-e_3} de_3 \cdots \int_{e_{J-3}}^{y+e_{J-3}} de_{J-2} e^{-3e_{J-2}} \\ &\quad= \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ &\quad= \frac{1}{(J-2)!}\big(1 - e^{-y}\big)\big(1 - e^{-2y}\big) \times \cdots \times \big(1 - e^{-(J-2)y}\big) \int_0^y e^{-e_1} de_1 \int_{e_1}^{y+e_1} e^{-(J-1)e_2} de_2 \\ &\quad= \frac{1}{(J-1)!}\big(1 - e^{-y}\big)\big(1 - e^{-2y}\big) \times \cdots \times \big(1 - e^{-(J-1)y}\big) \int_0^y e^{-Je_1} de_1 \\ &\quad= \frac{1}{J!}\big(1 - e^{-y}\big)\big(1 - e^{-2y}\big) \times \cdots \times \big(1 - e^{-Jy}\big). \end{aligned}$$

Then, we obtain (2.7).

□

Proof of Proposition 3. Such a result can be obtained by modifications of Propositions 1 and 2. Indeed, we begin by extending Proposition 1 in the case where $J_n \underset{n\to\infty}{\longrightarrow} \infty$ and

$J_n/\log n \underset{n\to\infty}{\longrightarrow} 0$. This is possible since $\Gamma_{n+1}/n = 1 + n^{-1/2}\varepsilon_n$ with $\varepsilon_n \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}(0,1)$ from usual Central Limit Theorem. Using the Delta-method, we also obtain $\log(\Gamma_{n+1}/n) = n^{-1/2}\varepsilon'_n$ with $\varepsilon'_n \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}(0,1)$. Hence, for any $j = 1, \cdots, J_n$,

$$log(n)\Big(\frac{G^{-1}\big(\Gamma_j/\Gamma_{n+1}\big)}{G^{-1}\big(\Gamma_{j+1}/\Gamma_{n+1}\big)} - 1\Big) \;=\; f_2(\Gamma_j) - f_2(\Gamma_{j+1}) + O\big(\frac{1}{\log(n)}\big).$$

Denote F_n the cumulative distribution function of $\big(\tau'_{n-J_n}, \cdots, \tau'_{n-1}\big)$, and $\widetilde{F}_n$ the one of $\big(f_2(\Gamma_{J_n}) - f_2(\Gamma_{J_n+1}), \cdots, f_2(\Gamma_1) - f_2(\Gamma_2)\big) = \big(C_2\log(\Gamma_{J_n+1}/\Gamma_{J_n}), \cdots, C_2\log(\Gamma_2/\Gamma_1)\big)$. Then for all $(x_1, \cdots, x_{J_n}) \in (0,\infty)^{J_n}$,

$$F_n(x_1, \cdots, x_{J_n}) = \widetilde{F}_n(x_1 + u_n^1, \cdots, x_{J_n} + u_n^{J_n}),$$

with $u_n^i = O\big(\frac{1}{\log(n)}\big)$. But it is clear that the probability measure of $\big(f_2(\Gamma_{J_n}) - f_2(\Gamma_{J_n+1}), \cdots, f_2(\Gamma_1) - f_2(\Gamma_2)\big)$ is absolutely continuous with respect to the Lebesgue measure on R^{J_n} . Thus, the partial derivatives of the function $\widetilde{F}_n$ exist. Then from the Taylor-Lagrange expansion,

$$\widetilde{F}_n(x_1 + u_n^1, \cdots, x_{J_n} + u_n^{J_n}) = \widetilde{F}_n(x_1, \cdots, x_{J_n}) + \sum_{j=1}^{J_n} u_n^j \times \frac{\partial}{\partial x_j} F_n(x'_1, \cdots, x'_{J_n}),$$

where $(x'_1, \cdots, x'_{J_n}) \in (0,\infty)^{J_n}$. Hence, we obtain $\Big|\sum_{j=1}^{J_n} u_n^j \times \frac{\partial}{\partial x_j} F_n(x'_1, \cdots, x'_{J_n})\Big| \leq C\sum_{j=1}^{J_n} u_n^j \leq C'\frac{J_n}{\log n}$. Consequently, we have:

$$F_n(x_1, \cdots, x_{J_n}) \underset{n\to\infty}{\sim} \widetilde{F}_n(x_1, \cdots, x_{J_n}).$$

Now, we are going back to the proof of Proposition 2 by computing $\widetilde{F}_n(x_1, \cdots, x_{J_n})$. This leads to compute the following integral:

$$\int_0^{y_1} e^{-e_1} de_1 \int_{e_1}^{y_2+e_1} e^{-e_2} de_2 \int_{e_2}^{y_3+e_2} e^{-e_3} de_3 \cdots \int_{e_{J-2}}^{y_{J-1}+e_{J-2}} e^{-e_{J-1}} de_{J-1} \int_{e_{J-1}}^{y_J+e_{J-1}} e^{-e_J} de_J,$$

with $y_i = x_i/C_2$, and with the same iteration than in the proof of Proposition 2, we obtain

$$F_n(x_1, \cdots, x_{J_n}) \underset{n\to\infty}{\overset{\mathcal{L}}{\sim}} \prod_{j=1}^{J_n} \big(1 - e^{-jx_{J_n-j+1}/C_2}\big).$$

Then, by considering the vector $((n-j)\tau'_j)_{n-J_n\leq j\leq n-1}$ and $x \geq 0$, we have

$$\Pr\big(\max_{j=n-J_n,\cdots,n-1}\{(n-j)\tau'_j\} \leq x\big) \underset{n\to\infty}{\sim} \big(1 - e^{-x/C_2}\big)^{J_n}.$$

To achieve the proof, we use the Slutsky Lemma. Indeed, since $\overline{s}_{J_n}$ converges to C_2 in probability, and from the law of large numbers the family $((n-j)\tau'_j)_j$ is asymptotically a family of i.i.d.r.v. with exponential distribution of parameter $1/C_2$ then $\frac{1}{\overline{s}_{J_n}} \max_{j=n-J_n,\cdots,n-1}\{(n-j)\tau'_j\}$ asymptotically has the same distribution than $\max_{j=n-J_n,\cdots,n-1}\{\frac{(n-j)}{C_2}\,\tau'_j\}$, which is the maximum of J_N i.i.d.r.v. with $\mathcal{E}(1)$ distribution. □

Proof of Proposition 4. We begin by considering the proof of Proposition 1. Hence, since $G^{-1} \in A_2$, we obtain for $k = 1, \cdots, J$,

$$\log\Big(\frac{G^{-1}\big(\Gamma_k/\Gamma_{n+1}\big)}{G^{-1}\big(\Gamma_{k+1}/\Gamma_{n+1}\big)}\Big) \;=\; -a\,\log\big(\Gamma_k/\Gamma_{k+1}\big) + o(1).$$

Then, we directly use the result of Proposition 2. □

Proof of Theorem 2.1. First consider the case $G^{-1} \in A'_1$. Using Proposition 1 and a Taylor expansion log function applied to (5.2), then

$$\log\Big(\frac{G^{-1}\big(\Gamma_j/\Gamma_{n+1}\big)}{G^{-1}\big(\Gamma_{j+1}/\Gamma_{n+1}\big)}\Big) = \frac{f_2(\Gamma_j) - f_2(\Gamma_{j+1})}{\log(n)} + O\big(\frac{1}{\log^2(n)}\big).$$

Consequently, using $G^{-1} \in A'_1$ and therefore the definition of f_2, we obtain:

$$\log(\tau_j) = -\frac{C_2}{\log(n)}\Gamma_j/\Gamma_{j+1} + O\big(\frac{1}{\log^2(n)}\big).$$

To prove (2.13), it is sufficient to use again the proof of Proposition 3, to normalize the numerator and denominator with $\log n$ and therefore to consider $\log n \times \widehat{L}_{J_n}$, which converges in probability to $\log 2 (C_2)^{-1}$ (indeed, the median of a sample of iidrv with $\mathcal{E}(\lambda)$ distribution is $\log 2/\lambda$).

When $G^{-1} \in A_2$, we can use the same argument that the ones of the proof of Proposition 3 with C_2 replaced by a (the reminder $1/\log n$ obtained from the definition of A_2 allows to achieve the proof when J_n is negligible compared to $\log n$). □

References

Abramowitz, M. and Stegun, I.A. (1964). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, New York: Dover Publications.

Barnett,V. and Lewis, T. (1994). Outliers in Statistical Data. Wiley Series in Probability & Statistics, Wiley.

Beckman, R.J. and Cook, R.D. (1983). Outlier....s, *Technometrics*, **25**, 119-149.

Beirlant, J., Vynckiera P. and Teugel, J. (1996) Tail Index Estimation, Pareto Quantile Plots Regression Diagnostics, *Journal of the American Statistical Association*, **91**, 1659-1667.

Blair, J.M., Edwards C.A. and Johnson J.H. (1976). Rational Chebyshev approximations for the inverse of the error function, *Math. Comp.* **30**, 827-830.

Breunig, M.M., Kriegel, H.-P., Ng, R.T. and Sander, J. (2000). LOF: Identifying Density-based Local Outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.*

Embrechts, P., Kleppelberg, C. and Mikosch, T. (1997). Modelling Extreme Events for Insurance and Finance. Springer.

Gnanadesikan, R. and Kettenring, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, **28**, 81-124.

Grubbs, F.E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, **11**, 1-21.

Hawkins, D.M. (1980). Identification of Outliers. Chapman and Hall

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**, 1163-1174.

Hubert, M., Dierckx, G. and Vanpaemel, D. (2012). Detecting influential data points for the Hill estimator in Pareto-type distributions. *Computational Statistics and Data Analysis*, **65**, 13-28.

Knorr, E.M., Ng, R.T. and Tucakov, V. (2000). Distance-based outliers: algorithms and applications, *The VLDB Journal*, **8**, 237-253.

Rousseeuw, P.J. and Leroy, A.M. (2005). Robust Regression and Outlier Detection. Wiley Series in Probability and Statistics, Wiley.

Tietjen, G.L. and Moore, R.H. (1972). Some Grubbs-Type Statistics for the Detection of Several Outliers, *Technometrics*, **14**, 583-597.

Tse, Y.K. and Balasooriya, U. (1991). Tests for Multiple Outliers in an Exponential Sample, *The Indian Journal of Statistics, Series B*, **53**, 56-63.

Chapitre 2

Le prix des VO. Modélisation et prédiction.

Nous posons

$$y = g(X) + \varepsilon$$

où pour un VO, y désigne le prix et $X^T = (X_1, \ldots, X_k)$ l'ensemble de caractères invariants dans le temps du véhicule (marque, modèle, type d'énergie, nombre de portes, ...), les caractéristiques mesurables (kilométrage, age, ...), les conditions de vente (lieu de vente, le type de marché ...). ε étant un terme aléatoire qui permet, contrairement à un VN, de prendre en compte les incertitudes dans la formation des prix d'un VO. Nous nous proposons de trouver une forme appropriée de g de telle sorte que la relation ainsi établie puisse satisfaire aussi bien les conditions d'optimalités perçues sous un angle mathématique que les contraintes imposées par la finalité commerciale du travail.
De par leur particularité, les VO de luxes et de prestiges sont exclus de notre étude.

2.1 Problématiques industrielle et académique

2.1.1 Les contraintes

Afin que les résultats soient admissibles du point de vue métier, plusieurs contraintes sur la modélisation ont été définies au préalable par l'intermédiaire des règles déterminées par les experts. En effet, le modèle que nous choisissons doit être le plus réaliste que possible. De plus, ce modèle doit être suffisamment flexible pour s'adapter tous les mois aux nouvelles données téléchargées, et cela, pour n'importe quel segment de véhicules.

Ces contraintes définies par les experts sont souvent incompatibles avec la rigueur mathématique usuelle dans le sens où des violations légères de certains postulats peuvent être admises.

Les contraintes réalistes

(**C1**) *Le prix doit refléter les effets simultanés de l'âge et du kilométrage.*
(**C2**) *L'âge et le kilométrage doivent évoluer respectivement en sens contraire avec le prix.*

Les contraintes commerciales

(**C3**) *Le modèle choisi, doit non seulement satisfaire les contraintes réalistes mais doit fournir des coefficients de prédiction pour un nombre suffisamment élevé de versions de*

véhicules disponibles dans la base Autobiz.

(**C4**) *Dans toutes les démarches qui seront entreprises, nous devons tenir compte de la finalité commerciale du modèle.*

(**C5**) *De plus, la capacité de prédiction du modèle choisi doit être perceptible lors de la mise en production et la validation des résultats doit être simple et facile à interpréter.*

Dans **(C1)**, il serait donc inadmissible de concevoir un modèle où le prix ne dépendrait que de l'une de ces variables.

REMARQUE 2.1.1. *Les contraintes (**C4**)et (**C5**) nous imposent à exclure tout modèle de type "boite noire".*

2.1.2 Formulation mathématique du problème

2.1.2.1 Formulation

Nous disposons de n réalisations $y_1, \ldots, y_n$ de la variable aléatoire y correspondant au prix d'un VO. Nous notons X la variable représentant l'ensemble des variables explicatives et $\mathscr{X}$ l'espace dans lequel elle prend ses valeurs.

Soit la fonction $g \in \mathscr{G}$ telle que

$$g : \mathscr{X} \longrightarrow \mathbb{R}$$

et

$$g(X) = \mathbb{E}[y \mid X = x]$$

Le problème peut donc se traduire de la manière suivante :

$$y = g(X) + \varepsilon \tag{2.1.1}$$

et nous supposons que les variables $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ sont mutuellement indépendantes et identiquement distribuées avec $\mathbb{E}[\varepsilon] = 0$.

La résolution de (2.1.1) consiste en une estimation d'une fonction $\hat{g}$ qui approche l'espérance au sens d'une erreur de prédiction que nous notons $L(\hat{g})$ conditionnellement aux variables explicatives. Cette espérance représente les grandes tendances de l'évolution du prix. L'écart entre le prix réel et cette fonction représente les fluctuations que nous ne pouvons pas expliquer qui englobent les perturbations économiques, le bruit de mesure lié au recueil des données et l'influence des variables dont nous ne disposons pas.

Nous évaluons la performance de la méthode d'estimation de $\hat{g}$ par sa capacité à choisir parmi une collection de modèles celui dont la quantité $L(\hat{g})$ est le plus faible [HTF09].

2.1.2.2 Les formes usuelles de $L(\hat{g})$

La quantité $L(\hat{g})$, qu'on appelle généralement *risque* ou *fonction coût* peut se présenter sous différentes formes dont les plus usuelles sont :

- $L(\hat{g}) = \mathbb{E}\left[(\hat{g}(X) - y)^2\right]$ (Sum of Square Errors (SSE)) : c'est la formulation qui domine la littérature pour ce qui est des problèmes de prédiction. Toutefois, l'optimalité des paramètres obtenus par la minimisation de ce critère repose sur certains postulats qui sont difficilement satisfaits par les données réelles.

- $L(\hat{g}) = \mathbb{E}\left[|y - \hat{g}(X)|\right]$ (Sum of Absolute Error (SAE) [BS71] [NW82]) : La minimisation de ce critère apparait comme une mesure plus satisfaisante du risque et les estimateurs obtenus sont moins sensibles aux outliers puisque le modèle s'ajuste à une partie des données en ignorant les valeurs extrêmes.

- $L(\hat{g}) = \mathbb{E}\left[\left|\frac{y-\hat{g}(X)}{y}\right|\right]$ (Sum of Relative Errors [NW77]) : Ce critère apparait comme le plus approprié pour être un critère de mesure de qualité de prédiction dans des études pratiques sachant que $y > 0$.

Chacun des critères cités précédemment est optimal sous certaines conditions. Une approche plus réaliste consisterait à combiner plusieurs critères [BK96]. La plupart des méthodes préconisées pour cela consiste à minimiser le carré des erreurs et l'erreur absolue.

Par ailleurs, plusieurs critères sont également disponibles pour apprécier la performance du modèle et nous pouvons citer [BLW09] :

- $RMSE = \left(\frac{1}{n}\sum_i^n (y_i - \hat{y}_i)\right)^{1/2}$ (Root Mean Square Errors).
- $MAE = \frac{1}{n}\sum_i^n |y_i - \hat{y}_i|$(Mean Absolute Error).
- $MaxE = \max|y_i - \hat{y}_i|$ (L'erreur maximale).
- $MAPE = \frac{1}{n}\sum_i |\frac{y_i - \hat{y}_i}{y_i}|$ (Mean Absolute Percentage Error(MAPE) [WA85] [Ray07] [TSB99]). Il correspond à la moyenne des écarts en valeur absolue par rapport aux valeurs observées et constitue un indicateur pratique de comparaison de modèles.

2.2 Le critère de prédiction choisi pour notre étude.

L'évaluation d'un modèle ne pourrait être dissociée de l'usage que l'on souhaite en faire.

Problématique 1. *Tenant compte de nos contraintes, il s'agit pour nous de faire un compromis entre deux formes d'évaluation du pouvoir prédictif du modèle et qui sont : « **avoir un écart-moyen acceptable pour toutes les données** » et « **avoir un écart minimum satisfaisant pour une proportion des données** ».*

Ainsi, dans le cas où en espérance, nous n'avons pas atteint le minimum voulu, en terme de fréquence, le modèle pourrait être acceptable.

Minimiser une espérance, bien que du point de vue mathématique, représente une facilité, n'a pas de significativité pratique et exploitable. Le respect d'une inégalité sur l'espérance ne garantit rien sur la fréquence des dépassements de cette inégalité.

Dans un cas concret, les contraintes à satisfaire avec une certaine probabilité ou fréquence sont généralement celles qui ont plus de signification pratique, sont facilement interprétables, mais aussi difficiles à traiter de façon mathématique.

Ainsi, l'idée serait de proposer une contrainte réaliste ayant une signification pratique qui va dans le sens d'un traitement mathématique plus facile comme par exemple la préservation de la convexité.

Reprenons le modèle en (2.1.1) et notons

$$\xi(\hat{g}(X_i), y_i) = \xi_i = \left|\frac{y_i - \hat{g}(X_i)}{y_i}\right| \tag{2.2.1}$$

une réalisation de la variable aléatoire ξ qui caractérise les erreurs relatives de prédiction.

Sous une forme mathématique, l'idée d'avoir un seuil de satisfaction pour un certain nombre de données peut se traduire par :

$$\mathbb{P}\left[\xi_i \leq \alpha\right] \geq p \tag{2.2.2}$$

où

- $p \in [0,1]$ désigne le niveau de probabilité requis pour la satisfaction de la contrainte.
- α désigne le niveau de la contrainte.

Pour Autobiz, on choisira $\alpha = 0.15$.

Notre recherche d'optimalité pour notre modèle se résume ainsi à résoudre le système suivant :

$$(S) : \begin{cases} \min_{g \in \mathscr{G}} \mathbb{E}\left[(g(X) - y)^2\right] \\ \\ \text{s. c } \mathbb{P}\left[\xi_i \leq \alpha\right] \geq p \end{cases} \tag{2.2.3}$$

2.3 Description des données utilisées dans la démarche expérimentale

La description s'attache premièrement à illustrer et en second lieu à découvrir des formes de régularité sur la corrélation et la dépendance entre les différentes variables ainsi que l'identification de certaines homogénéités.

2.3.1 Les données

Les données sur lesquelles se fondent les études menées dans ce travail correspondent à des annonces dédoublonnées du mois de *novembre* 2011 sur lesquelles nous avons appliqué les traitements dont les étapes sont décrites dans l'introduction.

Certaines variables ont été recodées pour faciliter leur interprétation et leur prise en compte dans le modèle. La variable "nombre de portes" qui initialement prend les valeurs "3 portes" ou "5 portes" a été recodée telle que

$$pt_5 = \begin{cases} 1 & \text{si 5 portes} \\ 0 & \text{sinon} \end{cases}$$

$$pt_3 = \begin{cases} 1 & \text{si 3 portes} \\ 0 & \text{sinon} \end{cases}$$

La considération de deux variables indicatrices du nombre de portes se justifie par le fait que dans notre échantillon, cette information n'est pas toujours explicite.

La présence des options jugées *a priori* pertinentes dans la formation du prix *(la climatisation automatique(CLIMA), Electronic Stability Program (ESP), Filtre à Particules(FAP), la jante en alliage d'aluminium JANTALU), Système de navigation embarquée(NAVI) , Toit ouvrant panoramique (PANORAMA), la sellerie en cuir (CUIR), le phare au Xenon(XENON), le siège sport adaptatif (SPORT))* ont également fait l'objet d'un recodage telle que pour chacune des k options (opt_k) on ait

$$opt_k = \begin{cases} 1 & \text{si option présente} \\ 0 & \text{si absente ou de série} \end{cases}$$

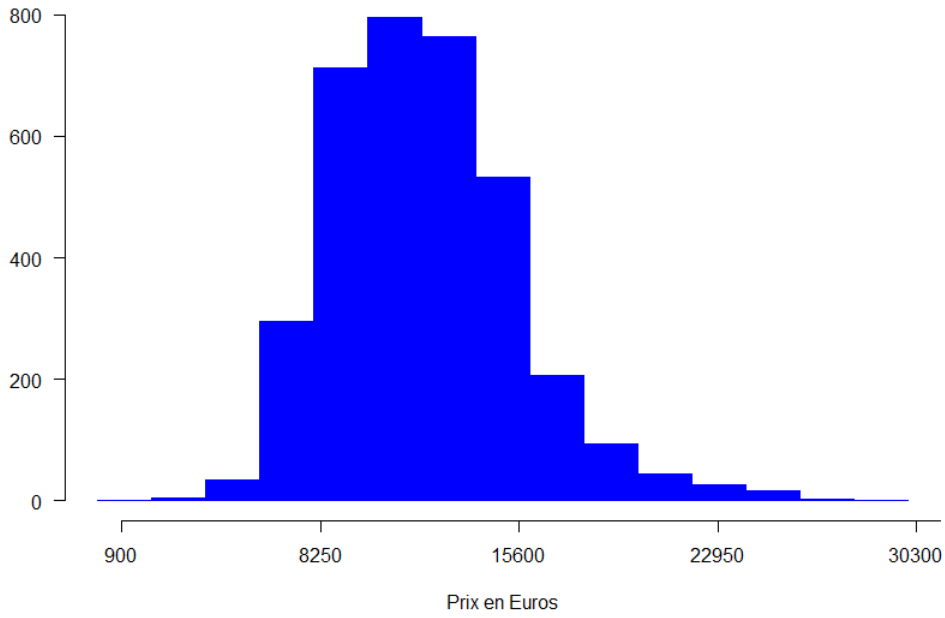

FIGURE 2.1 – DISTRIBUTION DU PRIX. *Il s'agit de la distribution des prix observés pour la marque Peugeot sur un échantillon de taille $N = 174933$ et composé de véhicules dont la PMEC est de 2008. Nous constatons que la dispersion autour de la moyenne est assez grande. Cela manifeste une hétérogénéité entre les prix d'une même marque de véhicule et de même tranche d'age. D'autres variables peuvent être la cause de cette hétérogénéité qui peut être atténuée par un raffinement de l'échelle d'observation.*

Le fait que la teinture du véhicule soit métallisée (mt) ou non a été recodé :

$$mt = \begin{cases} 1 & \text{si teinture métallisée} \\ 0 & \text{sinon} \end{cases}$$

2.3.2 Définition d'une maille pour la modélisation

La définition d'une maille pour la modélisation consiste en une démarche dans laquelle tous les facteurs exogènes pouvant influencer le prix sont contrôlés. Cette étape déterminera à quelle niveau d'agrégation des données la variabilité du prix sera étudiée. La méthode usuelle répondant à cet objectif est l'analyse de la covariance [AB12]. Suite à cette analyse statistique et sur les avis favorables des experts, nous proposons d'effectuer notre analyse à un maillage correspondant à une arborescence **Marque - Modèle - Energie - Carrosserie - Motorisation -Finition - Cylindrée** *(ex : Peugeot 207 1.5 HDI Berline)* que nous pouvons voir en FIG.2.2 puisque cette échelle d'analyse offre un haut degré de différenciation des prix VO et la prise en compte de l'hétérogénéité des prix.

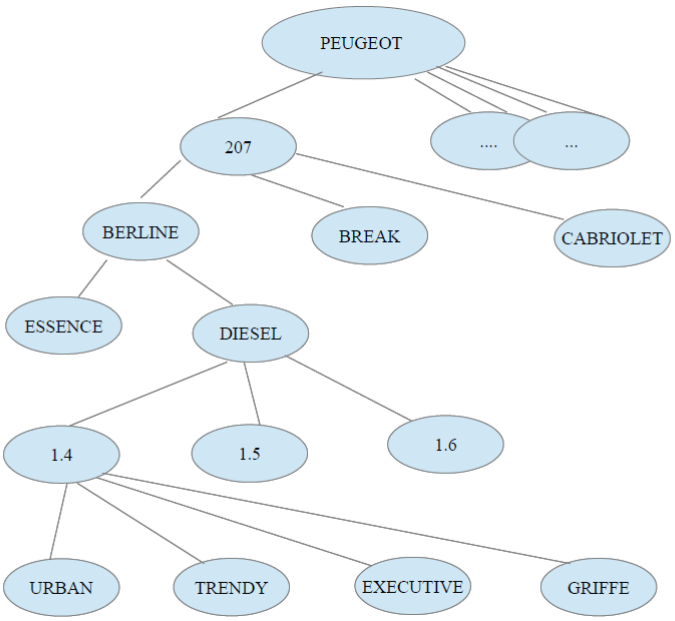

FIGURE 2.2 – STRUCTURE D'UNE MAILLE. *Nous soulignons que plusieurs générations ou phases d'un même modèle de véhicule peuvent se retrouver dans une maille maille.*

.

2.4 Première tentative : un modèle de régression linéaire

L'approche classique pour la construction d'un modèle statistique consiste à chercher le modèle le plus parcimonieux. Dans ce sens, minimiser le nombre de variables dans un modèle permet une plus grande stabilité du modèle et une dépendance moins grande du modèle par rapport aux données utilisées lors de l'apprentissage.

Ici, nous essayons de modéliser un phénomène pour lequel nous avons une connaissance approximative *a priori*. Nous souhaitons intégrer cette connaissance *a priori* sans cependant devoir faire plus d'hypothèses sur la configuration des mesures influentes. Le modèle paramétré doit être en cohérence avec des connaissances *a priori* qui fera que le choix du modèle optimal ne se base donc pas sur un souci de parcimonie mais sur l'expérience et le bon sens.

Dans cette optique, nous restreignons l'ensemble des variables explicatives au kilométrage (X_1) et l'âge (X_2). Les autres variables apparaitront comme des ajustements lors de la restitution des résultats.

2.4.1 Le modèle de base

Pour construire notre modèle, notre première approche est de poser des hypothèses simplificatrices sur la nature de g et de ε. Nous supposons que $\mathscr{G}$ se restreint à l'espace des fonctions candidates aux combinaisons linéaires dans $\mathscr{X}$ et le modèle associe donc un effet à chaque variable. De ce fait, une prédiction à partir d'une réalisation y correspond à la somme des effets de (X_1, X_2).

Suite à ces simplifications, l'équation (2.1.1) devient

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{2.4.1}$$

ou tout simplement, sous une écriture matricielle

$$y = X^T\beta + \varepsilon, \text{ où } X^T = (X_1, X_2) \text{ et } \beta = (\beta_0, \beta_1, \beta_2)^T. \tag{2.4.2}$$

La construction de $\hat{g}$ consiste à trouver un estimateur $\hat{\beta}$ de β que nous obtenons en minimisant une fonction de coût quadratique (SSE) et qui nous donne

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.4.3}$$

Nous pouvons nous reporter à [Gro03] pour les conditions d'existence, d'optimalité ainsi que les propriétés asymptotiques de l' estimateur $\hat{\beta}$.

Remarque 2.4.1. *Notre connaissance à priori de la nature du phénomène que nous étudions assure l'existence d'un terme constant dans le modèle de régression qui est représenté par β_0. En effet, la considération de la constante nous permet d'éviter, pour $i, i = 1, \ldots, N$, lorsque $X_{i1} = 0$ et $X_{i2} = 0$, une valeur $\hat{y}_i = 0$ qui ne serait pas admissible.*

2.4.2 Limites du modèle linéaire

Cette première tentative de régression a fait surgir un certain nombre de problématiques qui expriment les limites d'un modèle linéaire dans son application pour notre contexte. En effet, l'application des résultats théoriques dans la validation du modèle pose des problèmes pour le respect des contraintes définies précédemment.

2.4.2.1 Problème sur la validité du modèle

Dans l'analyse statistique de la régression, pour vérifier la validité globale du modèle nous testons l'influence des régresseurs non constants sur la variable dépendante en posant

$$\begin{cases} H_0 : & \beta_j = 0 \text{ pour } j \in \{1, 2\} \\ H_1 : & \beta_j \neq 0 \text{ pour au moins un } j \in \{1, 2\} \end{cases}$$

et sous l'hypothèse H_0, pour $l = j + 1$, la statistique

$$F = \frac{R^2}{1 - R^2} \frac{n - l}{l - 1}$$

suit une distribution de Fisher à $(l - 1, n - l)$ degrés de liberté, où

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Cette hypothèse sera repoussée en faveur de H_1 si la valeur calculée pour la statistique F est supérieure à la valeur théorique $f_{(l-1,n-l)}(1 - \alpha)$, $(1 - \alpha)$ étant le niveau de confiance du test que nous posons égale à 95%.

La figure FIG.2.3 est un exemple de relation pour laquelle la validité du modèle est remise en cause.

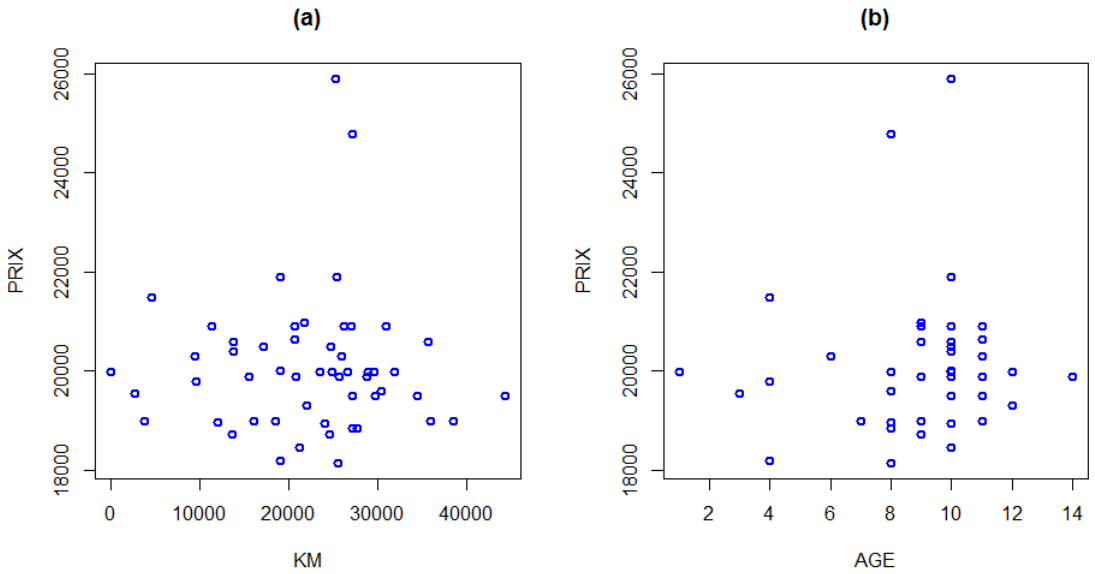

FIGURE 2.3 – PROBLÈME SUR LA VALIDITÉ DU MODÈLE. *Ces figures montrent respectivement les relations entre le Prix et le KM (a) et le Prix et l'âge (b) pour une maille composée de 53 individus pour laquelle la régression n'a pas été validée. Nous pouvons constater qu'il s'agit d'un échantillon constitué principalement de véhicules de moins d'un an. Dans la base de données utilisées dans la démarche expérimentale, ce phénomène a été observée pour* ***4.87% des mailles*** *disponibles.*

2.4.2.2 Problème sur la significativité de coefficients

Après la validation globale du modèle, l'analyse statistique se poursuit en un test de significativité des paramètres qui permet de tester pour chaque β_j les hypothèses,

$$\begin{cases} H_0 : & \beta_j = 0 : \text{ pas d'association entre la j-ième variable explicative et y.} \\ H_1 : & \beta_j \neq 0 : \text{ la j-ième variable explicative est linéairement associée à y.} \end{cases}$$

avec la statistique de test

$$T = \frac{\hat{\beta}_j}{\sqrt{s\left(\hat{\beta}_j\right)}}$$

qui suit sous H_0 une loi de Student à $(n-p)$ dégré de liberté. Nous rejetterons H_0 si l'observation de la statistique T est telle que

$$|T| > t_{n-p}(1-\alpha/2)$$

Il est alors d'usage de rejeter un paramètre lorsqu'il n'est pas significativement différent de zéro puisque cela affirme que la variable n'influe pas sur la réponse.

Toutefois , cette non significativité du paramètre peut surgir également lorsque les covariables sont fortement corrélées, c'est à dire, lorsque nous sommes confrontés à une relation exacte ou proche de la dépendance linéaire entre les variables explicatives (il existe un réel λ tel que $X_1 \approx \lambda X_2$), qui peut vouloir dire que la réponse y peut être parfaitement expliquée soit par X_1 soit par X_2.

Dans plusieurs situations, nous avons rencontré ce problème de colinéarité dont un exemple est rapporté en FIG. 2.4 et FIG.2.5, or la possibilité de rejeter l'une des deux variables proposées serait antinomique aux contraintes imposées par les experts.

2.4.2.3 Problèmes de cohérence avec la réalité

L'analyse métier est effectuée uniquement sur les mailles pour lesquelles la régression a été validée. Elle constitue environ **69%** des mailles disponibles sur la base de données utilisées pour la démarche expérimentale. La FIG.2.6 nous montre la distribution des R^2_{adj} obtenus pour chacune des mailles validées.

Deux problèmes de cohérence avec la réalité sont rencontrés. Le premier concerne le signe des coefficients et le second concerne les valeurs prédites pour le prix. En effet, il n'est pas acceptable que le coefficient associé à l'une des variables $\hat{a}ge$ ou Km pour un échantillon de taille convenable soit positif tout comme il est inadmissible d'observer un prix prédit négatif.

Nous pouvons voir en FIG.2.7 un exemple pour lequel cela s'est produit et nous voyons qu'il s'agit surtout d'un échantillon composé de véhicules récents.

La prédiction des prix négatifs est montré en exemple dans TAB.2.1. Notons que les équations de régression associées à ces mailles sont :

Équation 1 : $Prix = 21548 - 0.01841 * Km - 101.62120 * Age$
Équation 2 : $Prix = 25709 - 0.03364 * Km - 208.10672 * Age$
Équation 3 : $Prix = 14453 - 0.03495 * Km - 55.31392 * Age$
Équation 4 : $Prix = 24506 - 0.05373 * Km - 80.76995 * Age$
Équation 5 : $Prix = 20493 - 0.09155 * Km - 41.36174 * Age$

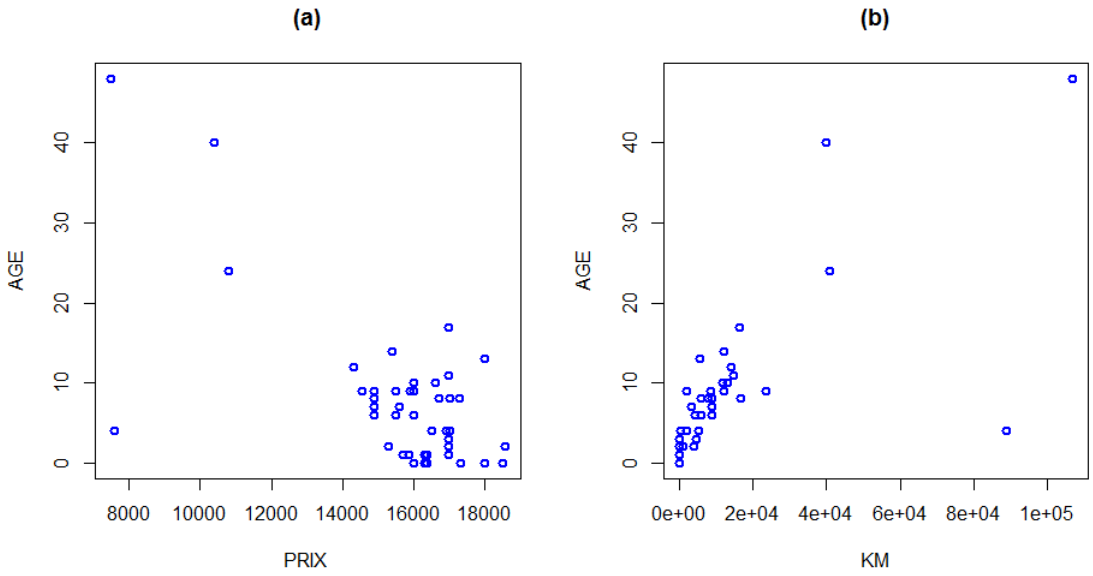

FIGURE 2.4 – PROBLÈME SUR LA SIGNIFICATIVITÉ DU COEFFICIENT LIÉ À LA VARIABLE ÂGE. *Il s'agit d'un exemple sur une maille constituée de 46 individus. Bien que la relation entre le Prix et l'Age(a) ici est mise en évidence, nous observons également une trop forte colinéarité entre l'âge et le KM (b) qui explique la non significativité du paramètre qui y est associé. Ce phénomène a été observé pour* ***17.35% des mailles*** *disponibles dans la base de données utilisée pour l'étude.*

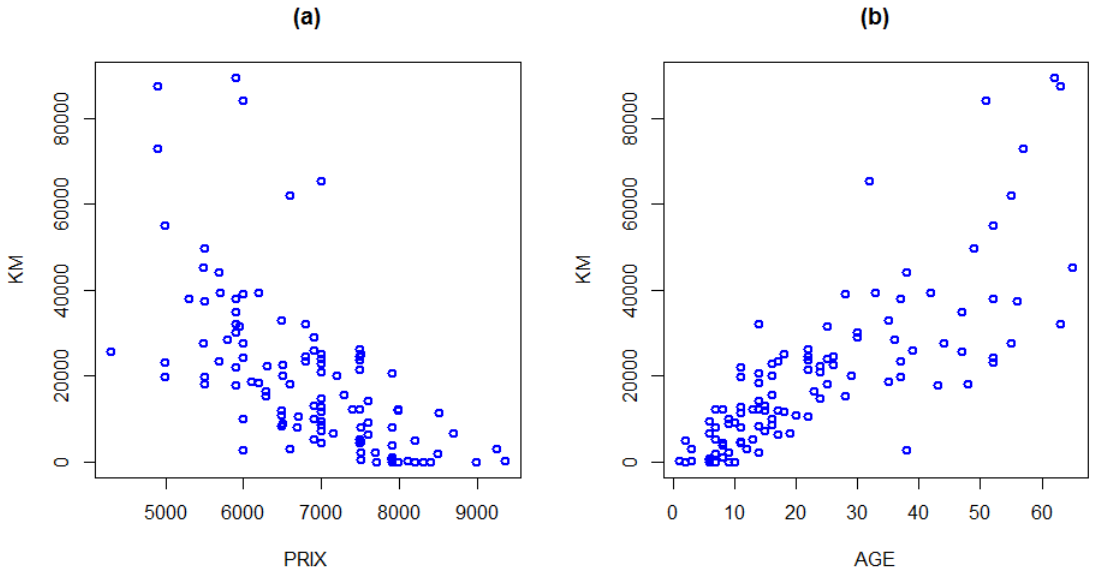

FIGURE 2.5 – PROBLÈME SUR LA SIGNIFICATIVITÉ DU COEFFICIENT LIÉ À LA VARIABLE KILOMÉTRAGE. *Exemple de maille constituée de 76 individus. Nous observons respectivement les relations entre le Prix et l'âge (a) et âge et le KM (b)et la non significativité du paramètre lié au KM ici est expliquée par la forte colinéarité entre l'âge et le Km.* ***9.29% des mailles*** *disponibles dans la base de données utilisée pour l'étude sont concernés par ce phénomène.*

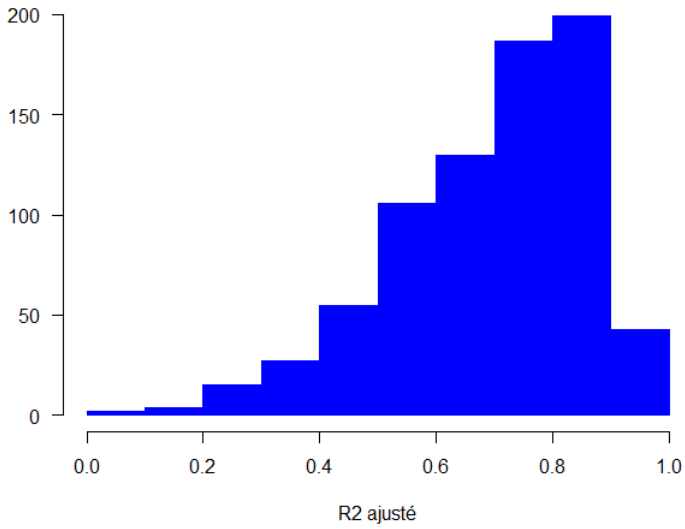

FIGURE 2.6 – Distribution des R^2_{adj} pour les mailles validées.

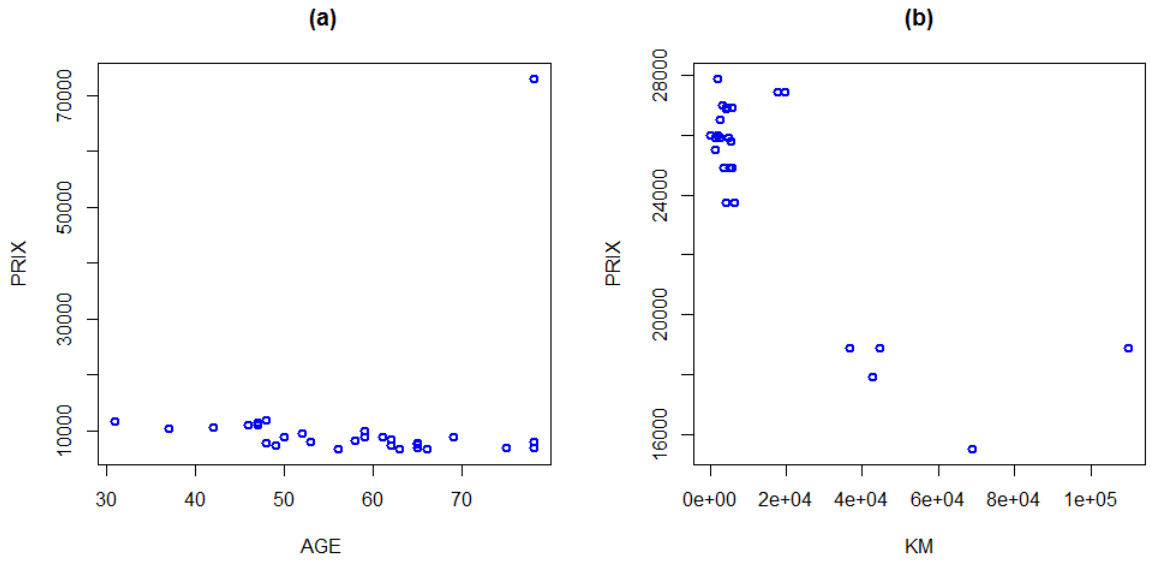

FIGURE 2.7 – PROBLÈME SUR LE SIGNE DES COEFFICIENTS. *Nous observons ici les relations entre l'âge et le Prix pour des mailles constituées respectivement de 30 (a) et de 46(b) individus. Le phénomène peut s'expliquer par la présence dans l'échantillon des observations présentant une relation discordante entre les variables comme par exemple pour la maille associée à (a) :* $(Prix; \hat{a}ge) = (73000\ euros; 78\ mois)$ *et qui ont échappé aux différents filtres appliqués lors des traitements de données.*

MARQUE	MODELE	PRIX	KM	ANNEE	MOIS	AGE	PRIX PREDIT
BMW	SERIE 3	3500	276000	1995	7	184	-2231
CITROEN	C5	4500	371854	2004	11	72	-1784
OPEL	MERIVA	2490	303000	2004	4	79	-506
RENAULT	GRAND ESPACE	1690	174000	1994	8	195	-593
RENAULT	GRAND SCENIC	3500	190000	2001	8	111	-1493
RENAULT	MEGANE	1800	267000	1996	1	178	-246
RENAULT	SCENIC	5800	250000	2006	10	49	-892

TABLE 2.1 – PRÉDICTION DE PRIX À VALEURS NÉGATIVES. *Nous observons dans ce tableau des exemples de VO pour lesquels les prix prédits par le modèle sont négatifs.*

Équation 6 : $Prix = 16739 - 0.01141 * Km - 78.31530 * Age$
Équation 7 : $Prix = 20492 - 0.05962 * Km - 132.22655 * Age$

En vérifiant les années de mise en circulation des véhicules composant chacune de ces mailles, nous constatons que cette situation se produit généralement lorsque l'échantillon est composé de véhicules âgés et /ou présentant un fort kilométrage.

2.4.2.4 Solutions envisagées

L'utilisation du logarithme du prix comme variable dépendante nous a permis de corriger ce problème de valeurs négatives associées aux prix prédits. La limitation de la cotation aux véhicules de moins de 10 ans peut également être une solution.

Pour ce qui est de la non significativité des paramètres, plusieurs méthodes statistiques permettant de pallier au problème de colinéarité sont disponibles dans la littérature lorsque l'on travaille en grande dimension. Principalement, ces méthodes consistent à ajouter un terme de pénalité à la fonction coût à minimiser. Mais cela ne peut être adapté à notre problème puisque notre modèle est parcimonieux.

Ainsi il nous faut trouver une méthode qui permette une régression avec une faible considération de cette colinéarité.

2.5 Un modèle additif

2.5.1 Le modèle proposé

En prenant le logarithme du prix comme variable dépendante, le modèle gagne en terme de réalisme mais ne permet pas d'ajuster les données pour toutes les versions de véhicules disponibles dans la base de données. Cela laisse à supposer que le mécanisme de formation du prix des VO serait différent pour chaque version ou groupe de versions de véhicules et que chaque segment de véhicules se déprécie selon une forme qui lui est propre.

Une régression non-paramétrique nous permettrait d'appréhender ce phénomène [Här90] [Eub88] [Tsy08] [DS11] [Gyö02] étant donné qu'elle ne suppose pas de structure pré–déterminée de la fonction de régression g. Toutefois, les méthodes non paramétriques ont des inconvénients majeurs qui limitent leur utilisation dans la pratique. En effet, elles souffrent d'un manque d'interprétabilité et exigent de très grandes tailles d'échantillons pour être performantes. De plus, les contraintes qui nous sont imposées ne nous laissent pas disposer d'autant d'outils et de règles méthodologiques pour conduire notre travail.

Une forme de modèle qui a été introduit par Stone (1985) [Sto85] et étudié entre autres par Hastie et Tibshirani(1986,1990) [HT86] et plus récemment par Opsomer et Rupper(1999) [Ops00] se rapproche plus de ce que nous cherchons. Ce modèle suppose une structure additive pour la fonction g de la forme

$$g(X) = \beta_0 + g_1(X_1) + g_2(X_2) \tag{2.5.1}$$

où pour $j = \{1, 2\}$, les g_j sont des fonctions unidimensionnelles.

Cette structure additive préserve la possibilité de représenter l'effet de chaque variable facilitant l'interprétation des résultats [BF85] et offre également plus de flexibilité au modèle [BHT89] [HL02]. En effet, si dans la régression linéaire (2.4.1), connaitre les β_j nous permet d'appréhender la variation de la prédiction de y selon les valeurs prises par chacune des X_j , il en est de même pour le cas additif puisque l'effet conjoint des variables explicatives sur la variable dépendante est exprimée comme une somme des effets individuels [LH96].

2.5.2 Les formes possibles pour les fonctions g_1 et g_2

REMARQUE 2.5.1. *Dans toute la suite y désigne le logarithme de la variable prix.* ($y = \log(prix)$).

L'analyse des graphiques en FIG.2.8 et FIG.2.9 nous incite à restreindre les formes de g_j, $j = \{1, 2\}$ à

$$g_j(X_j) = \beta_j X_j^p, p \in \mathbb{R},$$

qui simplifie l'écriture de notre modèle et nous ramène à

$$y = \beta_0 + \beta_1 X_1^{p_1} + \beta_2 X_2^{p_2} + \varepsilon \tag{2.5.2}$$

Le but est donc d'estimer $\beta = (\beta_0, \beta_1, \beta_3)^T$ et $p = (p_1, p_2)^T$ en tenant compte du fait que

(C_6) : pour toutes les valeurs prises par chaque X_j, $j = \{1, 2\}$ nous devons avoir

$$g_j'(X) \leq 0 \,(\text{ g décroissante})\,.$$

(C_7) : $\beta_0 \neq 0$.

(C_8) : $(\beta_0, \beta_1, \beta_2, p_1, p_2)^T \neq (0, 0, 0, 0, 0)^T$

En effet, si $(\beta_1, \beta_2)^T = (0, 0)^T$, le modèle se réduit à une constante ou ne satisfait pas la contrainte imposée dès le départ et le choix de p_1 et p_2 est sans conséquence. De façon similaire, si $p_1 = 0$ (resp. $p_2 = 0$), il n'existera aucun moyen d'identifier β_1 et β_2 séparément, et un nombre infini de paramètres peut être associé au même modèle.

Comme il a été toujours le cas dans ce travail, une connaissance *a priori* du phénomène nous pousse à limiter les valeurs des exposants p_1 et p_2 comprises dans l'intervalle $[-1; 1]$. Ce choix de l'ensemble de valeurs pour p_1 et p_2 nous garantit la monotonie de chaque fonction g_1 et g_2.

Simplifions l'écriture de (2.5.2) en le mettant sous une forme matricielle telle que

$$y = X(p)\beta + \varepsilon \tag{2.5.3}$$

Cette formulation permet de réduire le nombre de paramètre à estimer simultanément et implicitement, le temps de calcul.

La résolution de ce problème dont l'algorithme est détaillée en 2.2 se résume comme suit

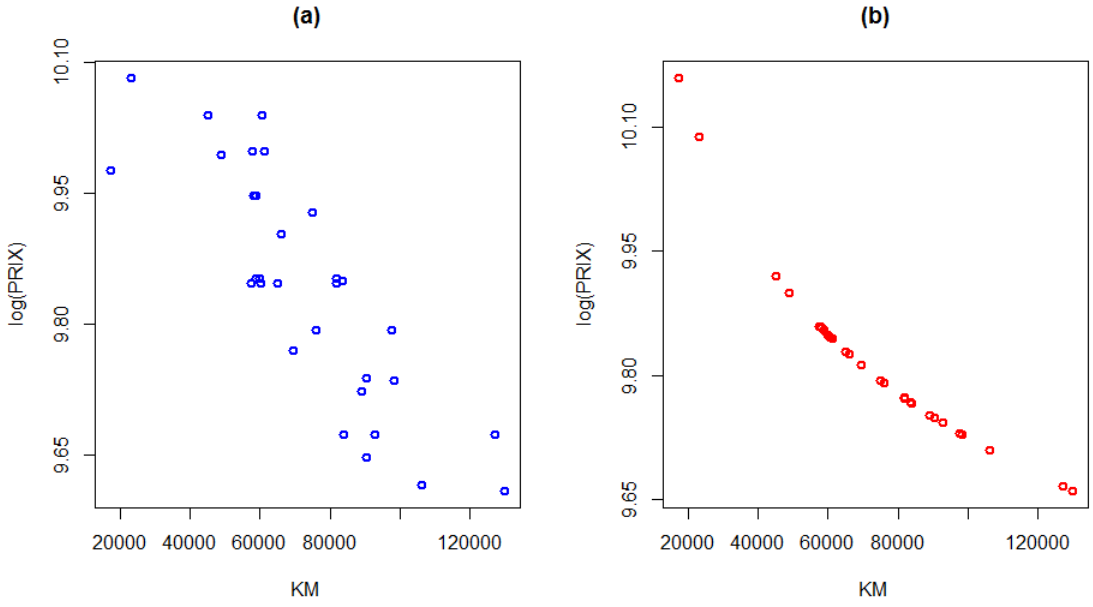

FIGURE 2.8 – FORME POSSIBLE DE LA RELATION ENTRE LES VARIABLES $Prix$ ET Km. Relation entre la réponse $Prix$ (le logarithme du prix) et les variables (Km, Age). *Pour une maille constituée de 31 véhicules, la figure (a) montre la dispersion des points exprimant la relation entre* $\log(Prix)$ *et* Km. *La figure (b) est une représentation de la fonction* $\log(Prix) = 12.967 Km^{-0.025}$. *Un rapprochement entre ces deux graphiques nous permet de supposer qu'une fonction de la forme* $g_1(x) = \beta x^p$ *peut être une des formes possibles de la relation entre les deux variables.*

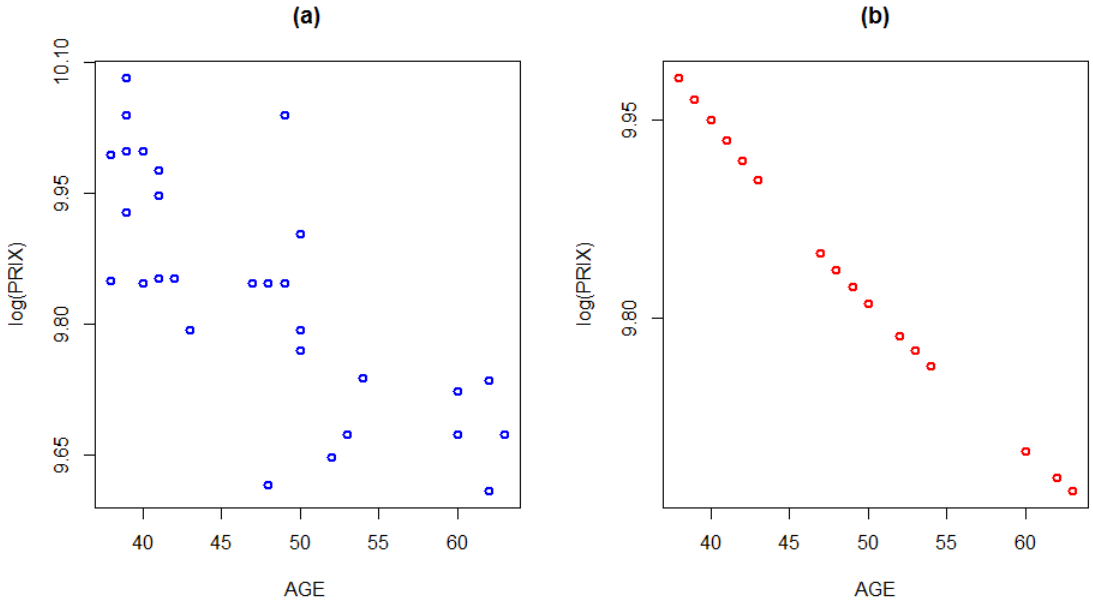

FIGURE 2.9 – FORME POSSIBLE DE LA RELATION ENTRE LES VARIABLES $Prix$ ET Age. *Avec les mêmes données utilisées en FIG.2.8, la figure (a) montre la dispersion des points exprimant la relation entre* $\log(Prix)$ *et* Age *et la figure (b) est une représentation de la fonction* $\log(Prix) = 12.553 Age^{-0.063}$.

1. **Entrée :**
   - y : logarithme du prix d'annonce.
   - X_1 : Km observé.
   - X_2 : $\hat{a}ge$ du véhicule.
3. **Déclaration** de p_1, p_2, β_0, β_1 et β_2.
2. **Définition** de $X'(p) = [1 \quad X_1^{p_1} \quad X_2^{p_2}]$ et de $\beta' = (\beta_0, \beta_1, \beta_2)$.
3. **Calcul** de $\widehat{\beta}(p_1, p_2)$ selon (2.5.4).
4. **Itération :**

   (a) Pour $p_1 = -1$ à 1 **faire** :

   (a1) **Pour** $p_2 = -1$ à 1 **faire** :

   (a11) **Actualiser** $\widehat{\beta}(p1, p_2)$ et $X(p)$.

   (a12) **Si** $\left(p_1\beta_1 X_1^{p_1-1} < 0\right)$ et $\left(p_2\beta_2 X_2^{p_2-1} < 0\right)$ **faire** :

   (a12a) **Calcul** de $\hat{y} = X'(p)\hat{\beta}$ et $\widehat{er}_i = \mid \dfrac{\mathrm{e}^{y_i} - \mathrm{e}_i^{\hat{y}}}{\mathrm{e}^{y_i}} \mid$

   (a12b) **Calcul** de $S(p_1, p_2) = \sum_{i=1}^{N} \mathbb{1}\left(\widehat{er}_i \leq \alpha\right)$.

   (a2) **Revenir** à (a1).
5. **Revenir** à [4.].
6. **Calcul** de $S_{\max} = \max_{p_1,p_2} S(p_1, p_2)$.
7. **Valeurs en sortie** : p_1, p_2, $S_{\max}$ et $\widehat{\beta}$.

TABLE 2.2 – Algorithme de résolution

Étape 1 : calcul de l'erreur de prédiction la plus petite dans le sens des moindres carrées qui nous permet d'obtenir

$$\hat{\beta}(p) = \left(X'(p)X(p)\right)^{-1} X'(p)y \tag{2.5.4}$$

Étape 2 : une exploration guidée renforce le choix $\hat{\beta}(p)$ en explorant la contrainte sur les erreurs relatives afin de trouver les valeurs appropriées pour p_1 et p_2 permettant d'atteindre

$$\max \left(\sum_{i=1}^{N} \mathbb{1} \left\{ \left(\mid \frac{\mathrm{e}^{y} - \mathrm{e}^{\hat{y}}}{\mathrm{e}^{y}} \mid \right) \leq \alpha \right\} \right), \alpha = 0.15 \tag{2.5.5}$$

2.6 La régression sur les quantiles comme alternative aux MCO

2.6.1 Performance des estimateurs l_1 en comparaison avec les estimateurs l_2

L'estimateur $\widehat{\beta}_{ls}$ décrit en Eq.(2.5.4) est le meilleur estimateur linéaire sans biais et de variance minimum lorsque entre autres, les hypothèses d'homoscédasticité et de normalité des

	constante	coefficient(Km)	coefficient(Age)
MCO	20568	-0.0429	-173.5
QR. $\tau = 0.05$	18588	-0.0357	-184.013
QR. $\tau = 0.10$	19389	-0.0369	-193.141
QR. $\tau = 0.25$	20196	-0.0346	-200.431
QR. $\tau = 0.50$	20630	-0.0372	-186.032
QR. $\tau = 0.75$	20996	-0.0494	-150.895
QR. $\tau = 0.90$	21530	-0.0627	-118.822
QR. $\tau = 0.95$	22050	-0.0723	-104.287

TABLE 2.3 – RÉSULTATS COMPARATIFS DES RÉGRESSIONS SUR LES QUANTILES ET MCO. *Ces coefficients peuvent fournir un intervalle de confiance pour la prédiction des prix étant donnés l'âge et le kilométrage.*

résidus sont satisfaites [AB12] [Hoc05] [JKP01]. De ce fait, si nous avons une raison quelconque de douter de la normalité des résidus, notamment en présence des observations extrêmes [Tuk70] [BDH82], la régression par la méthode des moindres carrés peut être remise en cause. Il est alors judicieux dans ce cas d'opter pour une régression robuste [RL05] [EE11], [HR11] [SL92] [Wil12] [AFO94]. La régression sur les quantiles [KB78] est connue pour avoir de bonnes propriétés sous des hypothèses assez restrictives [BD11]. De plus, de par ses principes fondamentaux, elle peut offrir des résultats plus riches et nuancés en ce qui concerne la perception des effets de variables sur une autre [KHM05]. Pour ce qui est du phénomène que nous étudions et au vu des résultats précédents, nous pouvons raisonner dans un sens où pour mieux comprendre les facteurs qui déterminent la variation du prix des VO, il est souhaitable d'estimer les effets des variables Age et Km sur la variable $Prix$ de façon conditionnelle en tenant compte de leur distribution respective [Buc94]. En effet, si ces variables ont un effet très différent sur le prix selon la plage de valeur à laquelle elles appartiennent, une analyse qui considère seulement le prix moyen pourrait conduire à une conclusion biaisée [SW07]. A titre d'illustration, nous pouvons voir en TAB.2.3 les différents coefficients de régression obtenus pour différents niveaux de quantiles et en comparaison avec ceux obtenus avec la régression des MCO dans un cas de régression linéaire.

Nous reprenons le modèle défini en Eq.(2.1.1) en supposant que g a une forme linéaire, c'est à dire

$$y = X'\beta_\tau + \varepsilon \tag{2.6.1}$$

et l'estimateur obtenu par la régression sur les quantiles est

$$\widehat{\beta}_\tau = \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - X_i'\beta) \tag{2.6.2}$$

où $\rho_\tau(.)$ est une fonction convexe [Koe05] [RZ08] telle que $\rho_\tau(z) = \begin{cases} \tau z & \text{si } z \geq 0 \\ (1-\tau z) & \text{si } z < 0 \end{cases}$.

En particulier, pour $\tau = 1/2$, $\hat{\beta}_\tau = \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} \mid y_i - X_i'\beta \mid$.

Les propriétés asymptotiques de $\widehat{\beta}_\tau$ sont délicates à établir car cet estimateur n'admet pas de formulation explicite qui serait similaire à celle de $\widehat{\beta}_{ls}$. Sous certaines conditions de régularités, un résultat principal sur la loi asymptotique de $\widehat{\beta}_\tau$ ainsi qu'un estimateur de la variance asymptotique ont été établis dans [KB78] [DJ00].

La performance de l'estimateur $\widehat{\beta}_{qr}$ de β comparée à $\widehat{\beta}_{ls}$ peut être évaluée par plusieurs critères dont principalement le *breakdown point* et l'*efficacité asymptotique relative*. Le *breakdown point*, comme il a été défini dans [DG07] [NW02] [RL05] correspond à la plus petite fraction de contamination qui peut produire pour $\hat{\beta}$ une valeur arbitrairement éloignée de β. Cette quantité, est exactement la même pour $\hat{\beta}_{ls}$ et $\hat{\beta}_{qr}$ et est égale à $\frac{1}{n}$ sachant que $\lim_{n\to\infty} \frac{1}{n} = 0$ et que la valeur limite optimale, atteinte pour certain estimateurs est de 1/2 [HR11] [EM92]. La comparaison de ces deux estimateurs reposera donc sur l'*Efficacité Relative Asymptotique* (ARE) [Bas56] par laquelle le comportement asymptotique du rapport de leurs variances sera évalué. Nous illustrons la performance d'un estimateur l_1 (la médiane m_d)) par rapport à un estimateur l_2 (la moyenne m) [NW02] pour un cas d'un échantillon de mélange Gaussien $(1-\delta)\mathcal{N}(0,\sigma^2)+\delta\mathcal{N}(0,k\sigma^2)$, où $nVar(m_d) \to \left(4f^2(0)\right)^{-1} \quad as \quad n \to \infty$ et $nVar(m) \to_{n\to\infty} V_f$, f étant la densité du modèle, on a lorsque n devient grand nous avons

$$Eff(m, m_d) = \frac{Var(m)}{Var(m_d)} = \frac{2}{\pi}\left(1+\delta(k^2-1)\right)\left[\frac{\delta}{k}+(1-\delta)\right]^2$$

Notons, que lorsque k croît $Var(m)$ tend vers l'infini, alors que $Var(m_d)$ reste bornée.

Les illustrations en FIG.2.10 nous montrent la sensibilité de l'estimation par la norme l_2 par rapport à l'écart à la normalité et à la présence des valeur aberrantes.

2.6.2 La régression non linéaire sur les quantiles

Réecrivons l'Eq.2.1.1

$$y = g_\theta(X) + \varepsilon \tag{2.6.3}$$

l'indice θ ayant été ajouté pour indiquer le caractère paramétrique de la fonction g.

Dans la section précédente (§.2.5.2), nous avons étudié les formes possibles de la fonction g_θ et nous retenons donc que g_θ est assimilée à une fonction non linéaire. L'algorithme présenté en 2.2 estime les paramètres de la fonction par la méthode des moindres carrés. Nous nous proposons maintenant d'estimer g_θ par une régression sur les quantiles qui ramène notre problème à une résolution d'une régression non linéaire sur les quantiles.

Tenant compte de la fonction $\rho(z)$, nous définissons l'erreur en généralisation du modèle telle que

$$M(g_\theta) := \mathrm{E}_{p(x,y)}\left[\rho_\tau\left(y - g_\theta(x)\right)\right] \tag{2.6.4}$$

et nous posons

$$g_\theta^\tau = \arg\min_\theta M(g_\theta) \tag{2.6.5}$$

où g_θ^τ est une fonction, non nécessairement unique, correspondant à la fonction théorique optimal pour le problème de régression sur les quantiles. $p(x,y)$ nous est inconnue, la résolution du problème de régression, pour un niveau de quantile $\tau \in (0,1)$, se réduit ainsi à la minimisation du risque empirique

$$M_n^\tau(g_\theta) = \sum_{i=1}^{n} \rho_\tau(y_i - g_\theta(x_i)) \tag{2.6.6}$$

et soit

$$\widehat{g}_\theta^\tau = \arg\min_\theta M_n^\tau(g_\theta) \tag{2.6.7}$$

La performance de cet estimateur peut être établie à travers les deux critères (Crit1) et (Crit2) suivant :

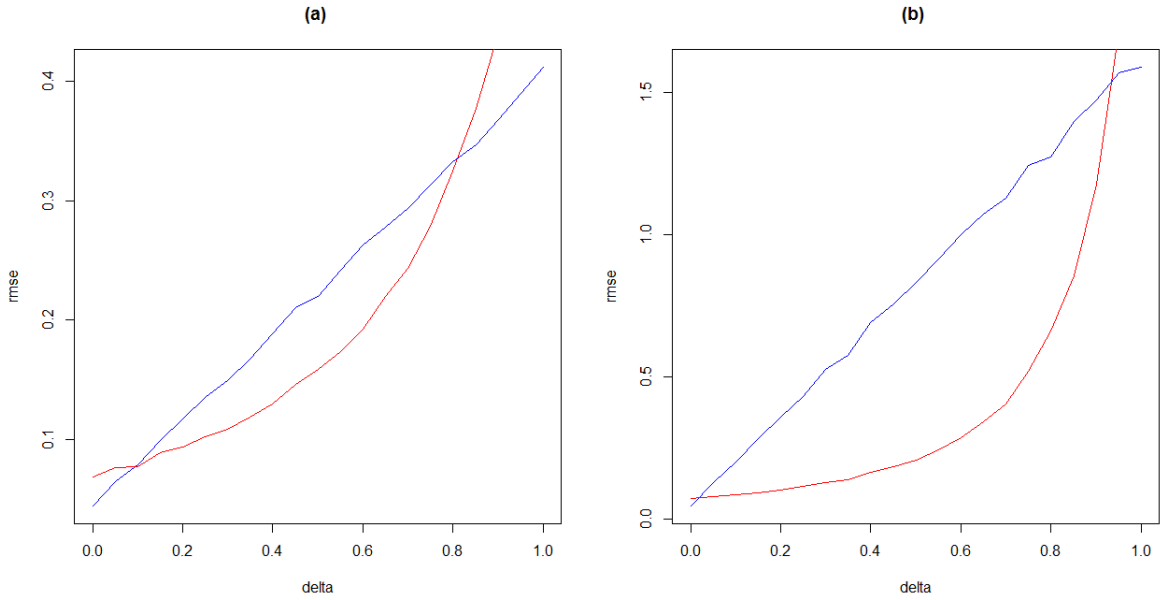

FIGURE 2.10 – COMPARAISON DES ESTIMATEURS L_1 ET L_2. *Nous représentons les valeurs du RMSE pour les deux modes d'estimation* l_1 (——) *et* l_2 (– – –) *pour différentes valeurs prises par* δ *dans* $\{0, 0.05 \ldots, 0.95, 1\}$ *pour le modèle* M_2. *Par cette expérience nous montrons que dans un cadre univarié dans lequel nous supposons que* $y_i = \mu + \varepsilon_i, \quad i = 1, \ldots, n.$ *où* μ *est un paramètre réel que nous voulons estimer, et* $\varepsilon \sim (1-\delta)\mathcal{N}(0, \sigma^2) + \delta\mathcal{N}(0, k\sigma^2)$. *Le graphique (a) correspond à une valeur de* $k = 3$ *et (b) à une valeur* $k = 6$. *Sauf dans le cas où l'on n'a pas de contamination, c'est -à dire nous sommes dans un cas gaussien, l'erreur moyenne de prédiction obtenue par la méthode* l_2 *reste essentiellement plus élevée que celle obtenue par la méthode* l_1. *L'estimateur* $\hat{\beta}_\tau$ *fournit une meilleure alternative à la moyenne empirique* $\hat{\beta}_{ls}$ *au sens de l'erreur quadratique moyen.*

(**Crit1**) Dans la mesure du possible, $\widehat{g}_\theta^\tau$ doit satisfaire les propriétés d'un quantile, c'est à dire

$$Pr_{X,Y}\left\{|Pr\{y < \widehat{g}_\theta^\tau(x)\} - \tau| \geq \epsilon\right\} \leq \delta \tag{2.6.8}$$

(**Crit2**) $g_{\hat{\theta}_\tau}$ doit être "proche" de g_θ. Comme g_θ est inconnue, nous pouvons prendre en considération le fait que g_θ^τ est la fonction qui minimise la quantité M. Nous pourrions donc évaluer les performances de $\widehat{g}_\theta^\tau$ en évaluant dans quelle mesure le minimum $M(g_\theta^\tau)$ est proche de l'estimateur $\widehat{g}_\theta^\tau$ et établir

$$Pr_{X,Y}\left\{M\left[\widehat{g}_\theta^\tau\right] - M\left[g_\theta^\tau\right] > \epsilon\right\} \leq \delta. \tag{2.6.9}$$

sachant que

$$\begin{aligned} M[\widehat{g}_\theta^\tau] - M[g_\theta^\tau] &\leq |M[\widehat{g}_\theta^\tau] - M_n[\widehat{g}_\theta^\tau]| + |M_n[g_\theta^\tau] - M[g_\theta^\tau]| && (2.6.10)\\ &\leq \sup_{g_\theta} |M[\widehat{g}_\theta^\tau] - M_n[\widehat{g}_\theta^\tau]| + |M_n[g_\theta^\tau] - M[g_\theta^\tau]| && (2.6.11) \end{aligned}$$

puisque $M_n[\widehat{g}_\theta^\tau] \leq M_n[g_\theta^\tau]$.

La plupart des résultats concernant les propriétés asymptotiques d'un estimateur de régression sur les quantiles dans le cas linéaire ont été étendus au cas non linéaire. La théorie qui y est associée dérive de celle des M-estimateurs [vdV00] et il a été établi que si les fonctions g_θ sont paramétriques, identifiables et suffisamment lisses et que si la fonction densité de probabilité du bruit existe et est positive alors la normalité asymptotique du M-estimateur peut être démontrée comme nous pouvons le voir dans [Wei91].

Nous proposons d'estimer les paramètres du modèles sous la forme présentée en Eq.(2.5.2) par une régression sur les quantiles . Nous expliciterons dans la section suivante l'apport de ce mode d'estimation.

NOTE 1. *Dans un travail en annexe à ce chapitre, nous avons montré pour la regression non linéaire sur les quantiles que des résultats asymptotiques peuvent être établis dans des cas de modèles non identifiables. L'étude a été conduite en considérant que la fonction de régression g_θ est une fonction de perceptrons multicouches (MLP).*

2.7 Solution industrielle retenue. Comparaison des modèles et application sur des données réelles

Nous comparons les quatre modèles suivants :

- le modèle linéaire $(ML) : y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
- le modèle linéaire avec le log du prix $(MLlog) : \log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
- le modèle avec puissance estimée par MCO avec le $\log(prix) : MAdd$.
- le modèle avec puissance estimée par une régression sur les quantiles avec le $\log(prix) : Mqr$.

L'évaluation de la performance de chaque modèle est étudiée. L'illustration de cette performance concerne deux aspects. Le premier aspect d'analyse s'attache à montrer la capacité du modèle à s'adapter aux différentes mailles de VO constituant la base de données. Le deuxième consiste à évaluer la capacité de prédiction du modèle au sein d'une maille considérée.

Nous prenons comme référence les résultats obtenus pour le modèle ML et nous nous proposons de répondre à plusieurs problématiques que nous énoncerons par la suite.

REMARQUE 2.7.1. *Le temps de calcul est un facteur que nous ne prenons pas en compte.*

	Apprentissage	Test
Mailles disponibles	2 399	2 399
Identifiant VO	150 696	63 058

TABLE 2.4 – DESCRIPTION DE LA BASE DE DONNÉES. *La base de données correspond à des annonces dédoublonnées parues au mois de février 2011 et contient, après traitement 213 754 identifiants VO et 2 399 mailles et dont le nombre d'individus constituant chacune des mailles est supérieur à 25.*

Modèle	Nb de mailles modélisées	Proportion
ML	1 364	56.86 %
$MLLog$	1 381	57.56 %
$MAdd$	2 380	99.21 %
Mqr	2 380	99.21 %

TABLE 2.5 – COMPARAISON DES PERFORMANCES GLOBALES. *Nous observons ici la proportion des mailles pour laquelle la regression a été validée parmi les mailles disponibles dans la base de données utilisée pour l'apprentissage. Les modèles $MAdd$ et Mqr ont été validés sur presque la totalité des mailles.*

Nous réservons 70% des données pour la construction des modèles (Ech_A) et 30% pour les tester (Ech_T). Chacune des méthodes est construite sur Ech_A et testée sur Ech_T afin de vérifier la capacité de chaque modèle à prédire le même phénomène sur de nouvelles données.

Nous calculons les valeurs prédites pour le prix à partir de Ech_T ainsi que les erreurs relatives ($|\frac{Prix_i - Prediction_i}{Prix_i}|$) qui leur sont associées.

Nous répétons cette étape un nombre $N = 100$ fois en effectuant un tirage aléatoire avec remise sur la base de données et nous estimons la performance d'une méthode en faisant la moyenne des performances obtenues sur les échantillons de test. Dans la production des résultats et pour chacun des cas, les chiffres traduisant une meilleure performance sont marqués en rouge.

Suite au fait que certaines mailles n'ont pu être modélisées selon la méthode proposée, la taille de la base Test ne sera pas la même selon le cas où l'on teste le modèle ML, $MLLog$, $MAdd$ ou Mqr que nous pouvons voir en TAB.2.6. Les proportions que nous déterminerons seront pondérées par la proportion de chaque base de données utilisée par rapport à la base de

Modèle	Nb de VO	Proportion
ML	45 956	72.88 %
$MLLog$	46 363	73.52 %
$MAdd$	62 571	99.23 %
Mqr	62 571	99.23 %

TABLE 2.6 – DESCRIPTION DE LA BASE TEST. *Il s'agit de la taille de la base qui servira de test pour chacun des modèles correspondant.*

Modèle	$\leq 5\%$	$\leq 10\%$	$\leq 15\%$	$\leq 20\%$	$\leq 25\%$	$\leq 30\%$
ML	21 034 (33,36%)	33 884 (53,74%)	40 183 (63,72%)	43 082 (68,32%)	44 435 (70,47%)	45 085 (71,50%)
$MLLog$	21 337 (33,84%)	34 810 (55,20%)	41 219 (65,36%)	44 021 (69,81%)	45 205 (71,68%)	45 748 (72,54 %)
$MAdd$	23 718 (37,61%)	41 402 (65,66%)	51 257 (81,29%)	55 965 (88,75%)	58 357 (92,55%)	59 548 (94,44%)
Mqr	33 427 (53,01%)	48 621 (77,11%)	55 926 (88,69%)	59 260 (93,98%)	60 847 (96,50%)	61 568 (97,64%)

TABLE 2.7 – CAPACITÉ DE PRÉDICTION DES MODÈLES. *Nous mesurons ici d'un point de vue global la capacité de prédiction de chaque modèle en appréciant les valeurs prises par les* $CVO(i), i = 1, \ldots, n$. *Dans la colonne* $"\leq 5\%"$ *, il y a* $33,36\%$ *d'individus pour lesquels les erreurs relatives obtenues sont plus petites que 5% lorsque l'on applique les coefficients du modèle* ML*, cette proportion est de* $33,84\%$ *avec les coefficients du modèle* $MLLog$ *tandis qu'elle s'élève à* $53,01\%$ *lorsque les coefficients du modèle* Mqr *sont utilisés.*

données initiale.

PROBLÉMATIQUE 2. *Comment sont distribués les écarts relatifs si nous observons la base test ?*

La réponse à cette problématique nous donne une vue globale sur la capacité de prédiction de chaque modèle et les résultats sont reportés en TAB.2.7.

Le seuil maximal est fixé à 0.30. En effet, un écart supérieur à 0.30 signifie d'une manière pratique que le prix prédit est bien loin du prix observé et que la qualité de l'estimation peut être remise en question. Une prédiction de ce genre perd tout son sens pratique et ne peut être exploitée à des fins commerciales.

La nouvelle structure de la relation entre le prix, le km et l'âge d'une part, l'utilisation de la régression sur les quantiles et l'application de la contrainte en probabilité d'autre part, ont permis d'obtenir des résultats plus satisfaisants. La solution proposée permet de prédire le prix des VO avec une plus grande précision pour une grande majorité des données.

PROBLÉMATIQUE 3. *Comment est distribuée la moyenne des écarts relatifs calculée pour chaque maille ?*

La réponse à cette problématique consiste à apprécier la performance de chaque modèle à l'échelle de la maille, les résultats correspondants sont traduits en TAB.2.9 et FIG.2.11. Notons que dans la production de ces résultats, nous avons exclu de notre base de données les individus pour lesquels l'erreur relative était supérieure à 0.5. Cette initiative nous est permise dans la mesure où nous pouvons supposer que ces observations ne sont autres que des cas atypiques (voir TAB.2.8), s'agissant des outliers qui ont échappé aux différents filtres appliqués lors du traitement des données. Ces observations pourraient biaiser le calcul de la moyenne que nous nous proposons de faire. Si l'on se place dans un cadre pratique, en tant qu'observateur, c'est cette forme de résultat qui donnera le plus de visibilité sur la capacité de prédiction de chaque modèle ainsi que le gain de performance obtenu pour chaque modèle.

Modèle	Nb de VO
ML	194
$MLLog$	135
$MAdd$	1448
Mqr	223

TABLE 2.8 – NOMBRE DE CAS ATYPIQUES. *Ces chiffres représentent pour chacun des modèles le nombre d'observations pour lesquelles l'erreur relative est supérieure à* 0.5.

	$\leq 5\%$	$\leq 10\%$	$\leq 15\%$	$\leq 20\%$	$\leq 25\%$	$\leq 30\%$
ML	138 (5,75%)	1 001 (41,73%)	1 270 (52,94%)	1 342 (55,94%)	1 353 (56,40%)	1 360 (56,69%)
$MLLog$	153 (6,38%)	1 095 (45,64%)	1 321 (55,06%)	1 367 (56,98%)	1 376 (57,35%)	1379 (57,48%)
$MAdd$	170 (7,09%)	1 303 (54,32%)	1 837 (76,58%)	2 011 (83,83%)	2 095 (87,33%)	2 131 (88,83%)
Mqr	773 (32,22%)	2 094 (87,29%)	2 332 (97,21%)	2 370 (98,79%)	2 378 (99,13%)	2 380 (99,21%)

TABLE 2.9 – APPRÉCIATION DE LA PERFORMANCE À L'ÉCHELLE DE LA MAILLE. *montre le gain relatif par rapport au modèle linéaire. Dans la colonne "* $\leq 10\%$ *" , il y a* $41,73\%$ *de mailles pour lesquelles la moyenne des erreurs relatives obtenues sont plus petites que 10% lorsque l'on applique les coefficients du modèle* ML*, cette proportion est de* $45,64\%$ *avec les coefficients du modèle* $MLLog$ *tandis qu'elle s'élève à* $87,29\%$ *lorsque les coefficients du modèle* Mqr *sont utilisés.*

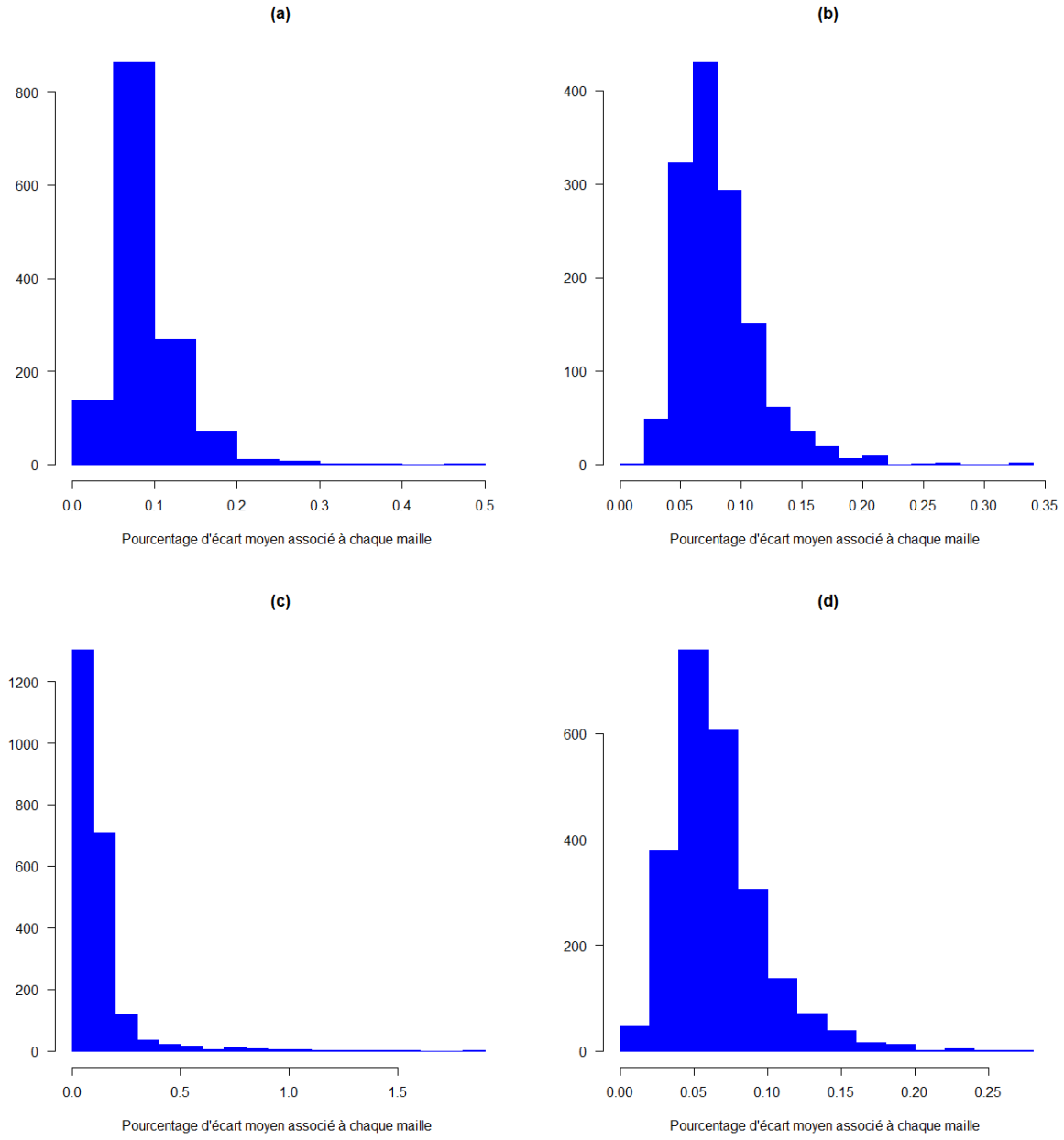

FIGURE 2.11 – APPRÉCIATION DE L'ERREUR RELATIVE À L'ÉCHELLE DE LA MAILLE. *Cette figure traduit ce que nous avons présenté en TAB.2.9. (a) correspond au modèle ML, (b) au modèle $MLLog$, (c) au modèle $MAdd$ et (d) au modèle Mqr.*

2.8 Conclusion

Dans ce chapitre, nous nous sommes proposés de mettre en place une méthodologie permettant de modéliser et prédire convenablement les prix des véhicules d'occasion.

Une analyse descriptive a permis d'apprécier la structure des corrélations qui existent entre la variable $Prix$ et les variables supposées expliquer sa dépréciation. Nous en avons conclu que le prix des VO est une fonction décroissante des variables Km et Age. Une trop grande dispersion des prix observés à travers les différentes marques de véhicule nous a poussé à effectuer notre analyse à une maille plus fine correspondant à une combinaison : **Marque- Modèle - Energie - Carrosserie - Motorisation** *(ex : Peugeot 207 1.5 HDI Berline).*

La modélisation du prix par un modèle de régression linéaire (ML) a fait apparaitre plusieurs problématiques limitant son exploitation. L'utilisation du Logarithme du prix ($MLLog$) a permis de traiter une partie des problèmes sans toutefois répondre complètement aux objectifs industriels. Un modèle inspiré des modèles additifs ($MAdd$) a été mis en place. La régression sur les quantiles a été abordée et ses propriétés théoriques ont été étudiées. L'application de la régression sur les quantiles (Mqr) sur le modèle ($MAdd$) a été proposée.

Par une étude comparative des quatre modèles effectuée sous différents aspects sur les données d'Autobiz, le modèle Mqr constitue une modélisation capable de traduire le mécanisme de dépréciation du prix, de tenir compte de la spécificité de notre problème et de répondre aux exigences industrielles. La cohérence du modèle est réellement satisfaisante et son apport par rapport aux autres modèles a été mise en évidence.

La validation par les experts ainsi que la mise en production d'un modèle issu de notre travail confirment la pertinence de notre approche. Les figures FIG.2.12 et FIG.2.13 sont des captures d'écran effectuées sur SystemVO et sont des exemples de la mise en production et de l'exploitation commerciale de ce projet de recherche.

Toutefois, le manque à gagner observé à travers les réponses aux différentes problématiques posées lors de la validation expérimentale nous apportent des informations sur la qualité des données ainsi que le manque d'efficacité des méthodes de détection des outliers. Les performances globales du modèle retenu pourraient être ainsi améliorées en proposant un meilleur programme de traitement des données.

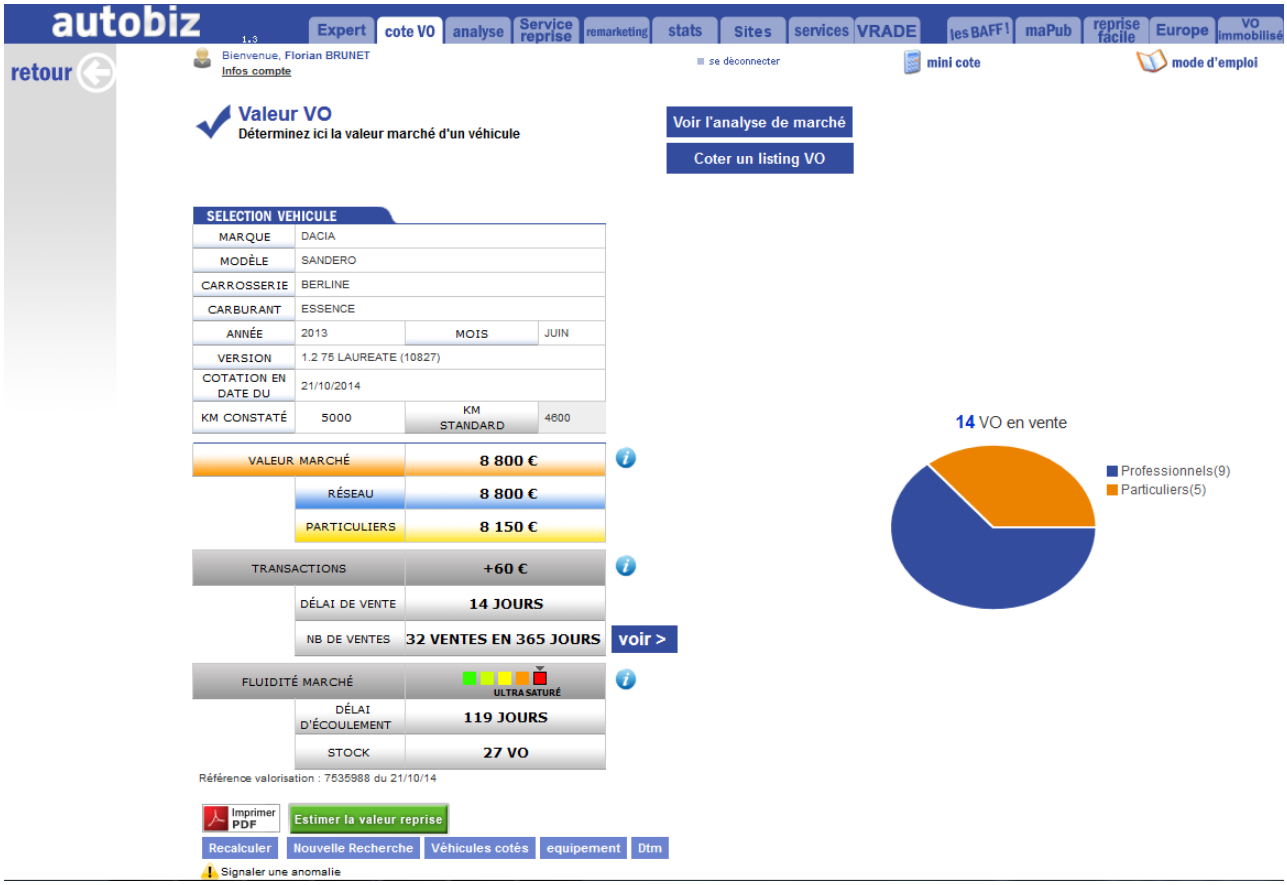

FIGURE 2.12 – RESTITUTION DE LA COTE VO EN TANT QUE VALEUR DE MARCHÉ. *Sur la demande d'un client, et sur les informations qu'il fournit, une valeur de marché lui est restituée. Cette valeur correspond au prix de transaction d'un professionnel à un particulier. Un prix ajusté correspondant au prix de transaction d'un particulier à un particulier lui est également présenté.*

OFFRE PROS

VENDEUR	ACTIVITÉ	CP	ANNÉE	KM	PRIX ANNONCE	VALEUR MARCHÉ	ÉCART
Garage Joostens	Réparateur agréé VP	80	2013	10.400	7 990 €	8 450 €	OK
en cours d'identification		62	2013	7.000	8 500 €	8 800 €	OK
en cours d'identification		34	2013	7.500	8 600 €	8 750 €	OK
en cours d'identification		62	2013	7.500	9 000 €	8 900 €	OK
Renault Chambray Les Tours	Conc. VP	37	2013	4.877	9 500 €	9 150 €	OK
Renault Chambray Les Tours	Conc. VP	37	2013	5.447	9 500 €	9 100 €	OK
en cours d'identification		78	2013	3.225	9 690 €	9 200 €	OK
Peugeot Bressuire	Conc. VP	85	2013	10.649	9 290 €	8 600 €	OK
Vent D'ouest Automobiles	Centre VO	44	2013	17.652	8 990 €	8 100 €	OK
Peugeot Nogent Le Rotrou	Conc. VP	28	2013	10.649	9 290 €	8 350 €	+ 950 €

1-10|11-14|

FIGURE 2.13 – VALEURS DE MARCHÉ ET PRIX D'ANNONCE. *Nous observons pour un extrait de stock chez un professionnel et pour une version de véhicule, le prix d'annonce ainsi que les valeurs de marché qui lui sont associés. Un écart relatif supérieur à* 0.10 *est signalé en rouge.*

General bound of overfitting for MLP regression models

S.-F. Dimby and J. Rynkiewicz

Universite Paris 1 - SAMM
90 Rue de Tolbiac, 75013 Paris - France

Abstract. We consider nonlinear quantile regression involving multilayer perceptrons (MLP). In this paper we investigate the asymptotic behavior of quantile regression in a general framework. First by allowing possibly non-identifiable regression models like MLP's with redudant hidden units, then by relaxing the conditions on the density of the noise. In this paper, we present an universal bound for the overfitting of such model under weak assumptions. The main application of this bound is to give a hint about determining the true architecture of the MLP quantile regression model. As an illustration, we use this theoretical result to propose and compare effective criteria to find the true architecture of such regression model.

1 Introduction

Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. Some q-quantiles have special names : The 2-quantile is called the median, the 4-quantiles are called quartiles and the 10-quantiles are called deciles.

We can define the quantile through a simple alternative expedient as an optimization problem. Just as we can define the sample means as the solution to the problem of minimizing a sum of squared residuals, we can define the median as the solution to the problem of minimizing a sum of absolute residuals. More generally, if $y_1, \cdots, y_n$ are observed values, solving

$$\min_{m \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - m) \tag{1}$$

where the cost function $\rho_\tau(z) = \tau \times (-z) \times \mathbf{1}_{\mathbb{R}^-}(z) + (1-\tau) \times z \times \mathbf{1}_{\mathbb{R}^+}(z)$ is the tilted absolute function. Having succeded in defining the unconditional quantiles as an optimization problem, it is easy to define conditional quantiles in an analogous fashion. To obtain an estimate of the conditional quantile, we simply replace the scalar m in the equation 1 by a function $f(x_i)$, where x_i are the covariate variables.

2 The model

The basic model is a possibly nonlinear regression model with an additive error. It is given by

$$Y_t = f_\theta(X_t) + \varepsilon_t \tag{2}$$

Where $(Y_t)_{1\leq t\leq n}$ are the observations, $(X_t)_{1\leq t\leq n}$ are random covariates and $(\varepsilon_t)_{1\leq t\leq n}$ are unobserved error term. The regression function f is assumed to be an MLP function with k hidden units can be written :

$$f_\theta(x) = \beta + \sum_{i=1}^{k} a_i \phi\left(w_i^T x + b_i\right),$$

with $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_{11}, \cdots, w_{1d}, \cdots, w_{k1}, \cdots, w_{kd})$ the parameter vector of the model and ϕ a bounded transfer function, usually a sigmoïdal function. θ belongs to $\Theta_k \subset \mathbb{R}^{k\times(d+2)+1}$, a compact (i.e. closed and bounded) set of possible parameters. The quantile regression estimator $f_{\hat{\theta}_\tau}$ is obtained by solving the optimization problem :

$$\begin{array}{l} \min_{\theta\in\Theta_k} M_n^\tau(f_\theta) \\ \text{with } M_n^\tau(f_\theta) = \sum_{i=1}^n \rho_\tau(y_i - f_\theta(x_i)) \end{array} \tag{3}$$

For a function $\rho_\tau(.)$ equal to

$$\rho_\tau(z) = \tau \times (-z) \times \mathbf{1}_{\mathbb{R}^-}(z) + (1-\tau) \times z \times \mathbf{1}_{\mathbb{R}^+}(z) \tag{4}$$

In the sequel, let f_{θ_τ} be a, possibly not unique, function such that

$$f_{\theta_\tau} = \arg\min_{\theta\in\Theta_k} M(f_\theta) \text{ with } M(f_\theta) = \int \rho_\tau(y - f_\theta(x)) dP(x,y). \tag{5}$$

f_{θ_τ} is the optimal function for the theoretical quantile regression problem.

2.1 Asymptotic distribution

If the possible functions f_θ are parametric, identifiable and smooth enough function and if the density of the noise exists and is positive then asymptotic normality of the M-estimator can be shown (see Koenker and Basset [1] for the linear case and Weiss [6] for the non-linear case and $\frac{1}{2}$-quantile). However it is possible to give more general results using emprical processes theory. In this paper we prove a general bound valid even if the optimal functions f_{θ_τ} are not unique and without assumptions on the density of noise, except moment conditions.

2.1.1 A general bound for $M_n^\tau(f_\theta)$

We will prove an inequality bounding the difference:

$$M_n^\tau(f_\theta) - M_n^\tau(f_{\theta_\tau}).$$

For an square integrable function $g(X,Y)$ the L_2 norm is:

$$\|g(X,Y)\|_2 := \sqrt{\int g^2(x,y) dP(x,y)}.$$

Let $\lambda > 0$ be a constant, the generalized derivative function is defined as:

$$d_\theta^\lambda(X,Y) = \frac{\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}}{\|\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}\|_2} = \frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}-1}{\|e^{-\lambda\rho_\tau(Y-f_\theta(X))-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}-1\|_2} \tag{6}$$

and let us define $\left(d_\theta^\lambda\right)_-(x,y) = \min\left\{0, d_\theta^\lambda(x,y)\right\}$. For now, let us assume that d_θ^λ is well defined, this point will be discuss later. We can state the following inequality:

Inequality:
for $\lambda > 0$,

$$\sup_{\theta\in\Theta_k}\left(M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)\right) \le \frac{1}{2\lambda}\sup_{f\in\mathcal{F}}\frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n\left(d_\theta^\lambda\right)_-^2(x_i,y_i)} \tag{7}$$

Proof:
The proof is very similar to the proof for the least square estimator obtained by Rynkiewicz [4]. We have

$$\begin{aligned}
&\left(M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)\right) = \\
&\tfrac{1}{\lambda}\textstyle\sum_{i=1}^n \log\left(1 + \|\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}\|_2 d_\theta^\lambda(x_i,y_i)\right) \\
&\le \sup_{0\le p\le\|\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}\|_2}\tfrac{1}{\lambda}\textstyle\sum_{i=1}^n\log\left(1+pd_\theta^\lambda(x_i,y_i)\right) \\
&\le \sup_{p\ge0}\tfrac{1}{\lambda}\left(p\textstyle\sum_{i=1}^n d_\theta^\lambda(x_i,y_i) - \frac{p^2}{2}\sum_{i=1}^n\left(d_\theta^\lambda\right)_-^2(x_i,y_i)\right).
\end{aligned}$$

Since for any real number u, $\log(1+u) \le u - \frac{1}{2}u_-^2$. Finally, replacing p by the optimal value, we found

$$\left(M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)\right) \le \tfrac{1}{2\lambda}\frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n\left(d_\theta^\lambda\right)_-^2(x_i,y_i)}$$

■

This inequality allows to prove that $M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)$ is bounded in probability under simple assumptions. This may be applied to model selection as discussed in the next section.

2.2 Application : selection of models

In this section, the set $\mathcal{F}$ of possible regression function will be set to

$$\mathcal{F} = \cup_{k=1}^K \mathcal{F}_k,$$

with $\mathcal{F}_{k_1} \subset \mathcal{F}_{k_2}$ for $k_1 < k_2$ and K is a, possibly huge, fixed constant. Let k^0 be the minimal dimension of the functional space needed to realize the true regression function f_τ. As examples, for parametric linear regression $\mathcal{F}_k$ may be seen as regression over k covariates and for multilayer perceptron $\mathcal{F}_k$ may be

perceptrons with k hidden units. We define the minimum-penalized estimator of k^0, as the minimizer $\hat{k}$ of

$$T_n(k) = \min_{f \in \mathcal{F}} \left(M_n^{\tau}(f_\theta) + a_n(k) \right) \tag{8}$$

Let us assume the following assumptions:

(A1) $a_n(.)$ is increasing, $n \times (a_n(k_1) - a_n(k_2))$ tends to infinity as n tends to infinity, for any $k_1 > k_2$ and $a_n(k)$ tends to 0 as n tends to infinity for any k.

(A2) It exists $\lambda > 0$ so that $\left\{ d_\theta^\lambda, f \in \mathcal{F} \right\}$ is a Donsker class (see van der Vaart [5]).

We now have:
Theorem:
Under **(A1)** *and* **(A2)**, $\hat{k}$ *converges in probability to the true dimension* k^0.

The proof of this theorem is exactly the same as in Rynkiewicz [4].

The assumption **(A1)** is fairly standard for model selection, in the Gaussian case **(A1)** will be fulfilled by the BIC criterion. The assumption **(A2)** is more difficult to check. First we note:

$$\begin{array}{l} \left(e^{-\lambda(\rho_\tau(Y - f_\theta(X)) - \rho_\tau(Y - f_{\theta_\tau}(X)))} - 1 \right)^2 = \\ e^{-2\lambda(\rho_\tau(Y - f_\theta(X)) - \rho_\tau(Y - f_{\theta_\tau}(X)))} - 2e^{-\lambda(\rho_\tau(Y - f_\theta(X)) - \rho_\tau(Y - f_{\theta_\tau}(X)))} + 1 \end{array}$$

So, d_θ^λ is well defined if $E\left[e^{-2\lambda(\rho_\tau(Y - f_\theta(X)) - \rho_\tau(Y - f_{\theta_\tau}(X)))} \right] < \infty$, so, if $\lambda > 0$ exists such that $E\left(e^{\lambda|\varepsilon|} \right) < \infty$ and $\sup_{f \in \mathcal{F}} E\left(e^{\lambda|f_\theta(X)|} \right) < \infty$ (i.e. ε and $f_\theta(X)$ admit exponential moments). Finally, using the same techniques of reparameterization as in Rynkiewicz [3], assumption **(A2)** can be shown to be true for linear regressions or MLP models with sigmoïdal transfer functions, if the set of possible parameters Θ is compact.

3 A little experiment

The theoretical penalization terms of the previous section can be chosen among a wide range of functions (see condition **A1**). In the sequel, a little experiment is conducted to assess the right rate of penalization to guess the "true" architecture of a model.

Consider a simulated model:

$$Z_t = F_{\theta^0}(X_t, Y_t) + \varepsilon_t, t = 1, \cdots, n,$$

with $((X_1, Y_1), \cdots, (X_n, Y_n))$ i.i.d., $(X_t, Y_t) \sim \mathcal{N}(0_{\mathbb{R}^2}, 3 \cdot I_2)$,
$(\varepsilon_1, \cdots, \varepsilon_n)$ i.i.d., $\varepsilon_t \sim \mathcal{U}[-1, 1]$, the uniform law in $[-1; 1]$ and

$$\begin{array}{l} F_{\theta^0}(x, y) = \tanh(6 \cdot x - 2 \cdot y) + 2 \cdot \tanh(8 - x + 3 \cdot y) \\ -3 \cdot tanh(2 - 6 \cdot x - 2 \cdot y) + 1.5. \end{array} \tag{9}$$

Here, the true model is an MLP with 2 inputs, 3 hidden units and one output. In order to avoid too long time of computation, the number of hidden units is assumed to be between 1 and 10.

Let D be the size of the parameter vector (the dimension of the model or the number of weights of the MLP)

We will compare 3 criteria, from the least penalized (AIC like) to the most penalized (Very Strong Penalization), the following penalized criteria are assessed:

- AIC like: $\frac{1}{n}\sum_{t=1}^{n}\rho_{0.5}\left(z_t - F_\theta(x_t, y_t)\right) \times \left(1 + \frac{2}{n}\right)$
- BIC like: $\frac{1}{n}\sum_{t=1}^{n}\rho_{0.5}\left(z_t - F_\theta(x_t, y_t)\right) \times \left(1 + \frac{\log n}{n}\right)$
- SP (Strong Penalization): $\frac{1}{n}\sum_{t=1}^{n}\rho_{0.5}\left(z_t - F_\theta(x_t, y_t)\right) \times \left(1 + \frac{\sqrt{n}}{n}\right)$

We simulate $n = 100$, $n = 500$ and $n = 1000$ data according to the true model (9), for each n the experiment is repeated 100 times.

The following architectures are chosen by the penalized criteria :

- n=100

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**13**	**10**	**5**	**6**	**2**	**10**	**21**	**33**
BIC like	models sel.	0	**9**	**86**	**3**	0	**1**	0	0	0	**1**
SP	models sel.	**3**	**36**	**61**	0	0	0	0	0	0	0

- n=500

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**62**	**19**	**13**	**5**	**1**	0	0	0
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	**2**	**98**	0	0	0	0	0	0	0

- n=1000

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**72**	**13**	**7**	**6**	0	**2**	0	0
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0

The BIC like criterion and the Strong Penalization chose often the true architecture even for a small number of data. According to the theory, AIC like criterion is not consistent (see condition **A1**) and the chosen architecture is always too large. The Strong penalization chose a too small architecture when the number of data is small ($n = 100$), however it is a consistent criterion, so its behavior is correct for larger number of data ($n = 500$ and $n = 1000$). The BIC like criterion seems to be the best for this cost function.

4 Conclusion

The conventional least squares estimator may be seriously deficient in case of non-Gaussian errors. It seems reasonable to pay a small premium in the form of sacrified efficiency, in order to get more robust regression models. The class of statistics model called "regression quantiles" are known to have good properties under some restrictive assumptions. In this paper we have shown that some results may be obtained under more general assumptions. We have proven an inequality showing that overfitting of theses models is moderate if the noise admits exponential moments. This bound justifies the use of penalized criterion similar to the BIC criterion in order to fit the dimension of models. Finally, a more challenging task may be to get a more precise tuning of penalization term which, according to our result, can be chosen among a wide range of functions.

References

[1] Koenker, R. and Basset, G., Regression quantiles. *Econometrica*, 46:1, pages 33-50, 1978

[2] Engel, E., Die produktions- und Konsumptionverhaltnisse des Konigreichs Sachsen.*International Statistical Institute Bulletin*, 9, pages 1-125, 1857

[3] J. Rynkiewicz, Consistent estimation of the architecture of multilayer perceptrons. In M. Verleysen, editor, *proceedings of the* 14^{th} *European Symposium on Artificial Neural Networks* (ESANN 2006), d-side pub., pages 149-154, April 28-30, Bruges (Belgium), 2006.

[4] J. Rynkiewicz, General bound of overfitting for MLP regression models. *Neurocomputing* to appear.

[5] A.W. van der Vaart, *Asymptotic statistics*, Cambridge university Press, Cambridge, 1998.

[6] Weiss, A., Estimating nonlinear dynamic models using least absolute error estimation.*Econometric Theory*, 7, pages 46-68, 1991 *Econometrica*, 46:1, pages 33-50, 1978

Chapitre 3

Les délais de vente des VO. Modélisation et prédiction.

Connaitre la capacité d'absorption du marché constitue un dernier élément qu'il nous faut étudier pour disposer d'un modèle complet sur les VO. La complexité du marché de l'occasion confère aux délais de vente des VO une nature aléatoire qui justifie l'utilisation des méthodes statistiques pour leur étude. L'analyse conduite dans ce chapitre se propose de résoudre les problématiques suivantes :

(P0) segmenter les VO selon leur délai de vente.

(P1) modéliser et prédire les délais de vente d'un VO.

(P2) modéliser le nombre $N = N(T)$ de VO vendus à travers K points de vente pendant une période T.

3.1 Préparation et description des données

3.1.1 Préparation de la base de données

Soit $a^{(l)}_{(m_i,w_j)}$ une annonce publiée au mois m_i à partir d'un site w_i. Pour traiter les problématiques posées ci-dessus, nous constituons une base de données que nous appelons $\mathscr{A}$ et telle que $\mathscr{A} = \{a^{(l)}_{(m_i,w_j)} : l = 1,\ldots,L; i = 1,\ldots,I; j = 1,\ldots,J\}$. Cette structure de $\mathscr{A}$ fait qu'il est possible de trouver pour $l \neq l^{'}$, $i \neq i^{'}$ et $j \neq j^{'}$ au moins deux annonces $a^{(l)}_{(m_i,w_j)}$ et $a^{(l^{'})}_{(m^{'}_i,w^{'}_j)}$ qui seront associées à un même véhicule. Ainsi, pour que nos données soient exploitables pour notre étude, il nous faut dans un premier temps identifier parmi les annonces disponibles celles qui sont associées à un seul véhicule et repérer par la suite les dates D_l à laquelle chacune de ces annonces ont été publiées.

Notons $V_l\left(C^l_1,\ldots,C^l_k,KM^l,Prix^l,Age^l\right)$ le véhicule associé à l'annonce $a^{(l)}_{(m_i,w_j)}$ où $C^l_1,\ldots,C^l_k$ désignent ses caractéristiques invariantes. Pour identifier les annonces associées à un même véhicule, nous procédons par deux étapes (**Etape 1** et **Etape 2**) dans lesquelles nous proposons à chaque fois une règle d'identification des annonces.

Étape 1 $a^{(l)}_{(m_i,w_j)}$ et $a^{(l^{'})}_{(m_i,w_z)}$ désignent le même véhicule si pour chaque m_i, i fixé et $j \neq z$, nous avons $\left(C^l_1,\ldots,C^l_k\right) = \left(C^{l'}_1,\ldots,C^{l'}_k\right)$ et que $Km_l = Km_{l'}$ et $Px_l = Px_{l'}$.

Étape 2 $a^{(l)}_{(m_i,w_z)}$ et $a^{(l^{'})}_{(m_j,w_z)}$ désignent le même véhicule pour $i \neq j$ nous avons a $\left(C^l_1,\ldots,C^l_k\right) = \left(C^{l'}_1,\ldots,C^{l'}_k\right)$ et que $Km_l = Km_{l'} + dKm$ et $Px_l = Px_{l'} + dPx$.

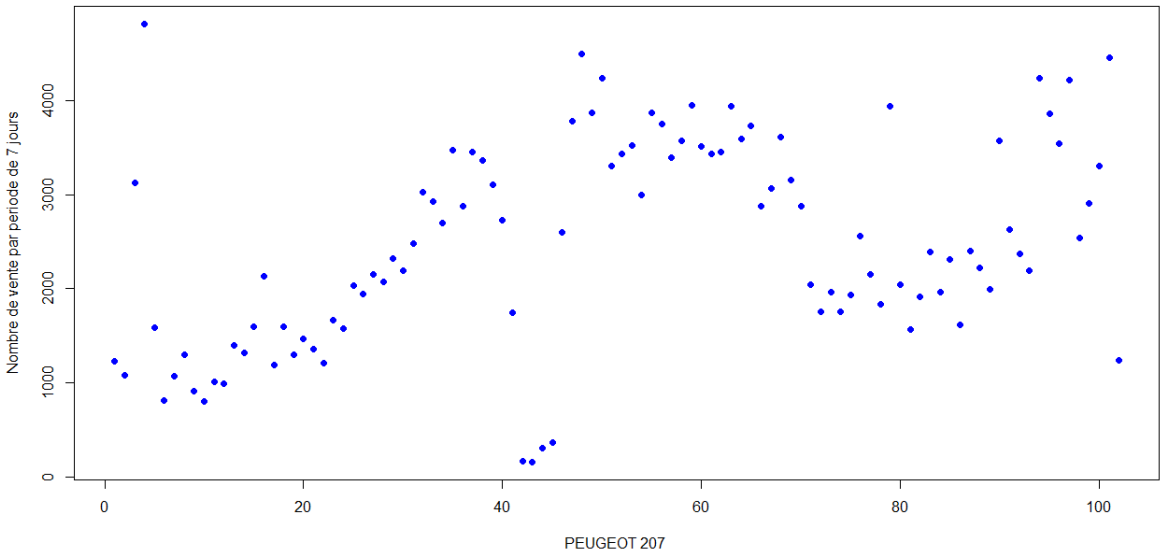

FIGURE 3.1 – NOMBRE DE VO VENDUS PAR PÉRIODE DE 7 JOURS. *Cette figure concerne la dispersion des délais de vente associés à la Peugeot 207 HDI Berline pour des annonces observées sur la période de janvier 2010 à février 2011. Nous voyons à travers ces figures qu'il y a des périodes où des observations se démarquent significativement des autres. Nous pouvons supposer que les dates correspondant à ces observations sont pour la plupart des dates de retrait après une longue publication de l'annonce et non des dates de vente. Nous avons supprimé ces observations.*

où $dKm \in [0;\alpha]$ et $dPx \in [0;\beta]$. α et β sont respectivement les variations des prix et des kilométrages admises et déterminées par les experts.

Lorsque nous aurons identifié à travers plusieurs annonces le même véhicule V_l , nous pourrons constituer un ensemble $\mathscr{D}_l = \{D_1, \ldots, D_{N_i}\}$ correspondant à l'ensemble des dates de mise en ligne de N_i de ces annonces.

Nous faisons les hypothèses suivantes :

HYPOTHÈSE 1. *La date de mise en vente D_d du VO correspond à la date à laquelle l'annonce a été aperçue pour la première fois dans un des sites w_j, soit $D_d = min\,(D_1, \ldots, D_{N_i})$*

HYPOTHÈSE 2. *La date de vente τ_f du VO correspond à la date à laquelle l'annonce a été aperçue pour la dernière fois, soit*

$$D_f = max\,(D_1, \ldots, D_{N_i}) \tag{3.1.1}$$

Ce travail se positionne plus dans une logique commerciale que mathématique. De ce fait, nous nous sommes permis de considérer une approche hypothétique et non rigoureuse dans la définition des variables liées aux délais de vente des VO. Nous ne tenons donc pas compte de la possibilité que le véhicule soit retiré du marché sans qu'il ne soit vendu. Toutefois, dans la démarche expérimentale, il nous est amené de supprimer certaines observations suspectes telles qu'elles apparaissent en FIG.3.1.

3.1.2 Définition des variables d'intérêt

Définition 3.1.1. *Sous les hypothèses H.1 et H.2, le délai de vente τ du VO est défini par*

$$\tau = D_f - D_d \tag{3.1.2}$$

et correspond à la période (en nombre de jours) de publication de l'annonce.

Nous retenons également la date à laquelle le prix a été modifié pour la dernière fois, soit D_{modif} cette date.

Définition 3.1.2. *Le temps de vente ajusté du VO, notée τ_a se calcule par $\tau_a = D_f - D_{modif}$.*

Définition 3.1.3. *L'âge considéré dans l'étude, que nous appelons $AgFin$, est l'âge en mois du véhicule en date de fin d'observation*

$$AgFin = Ag + \tau \tag{3.1.3}$$

Définition 3.1.4. *La variable $ClassAg$ est une variable catégorielle telle que*

$$ClassAg = \begin{cases} 1 & si\ Age < 36\ mois \\ 2 & si\ 36\ mois \leq Age < 60\ mois \\ 3 & Age\ \geq 60\ mois \end{cases} \tag{3.1.4}$$

La définition des modalités s'est faite en s'appuyant sur l'analyse des kilométrages.

Pour chaque véhicule présent dans notre base de données, nous récupérons le prix prédit ($CoteAutobiz$) par le modèle mis production explicité en chapitre 2.

Définition 3.1.5. *La variable r_0 correspond à un écart relatif du prix prédit par rapport au prix observé. Elle est donnée par*

$$r_0 = \frac{Prix - CoteAutobiz}{Prix} \tag{3.1.5}$$

Remarque 3.1.1. *Nous avons $(r_0 \leq 0)$ ou $(r_0 \geq 0)$ selon le cas ou le prix annoncé est plus petit ou plus grand que le prix prédit.*

Deux variables indicatrices sont liées au prix du véhicule.

Définition 3.1.6. *La variable r_0ind donne la position du prix de l'annonce pour le véhicule par rapport à la valeur de marché et est définie par*

$$r_0ind = \begin{cases} 1 & si\ r_0 > 0 \\ 0 & sinon \end{cases} \tag{3.1.6}$$

Définition 3.1.7. *La variable $modif$ précise si le prix d'annonce du véhicule a été modifié au moins une fois pendant toute la période de publication*

$$modif = \begin{cases} 1 & si\ D_{modif} \neq D_d \\ 0 & sinon \end{cases} \tag{3.1.7}$$

3.2 Caractérisation des délais de vente

Soit τ_i le délai de vente observé pour un véhicule i, $i = 1, \ldots, L$.

hypothèse 3. *$(\tau_1, \tau_2, \ldots, \tau_{L-1}, \tau_L)$ correspondent à L réalisations indépendantes de la variable aléatoire $\mathscr{T}$ et que pour tout $i \in \{1, \ldots, L\}$, $Prob(\mathscr{T} \leq \tau_i) = F_{\mathscr{T}}(\tau_i)$.*

3.2.1 Échelle d'observation de la variabilité des délais de vente

Une grande dispersion dans les valeurs observées pour les délais de vente que nous pouvons voir en FIG.3.2 pourrait indiquer une fraction de variabilité qui n'est pas aléatoire et résultant donc d'un mécanisme déterministe qui laisse à supposer que : "*certains segments de véhicules ayant des caractéristiques bien définies présentent des délais de vente plus longs que d'autres*". Il est alors important de choisir à quelle échelle les délais de vente doivent être analysés pour que cette variabilité puisse être réduite. Pour cela, nous testons si pour un niveau de facteur tel que la marque du véhicule par exemple, nous pouvons observer un délai de vente moyen différent pour chaque modalité. Cela revient à tester

$$H_0 : \mu_{Mq_1} = \ldots = \mu_{Mq_J} \text{ contre } H_1 : \exists i, l, i \neq l, \text{ t. q } \mu_{Mq_i} \neq \mu_{Mq_l}$$

où μ_{Mq_j} désigne le délai de vente moyen associé aux véhicules de la marque Mq_J. La démarche correspond à un modèle d'analyse de la variance (ANOVA) et nous pouvons nous référer à [AB12] pour plus de détails. La FIG.3.3 nous montrent quelques valeurs estimées à partir de notre base de données des délais moyens pour différentes marques et différentes familles au sein d'une même marque. En accord avec les experts et dans le souci d'une facilité de mise en production, nous fixons notre maille à une association hiérarchique **"Famille - Carrosserie - Energie "** et si Fam, $Enrg$ et $Carsr$ désignent respectivement l'ensemble des modalités des différentes familles, type d'énergie et type de carrosserie disponibles dans notre base de données $\mathscr{A}$, alors le nombre K de mailles constituées sera

$$K = |Fam| \times |Enrg| \times |Carsr|$$

et $\mathscr{A}$ peut s'écrire telle que

$$\mathscr{A} = \{\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_K\}$$

3.2.2 Segmentation du marché VO selon des profils de des délais de vente

Afin de fournir une connaissance satisfaisante du marché VO, la finalité commerciale dans laquelle s'inscrit ce travail nous oblige à établir des profils types des délais de vente et donc de définir des indicateurs permettant de résumer l'hétérogénéité de ce marché VO. L'idée consiste à regrouper des véhicules présentant des caractéristiques communes dans leur délai de vente. Soit $F_{m_k} = P\left(\mathscr{T}_{m_k} < q_{m_k}\right)$ la distribution des délais de vente observés au sein d'une maille m_k. Nous résumons cette distribution F_{m_k} à partir de quatre indicateurs associés à $u = \{0.20, 0.40, 0.60, 0.80\}$ et pour chaque u, nous déterminons q_{m_k} tel que $P\left(\tau_{m_k} < q_{m_k}(u)\right) \geq u$. Des indicateurs de référence $I_R(u)$ sont définis par la suite en calculant la moyenne des q_{m_k} sur toutes les K mailles, soit

$$I_R(u) = \frac{1}{K} \sum_{i=1}^{K} q_{m_i}(u) \tag{3.2.1}$$

Nous obtenons cinq profils-type de véhicules selon leur délai de vente. Il est alors possible par la suite de définir à des niveaux d'analyses intermédiaires tel que le type d'énergie, le type de carrosserie, la marque *ect* ... des indicateurs par une formule analogue à (3.1). Le calcul sur la base de données que nous avons utilisée pour la démarche expérimentale a produit les différents indicateurs que nous pouvons voir en TAB.3.1 pour ce qui est des indicateurs de réference et

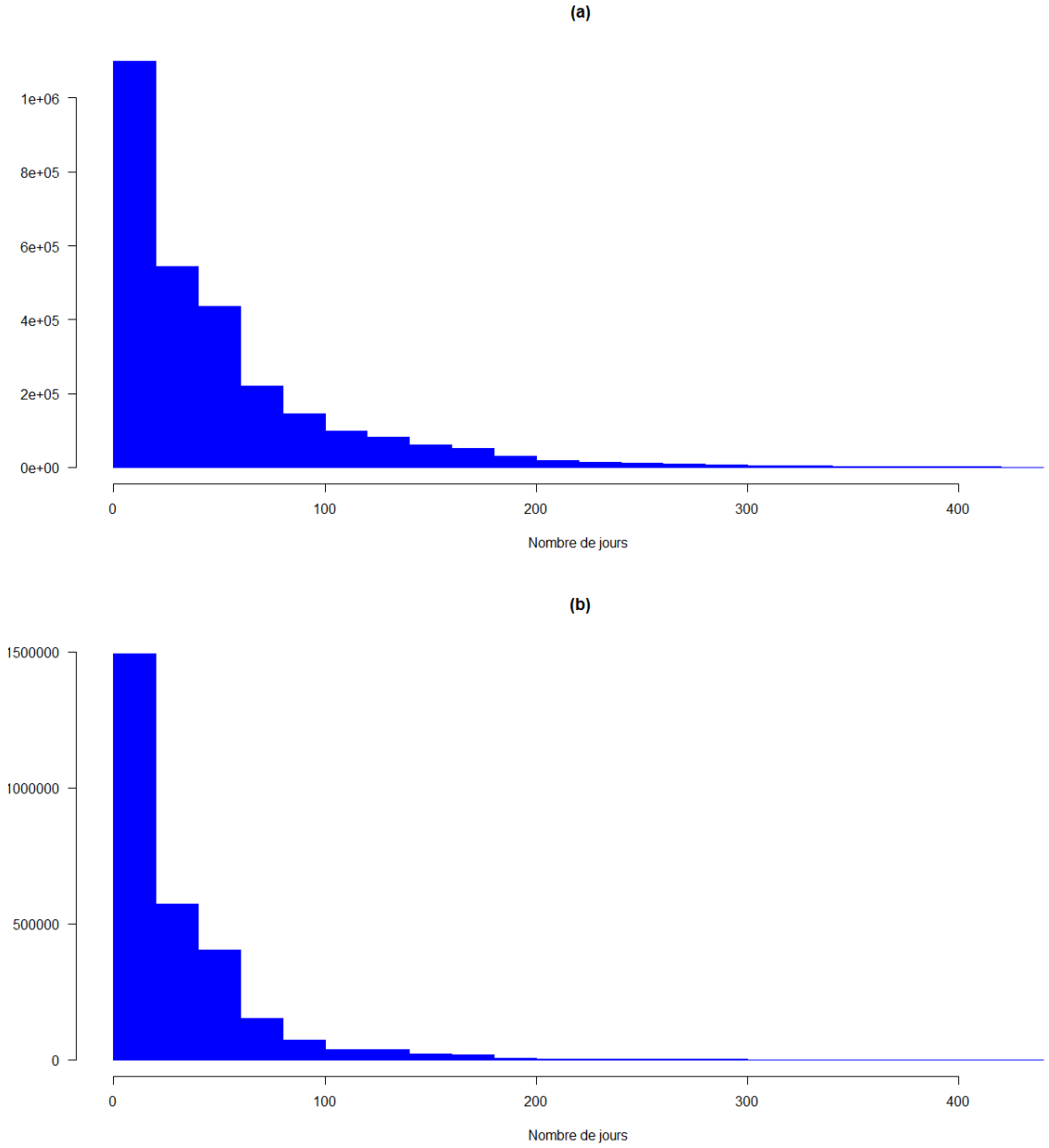

FIGURE 3.2 – DISTRIBUTION DES DÉLAIS DE VENTE. ***(a)*** *représente les délais de vente observés pour 2 839 069 véhicules vendus sur la période de janvier 2010 à février 2011 et* ***(b)*** *le délai ajusté, c'est-à- dire après une dernière modification de prix de vente. Ces figures nous montrent bien-sûr que certains VO se vendent plus rapidement que d'autres. Le délai de vente moyen respectif pour chaque cas considéré est de* ***52 jours*** *pour* ***(a)*** *et* ***18 jours*** *pour (b) avec* ***écart-type*** *respectif de* ***67*** *et* ***44*** *jours.*

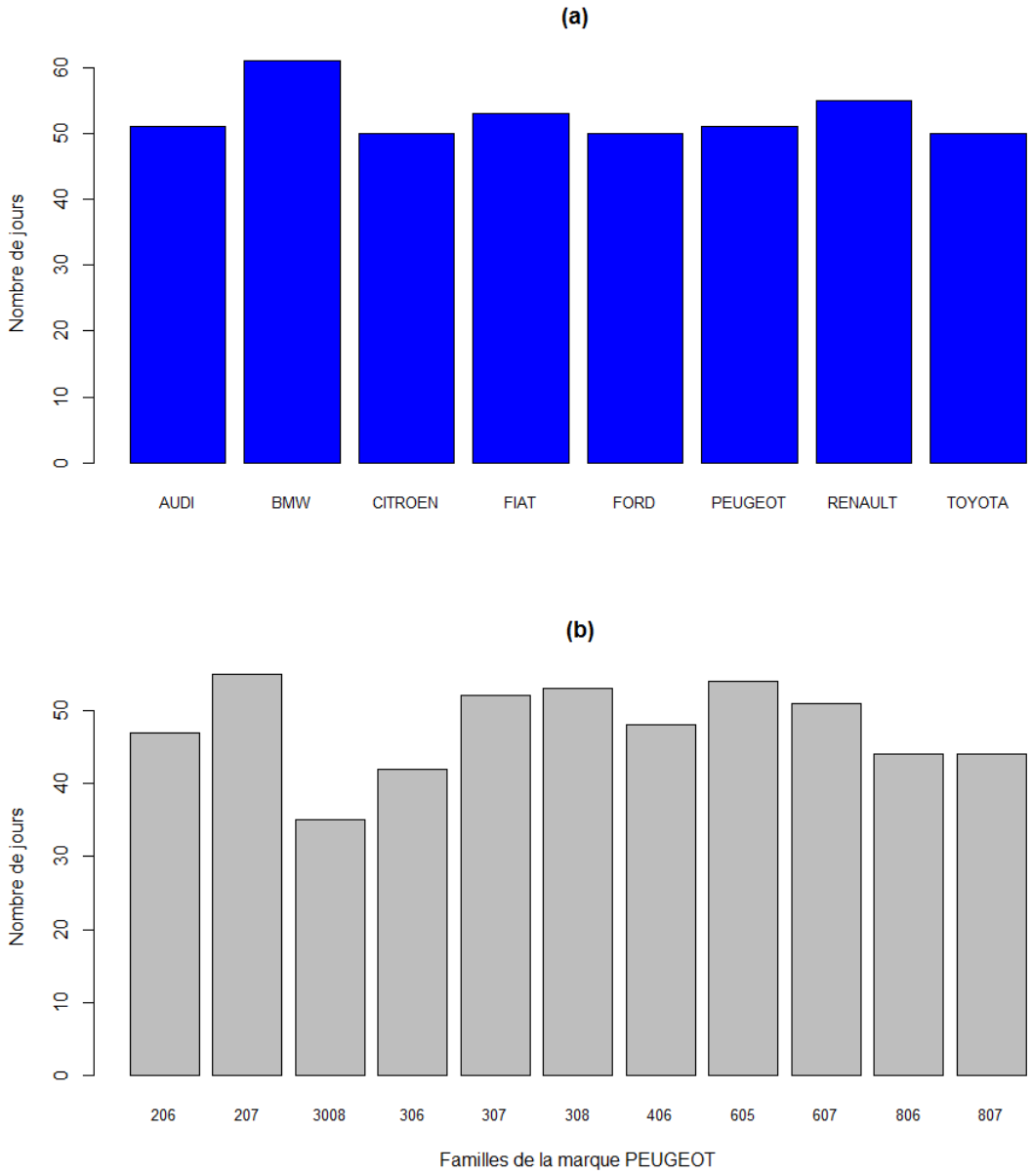

FIGURE 3.3 – DÉLAI DE VENTE MOYEN. *Ces figures nous témoignent du délai de vente moyen par marque* ***(a)*** *et par famille au sein d'un marque* ***(b)***.

u	$I_\tau(u)$	délais de vente	Profil
0.20	11	inférieur à 12 jours	r1 : très rapide
0.40	24	entre 12 et 24 jours	r2 : rapide
0.60	44	entre 25 et 44 jours	r3 : moyen
0.80	75	entre 45 et 75 jours	r4 : lent
		supérieur à 75	r5 : très lent

TABLE 3.1 – PROFILS DE RÉFÉRENCE POUR LES DÉLAIS DE VENTE DE VO.

Marque	$I_\tau(0.20)$	$I_\tau(0.40)$	$I_\tau(0.60)$	$I_\tau(0.80)$
Citroen	10	22	43	76
Ford	10	23	45	76
Opel	9	20	39	62
Peugeot	12	27	52	91
Renault	10	23	46	81
Volkswagen	8	19	37	61

Carrosserie	$I_\tau(0.20)$	$I_\tau(0.40)$	$I_\tau(0.60)$	$I_\tau(0.80)$
Berline	10	23	45	79
Break	12	27	52	90
SUV	8	18	37	64
7	10	23	45	73
9	10	23	43	69
Monospace	9	22	41	72
VO de Société	9	20	41	69

Energie	$I_R(0.20)$	$I_\tau(0.40)$	$I_\tau(0.60)$	$I_\tau(0.80)$
Essence	9	20	39	65
Diesel	10	24	47	82

TABLE 3.2 – INDICATEURS PAR CATÉGORIE. *Ces chiffres correspondent aux indicateurs calculés en nombre de jours pour les marques, carrosseries et types d'énergie. Ces indicateurs permettent pour chaque catégorie de définir le profil de vente associé aux VO.*

TAB.3.2 pour les indicateurs associés à la marque, au type de carrosserie et type d'énergie que nous avons également comparés aux indicateurs de référence tel qu'ils sont présentés en FIG.3.1.

La pertinence et l'utilité effective des indicateurs de référence que nous avons définis ont été mises en évidence en répondant à la question posée en 3.2.1.

Question 3.2.1. Y-a-t-il une différence significative entre le profil de référence et le profil défini pour les différentes catégories ?

La réponse à Q.3.2.1 consiste à repérer s'il existe des écarts de distribution significatifs de délais de vente entre le profil obtenu par les indicateurs de référence et le profil obtenu par les indicateurs associés aux différentes catégories. En effet, si les profils des délais de vente associés à ces derniers sont similaires à ceux associés aux indicateurs de référence, nous devrions avoir une même proportion de VO pour chaque profil tel que nous le présentons en TAB.3.3 et ceci sera validé par un un test statistique de χ^2 [Man86] [Kri96] sur le même échantillon pour chaque cas considéré en proposant pour tout $u = \{0.20, 0.40, 0.60, 0.80\}$ les différentes hypothèses suivantes :

$$H_0 : n_{R_1}(e) = n_{E_1}(e), \ldots, n_{R_5}(e) = n_{E_5}(e)$$

contre

H_1 : au moins une des égalités n'est pas vérifiée

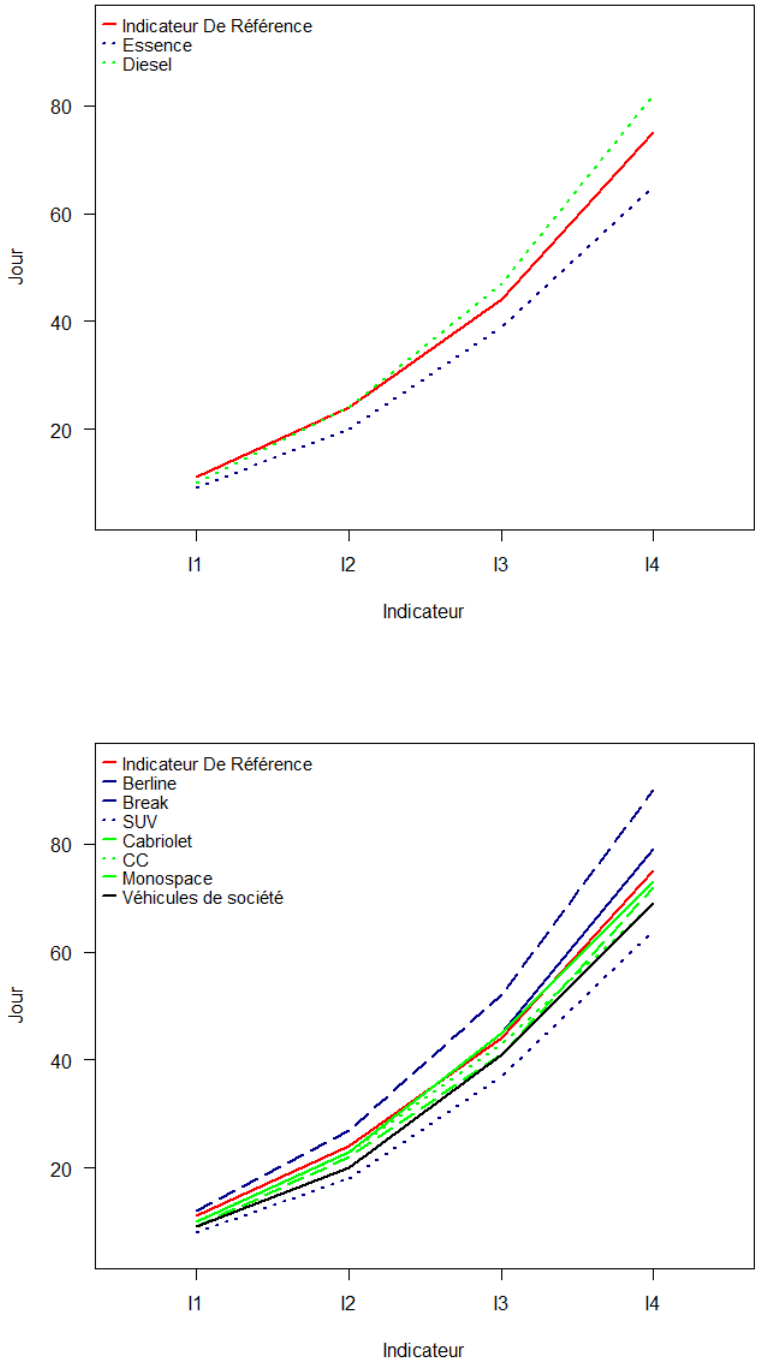

FIGURE 3.4 – COMPARAISON AVEC LES INDICATEURS DE PROFIL DE RÉFÉRENCE $I_R(u)$. *Ces figures ont l'avantage de montrer simplement la forme des différents profils en proportions cumulées par rapport à la référence du marché pour les différentes marques, différents types de carburant et les différents types de carrosserie. Dans les trois figures, la courbe en trait rouge continu représente le profil de référence $I_R(u)$. Ainsi, si une courbe est située en dessus de celle-ci, c'est que la marque ou le type de carburant ou le type de carrosserie en question* ***"n'est pas bon"*** *en terme de distribution du temps de vente. Autrement dit, ce sont des catégories de véhicules qui se vendent plus lentement par rapport à la moyenne. Par ailleurs, les VO essence se vendent en des délais plus courts que les VO Diesel.*

	$r1$	$r2$	$r3$	$r4$	$r5$
$I_E(u)$	n_{E1}	n_{E2}	...	...	n_{E5}
$I_R(u)$	n_{R1}	n_{R2}	...	...	n_{R5}

TABLE 3.3 – COMPARAISON DES EFFECTIFS. *Ce tableau donne la forme de comparaison des effectifs dans chaque profil de délai de vente obtenus par les indicateurs de référence ($I_R(u)$) et par les indicateurs associés à la marque, au type de carburant et au type de carrosserie ($I_E(u)$) . n_{Ei} correspond à l'effectif de VO appartenant au profil de rotation ri obtenu par les indicateurs associés à chacun des échantillon E et n_{Ri} correspond à l'effectif obtenu en utilisant les indicateurs de références avec $i = \{$ r1=très rapide , r2=rapide, r3=moyen, r4=lent, r5=très lent $\}$. Si les différents indicateurs obtenus par marque, par type de carburant et par type de carrosserie étaient similaires à l'indicateur de référence, on devrait trouver à chaque fois pour tout i, $n_{Ei} = n_{Ri}$.*

pour la statistique de test d^2 telle que

$$d^2 = \sum_{i=1}^{5} \frac{(n_{R_i} - n_{E_i})^2}{n_{E_i}} \tag{3.2.2}$$

Sous l'hypothèse H_0, d^2 est une réalisation d'une variable aléatoire D^2 suivant asymptotiquement une loi du χ^2 à 4 degrés de liberté. Avec un seuil $\alpha = 0.05$, l'hypothèse nulle devrait être rejetée pour des valeurs de d^2 supérieures à $Q_{\chi^2(4)}(0.95) = 9.49$. Les résultats sont reportés en TAB.3.4.

3.3 Prédiction par un modèle de régression linéaire

Nous avons introduit dans les sections précédentes une caractérisation du délai de vente, une quantification qui nous a permis de définir des profils de VO selon leur délai de vente et d'identifier dans une certaine mesure les facteurs expliquant la variabilité de ce délai de vente.

Le problème traité ici est celui de la prédiction du délai de vente à partir des variables quantitatives.

3.3.1 Décomposition du problème de prédiction

Nous supposons que pour une maille m_k donnée (une association Marque - Famille - Énergie - Carrosserie) le délai de vente est fonction des variables caractéristiques évoluant dans le temps du véhicule et nous écrivons :

$$\tau_{m_k} = f_{m_k}(V_{age}, V_{km}, V_{prix}) + \varepsilon_k \tag{3.3.1}$$

où ε_k désigne les conditions de vente, les effets de conjoncture, les erreurs de saisie ainsi que les effets des autres facteurs que nous ne connaissons pas et V_{age}, V_{km} et V_{prix} sont respectivement des variables liées à l'âge, au kilométrage et au prix du véhicule. Par ailleurs, nous supposons

Citroen	r1	r2	r3	r4	r5
I_R	120 152	87 244	83 007	84 088	92 096
I_E	92 784	99 424	88 261	94 022	92 096

Ford	r1	r2	r3	r4	r5
I_R	51 632	36 580	35 812	39 141	40 054
I_E	40 260	41 494	40 977	40 434	40 054

Opel	r1	r2	r3	r4	r5
I_R	55 734	36 559	33 781	38 456	29 843
I_E	33 483	44 666	39 723	38 324	38 177

Renault	r1	r2	r3	r4	r5
I_R	141 847	117 842	122 072	122 850	170 372
I_E	131 107	142 181	134 009	134 417	133 269

Peugeot	r1	r2	r3	r4	r5
I_R	257 205	171 751	168 061	182 942	210 960
I_E	181 974	216 761	198 282	195 960	197 942

Volkswagen	r1	r2	r3	r4	r5
I_R	94 965	58 551	53 538	57 539	44 395
I_E	58 248	68 385	58 939	63 219	60 197

Berline	r1	r2	r3	r4	r5
I_R	601 323	416 806	403 791	424 587	477 090
I_E	421 040	523 881	462 462	455 787	460 427

Break	r1	r2	r3	r4	r5
I_R	54 922	45 920	47 165	50 137	65 255
I_E	50 628	55 612	51 916	53 035	52 208

Cabriolet	r1	r2	r3	r4	r5
I_R	332	201	178	192	181
I_E	200	249	212	211	212

Coupé	r1	r2	r3	r4	r5
I_R	24 948	17 865	17 944	20 813	18 932
I_E	19 315	21 525	19 917	19 692	20 053

SUV	r1	r2	r3	r4	r5
I_R	19 645	14 711	15 223	16 239	13 873
I_E	15 153	17 536	15 672	15 556	15 774

Monospace	r1	r2	r3	r4	r5
I_R	829	541	573	536	563
I_E	593	627	617	598	607

Veh.Sociétés	r1	r2	r3	r4	r5
I_R	19 536	12 483	11 397	12 512	11 826
I_E	11 869	15 336	13 905	13 122	13 522

Essence	r1	r2	r3	r4	r5
I_R	134 060	94 078	88 285	90 251	78 512
I_E	93 039	104 588	93 489	97 353	96 717

Diesel	r1	r2	r3	r4	r5
I_R	587 475	414 449	407 986	434 765	509 208
I_E	458 814	496 343	468 399	462 097	468 230

TABLE 3.4 – LES EFFECTIFS ASSOCIÉS AUX DIFFÉRENTES CATÉGORIES. *Les valeurs de la statistique d^2 calculée pour chacune des catégories ci dessus sont largement supérieures à $Q_{\chi^2(4)}$. Nous pouvons nous permettre de dire qu'il existe une différence significative entre les indicateurs de référence et ceux calculées individuellement pour les marques, les types de carburant et les types de carrosserie.*

que ces ε_k sont mutuellement indépendants et identiquement distribués selon une loi $\mathscr{D}$ et que f_{m_k} est une combinaison linéaire des variables V_{age}, V_{km} et V_{prix}.

Nous estimons la fonction f_{m_k} par apprentissage statistique de son comportement. La fonction estimée $\widehat{f}_{m_k}$ est choisie dans un espace $\mathscr{F}$ suffisamment riche pour que $\widehat{f}_{m_k}$ puisse traduire de façon la plus fidèle possible les caractéristiques du phénomène étudié sur la base des données utilisées. Cette estimation consiste à minimiser des critères de prédiction $L\left(\widehat{f}_{m_k}\right)$.

En proposant une forme linéaire pour f_{m_k} et l'Eq.(3.3.1) devient

$$\tau_{m_k} = \beta_0 + \beta_{1,k} V_{age} + \beta_{2,k} V_{km} + \beta_{3,k} V_{prix} + \varepsilon_k \tag{3.3.2}$$

et l'idée étant donc d'estimer les β_i. le Les méthodes d'ajustement sont nombreuses et ici, nous utilisons la méthode des MCO [Gro03]. Particulièrement, nous souhaitons que pour tout $i \in \{0,1,2,3\}$ on ait $\beta_i \neq 0$.

3.3.2 Exploration de l'ensemble des modèles candidats. Choix du modèle

Nous disposons de quatre groupes de variables explicatives :
- $V_{age} = \{Ag, \log(Ag)\}$ (variables liées à l'âge.)
- $V_{km} = \left\{Km, \log(Km), \sqrt{Km}, \sqrt[3]{Km}\right\}$(variables liées au kilométrage.)
- $V_{prix} = \left\{Px, \log(Px), \sqrt{Px}\right\}$ (variables liées au prix.)
- $V_r = \{r_0, r_0ind, modif\}$ (variables indicatrices contenant des informations liées au prix, au Km et aussi à l'âge.)

La définition des éléments de chaque groupe s'est basée sur des études antérieures, notamment pour la prédiction du prix.

Chaque modèle candidat $M_1, \ldots, M_k$ utilise un élément des variables explicatives V_{age} , V_{km}, V_{prix} et les variables r_0, r_0ind et $modif$ que nous avons défini en début de ce chapitre. Les différents modèles possibles pour une maille m_k correspondent au produit cartésien

$$\begin{pmatrix} Ag \\ \log(Ag) \end{pmatrix} \times \begin{pmatrix} Km \\ \log(Km) \\ \sqrt{Km} \\ \sqrt[3]{Km} \end{pmatrix} \times \begin{pmatrix} Px \\ \log(Px) \\ \sqrt{Px} \end{pmatrix} \times r_0 \times r_0ind \times modif \tag{3.3.3}$$

dont le nombre sera

$$J = \text{card}(V_{age}) \times \text{card}(V_{km}) \times \text{card}(V_{prix}).$$

Le choix du modèle qui sera associé à la maille m_k se fait en deux temps :

Etape 1 pour chacun des différents modèles possibles $M_1, \ldots, M_J$ nous effectuons une sélection pas à pas (ou stepwise selection) [RPD98] [WJ03] pour déterminer parmi les variables explicatives composant chaque modèle proposé celles pour lesquelles le critère BIC [Kad11] est le plus faible. Ce sera le modèle retenu, soit M_R.

Etape 2 Nous renouvelons l'Etape 1 N fois sur des échantillons obtenus par tirage aléatoire avec remise à partir de la maille m_k. Nous obtenons alors N_0 ($N_0 \leq N$) modèles $M_{R_{N_1}}, \ldots, M_{R_{N_0}}$ qui correspondent au modèle retenu à la fin de chaque expérience.

Le modèle choisi pour la maille m_k sera celle qui aura la plus grande occurrence d'apparition [BASH06], c'est à dire

$$M_{opt}(m_k) = \max \sum_{i=1}^{N} 1_{\{\text{modèle retenu}=M\}}. \tag{3.3.4}$$

3.3.3 Les résultats associés aux données choisies pour l'étude

A partir d'une concaténation de toutes les annonces mises en ligne entre le 03 janvier 2010 et le 17 février 2012 nous constituons une base de données que nous traitons de façon à les rendre exploitables pour notre étude. De restrictions ont été appliquées lors de la constitution du fichier et la base de données finale contient les six marques figurant parmi les plus présentes sur le marché de l'occasion et qui sont Citroën, Ford, Opel, Peugeot, Renault et Volkswagen ; deux types d'énergie, Diesel et Essence, et sept types de carrosserie - lorsqu'ils sont associés aux marques et types de carburant cités précédemment - qui sont Berline, Break, SUV, Compact, Coupé, Coupé-Cabriolet et Véhicule de société. Dans cette étude, nous testons donc $J = 24$ modèles différents sur chacune des mailles $m_k : k = 1, \ldots, K$.

Nous séparons la base de données en apprentissage 70% $\mathscr{L}_A$ et base test 30% $\mathscr{L}_T$. Nous calculons les coefficients de la régression et analysons la validité par les critères usuels. Nous appliquons les étapes **Etape 1** et **Etape 2** décrites précédemment. Un résultat pour une maille particulière est reporté en TAB.3.5.

L'interprétation des coefficients ici n'a pas vraiment d'utilité pour notre étude. Nous nous intéressons plutôt à la prédiction du délai de vente par le modèle. Notons qu'un modèle prédictif n'est utile que s'il décrit les nouvelles données avec une précision suffisante. Nous évaluons la capacité de prédiction du modèle à prédire le bon profil pour le véhicule et les précisions de la prédiction pour la maille prise en exemple sont reportées en TAB.3.6.

REMARQUE 3.3.1. *Le temps de calcul moyen n'est pas pris en compte.*

3.4 Modélisation du nombre de véhicules vendus dans un intervalle de temps T

Nous nous proposons de modéliser le nombre n de VO similaires (associés à une maille) vendus dans un intervalle de temps T. Si nous notons d_i la date à laquelle pour le site i la vente débute, nous supposons dans un premier temps que pour tout $i = 1, \ldots, K$: $\quad d_i = T_0 = 0$.

3.4.1 Le modèle proposé

Soit $n = n(T)$ le nombre total de VO vendus à partir de K sites pendant une période T et soit n_i le nombre de VO vendus à partir du site i, nous avons

$$n = \sum_{i=1}^{K} n_i. \tag{3.4.1}$$

Nous disposons des $n(T_1), \ldots, n(T_J)$ obtenus sur les périodes $T_1, \ldots, T_J$.

HYPOTHÈSE 4. *Le nombre de VO vendus dans la période T est la réalisation d'une variable aléatoire N d'espérance finie $\mathbb{E}(N) = \lambda$.*

Cette hypothèse H.4 ouvre la voie à une première considération. En effet, la loi usuelle pour modéliser la distribution du nombre d'évènements dans un intervalle de temps est la loi de Poisson $\mathscr{P}$ telle que pour tout $n \in \mathbb{N}^* : p(n) = \mathbb{P}(N = n) = \mathrm{e}^{-\lambda}\dfrac{\lambda^n}{n!} \quad , \lambda > 0$ où $\mathbb{E}(N) = \mathbb{V}ar(N) = \lambda$.

Source	DF	Sum of Squares	Mean Square	F Value	$Pr > F$
Model	6	705003689	117500615	1434501	<.0001
Error	190321	15589273	81.91042		
Corrected Total	190327	720592962			

Root MSE	9.05044	R-Square	0.9784
Dependent Mean	60.35152	Adj R-Sq	0.9784
Coeff Var	14.99620		

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr > \|t\|$
Intercept	1	-341.89003	2.07554	-164.72	< .0001
Ag	1	0.98781	0.00037890	2607.05	< .0001
$\log(prix)$	1	35.70973	0.21443	166.54	< .0001
$\sqrt{km}$	1	-0.07906	0.00050300	-157.18	< .0001
r_0	1	-28.23579	0.35881	-78.69	< .0001
$r_0 ind$	1	-0.87088	0.06336	-13.75	0.0494
$modif$	1	1.90731	0.04718	40.43	< .0001

TABLE 3.5 – ESTIMATION DU MODÈLE DE RÉGRESSION. *Ces résultats concernent N=190 328 données de la Peugeot 207 HDI Berline. Le modèle est valide, c'est-à-dire l'hypothèse selon laquelle tous les coefficients sont nuls, est rejetée. De plus, le pouvoir explicatif du modèle exprimé par le* $R^2 \simeq 0.98$ *confirme la pertinence des variables retenues. Tous les coefficients son significatifs au seuil de* 0.05*. La constante en principe, correspond à une valeur moyenne prise par* τ *lorsque tous les autres paramètres sont nuls. La variable associée au prix ne pouvant être nulle, ce coefficient réajuste donc la valeur moyenne pour une valeur minimum du prix. Donc, pris individuellement ici, aussi bien le signe que la valeur de la constante (Intercept) n'ont aucun sens.*

Profil prédit / Profil observé	"très rapide"	"rapide"	"moyen"	"lent"	"très lent"
"très rapide"	68,54%	25,18%	6,25%	0,03%	0,00%
"rapide"	26,75%	43,86%	28,36%	1,03%	0,00%
"moyen"	1,14%	16,51%	64,87%	17,44%	0,04%
"lent"	0,00%	0,05%	11,09%	80,27%	8,59%
"très lent"	0,00%	0,00%	0,00%	5,14%	94,86%

TABLE 3.6 – ÉVALUATION DE LA PRÉDICTION SUR LE PROFIL POUR LA MAILLE m_k. *Nous avons ici la proportion des VO dont les délais de vente prédits sont répartis selon leur profil d'appartenance. Nous constatons que dans plus de 73% des cas, les délais prédits se retrouvent dans le bon profil de rotation.*

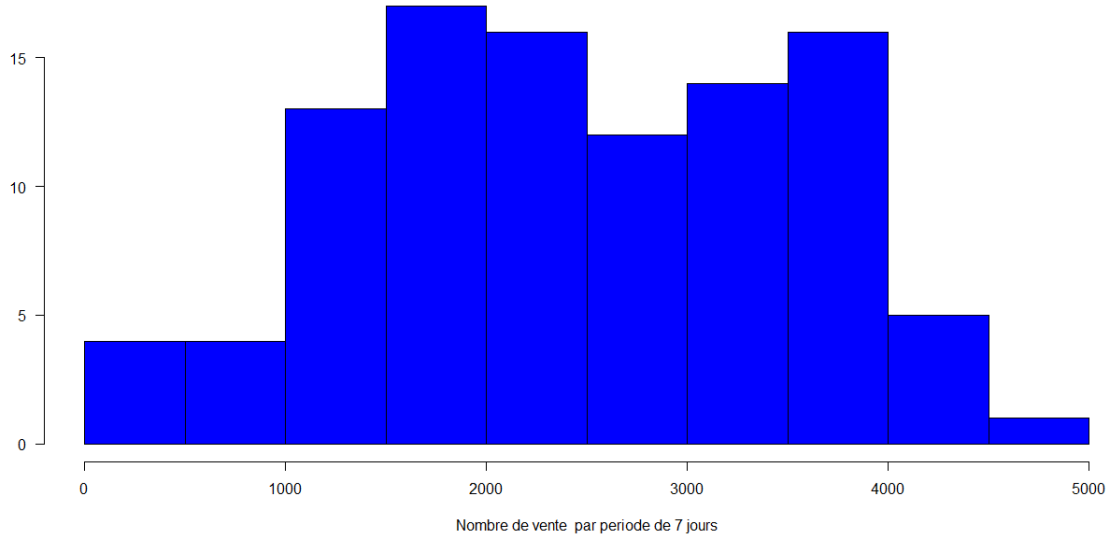

FIGURE 3.5 – NOMBRE DE VENTES PAR PÉRIODE $T = 7$ JOURS .*Cette figure représente le nombre de ventes observées sur la période du 01 Janvier 2010 au 16 mars 2012 pour la Peugeot 207 HDI Berline par période de 7 jours. Le moyenne est de* $\mu_{P.207} = 2467$ *et l'écart-type est* $\sigma_{P.207} = 1086.72$.

Comme nous le voyons en FIG.3.5 le nombre de VO vendus dans une période de $T = 7$ jours fait apparaitre une variance plus grande que la moyenne. Dans de telles circonstances, nous pouvons supposer que le paramètre λ est soumis à une variabilité inhérente [Law80] que nous pouvons associer, dans un premier temps aux sites de vente *(exemple : régions administratives du territoire national français)*. Nous pouvons observer en FIG.3.6 le nombre de vente moyen par période de 7 jours observé à travers les différents sites de vente. Dans un cas pratique et plus réaliste, les sites peuvent être associés à des points de vente, des sites web ou des concessionnaires, qui appartiennent ou pas à un même groupe de distribution (exemple : PGA Motors).

REMARQUE 3.4.1. *Bien qu'en réalité, la dynamique du marché est beaucoup plus complexe, la variabilité liée à d'autres composantes comme les conditions de vente (exemple : la saison ou le mois de vente) ne sera pas considérée dans cette étude et sera laissée à un travail futur.*

L'hypothèse H.4 n'offre donc qu'une utilité limitée pour notre problème. La distribution binomiale négative apparait comme une alternative à la loi de Poisson lorsque λ correspond à une réalisation d'une variable aléatoire Λ [Fri80] [BGP96] [Cox55] [CI80b] [JKK92].

HYPOTHÈSE 5. $\lambda = \sum_{i=1}^{K} \lambda_i$ *où chaque* λ_i *est une réalisation d'une variable aléatoire distribuée selon une loi Gamma de paramètres* (α, β).

Propriété 3.4.1. *[JKK92] Si* $\Lambda \equiv Ga(\alpha K, \beta)$ *et* $(N \mid \Lambda = \lambda) \equiv \mathscr{P}(\lambda)$ *alors la variable aléatoire* N *décrivant le nombre de VO vendus dans la période* T *suit une loi binomiale négative* $\mathscr{BN}\left(r = \alpha K, p = \dfrac{\beta}{\beta + 1}\right)$ *telle que*

	Région		Région		Région
s1	Alsace	s8	Champagne-Ardenne	s15	Midi-Pyrenees
s2	Aquitaine	s9	Franche-comte	s16	Nord-Pas-De-Calais
s3	Auvergne	s10	Haute-Normandie	s17	Paca
s4	Basse-Normandie	s11	Ile-De-France	s18	Pays De la Loire
s5	Bourgogne	s12	Languedoc-Roussillon	s19	Picardie
s6	Bretagne	s13	Limousin	s20	Poitou-Charentes
s7	Centre	s14	Lorraine	s21	Rhone-Alpes

TABLE 3.7 – LES DIFFÉRENTS SITES DE VENTE CONSIDÉRÉS. *Ce tableau nous donne la correspondance des libellés pour les sites de vente indiqués en FIG.3.6.*

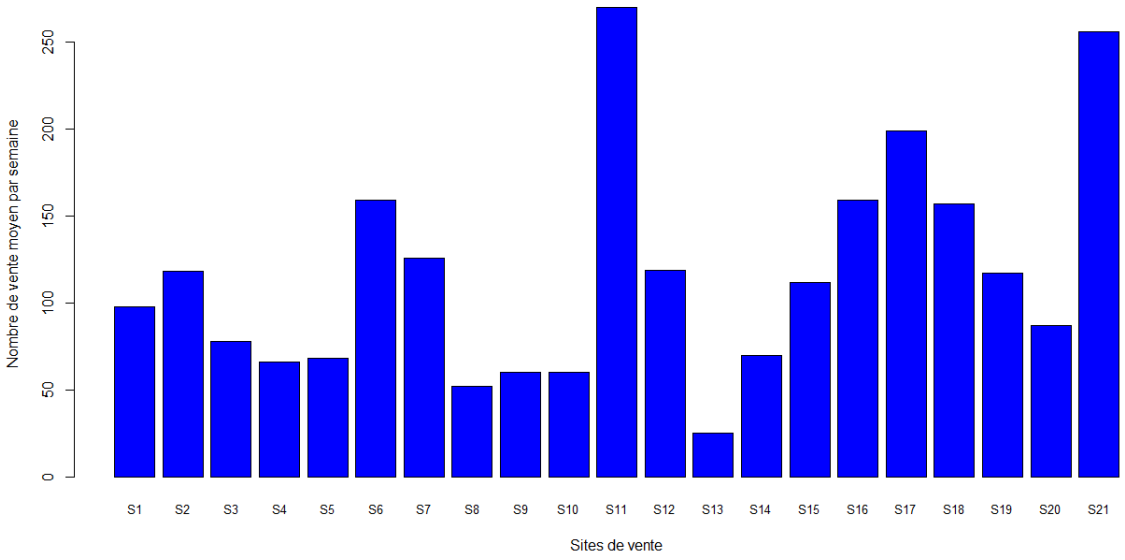

FIGURE 3.6 – NOMBRE DE VENTE PAR RÉGION POUR LA PEUGEOT 207. *Il s'agit du nombre de vente moyen observé par période de 7 jours par site de vente et dans la période du 01 Janvier 2010 au 16 mars 2012. Cette figure laisse à supposer la variabilité du nombre de vente pouvant être expliquée par le site de vente.*

Maille	*size*	μ	$\hat{D}$	P-value
Peugeot 207 Diesel Berline	3.69	2467.36	0.1373	0.2919
Citroen C3 Essence Berline	3.32	1799.10	0.1359	0.2973

TABLE 3.8 – TEST D'ADÉQUATION À LA LOI BINOMIALE NÉGATIVE. *Ces test ont été effectués sur des ventes sur 102 semaines. Par un test de Kolmogorov-Smirnov, nous voyons que la probabilité de rejeter à tort l'hypothèse selon laquelle le nombre de VO vendus suit une loi Binomiale Négative pour les deux jeux de données avoisine les 30%. L'ajustement par une loi Binomiale Négative convient pour modéliser nos données.*

$$P(N = n) = p(n) = \binom{n + r - 1}{r - 1} p^n (1 - p)^r$$

où r *est un entier et* $0 \leq p \leq 1$, $Var(N) = \dfrac{rp}{(1-p)^2}$ *et* $\mathbb{E}(N) = \dfrac{rp}{(1-p)}$.

De la propriété (3.4.1), nous proposons que la variable aléatoire de $N_i : i = 1, \ldots, K$ décrivant le nombre de VO vendus dans la période T à partir du site i soit modélisée selon une loi Binomiale Négative de paramètres $\mathscr{BN}\left(\alpha, \dfrac{\beta}{\beta + 1}\right)$.

3.4.2 Adéquation des données aux différents modèles proposés

Nous effectuons un test sur la validation de l'ajustement pour s'assurer dans une certaine mesure que les données sur le nombre de vente de VO s'ajuste bien à la distribution Binomiale Négative. Il s'agit de mesurer l'écart entre les données observées et les valeurs théoriques espérées pour une distribution Binomiale Négative en proposant un test statistique de Kolmogorov Smirnov dont l'hypothèse nulle stipule la validité du modèle [CvC88]. Un ajustement graphique a été également effectué. Les tests ont été effectués pour la "Peugeot 207 Diesel Berline" et la "Citroen C3 Essence Berline" et les résultats sont rapportés en Tab. 3.8 et en Fig.3.7.

3.5 Conclusion

Nous avons cherché à répondre au problème de la prédiction du délai de vente des véhicules d'occasion. Nous n'avons trouvé aucun ouvrage pouvant nous servir de référence en ce qui concerne la résolution de cette problématique. Ainsi, notre conclusion s'attachera uniquement à décrire ce que nous avons développé et en apprécier l'efficacité sans aucune autre idée d'appréciation, nous ne pouvons pas affirmer que notre approche est meilleure ou non.

Après une agrégation convenable de nos données, une méthodologie permettant de construire un profil de référence et de quantifier les délais de vente a été proposée. La quantification des délais de vente permet de comprendre le marché VO de manière objective. L'intérêt d'avoir un profil de référence ainsi que la pertinence des indicateurs calculés ont été confirmés numériquement.

Nous avons distingué un modèle de régression linéaire à travers lequel nous avons supposé que le délai de vente résulte de l'association des différentes grandeurs liées au *prix*, à l'*âge* et au *kilométrage* du véhicule. Le problème étant spécifique, l'évaluation de la qualité de prédiction

a été adaptée à ce contexte. Les résultats ont été satisfaisants et les types de modèles obtenus sont facilement interprétables.

Une approche permettant de modéliser le nombre de vente de VO en une période T ainsi que de prédire le délai nécessaire à la vente d'un certain nombre de VO a été développée par la suite. Le nombre de ventes de VO a été ajusté par une loi Binomiale Negative. Ce modèle reflète la variation naturelle des taux de vente moyen observée dans la pratique. L'analyse statistique ainsi que les simulations des ventes réalisées montre que ce modèle est en adéquation avec les données disponibles et peut, par conséquent servir de modèle de base pour les ventes de VO.

Plusieurs axes de recherches peuvent être encore exploités à partir de ces résultats, déterminer entre autres le délai nécessaire à la vente de n VO ou le nombre de zones de vente convenable pour vendre un certain nombre de VO en un intervalle de temps T. Combiner ces modèles avec d'autres qui prennent en compte l'offre de VO similaires sur le marché permettrait d'étendre son utilisation pour résoudre d'autres problématiques.

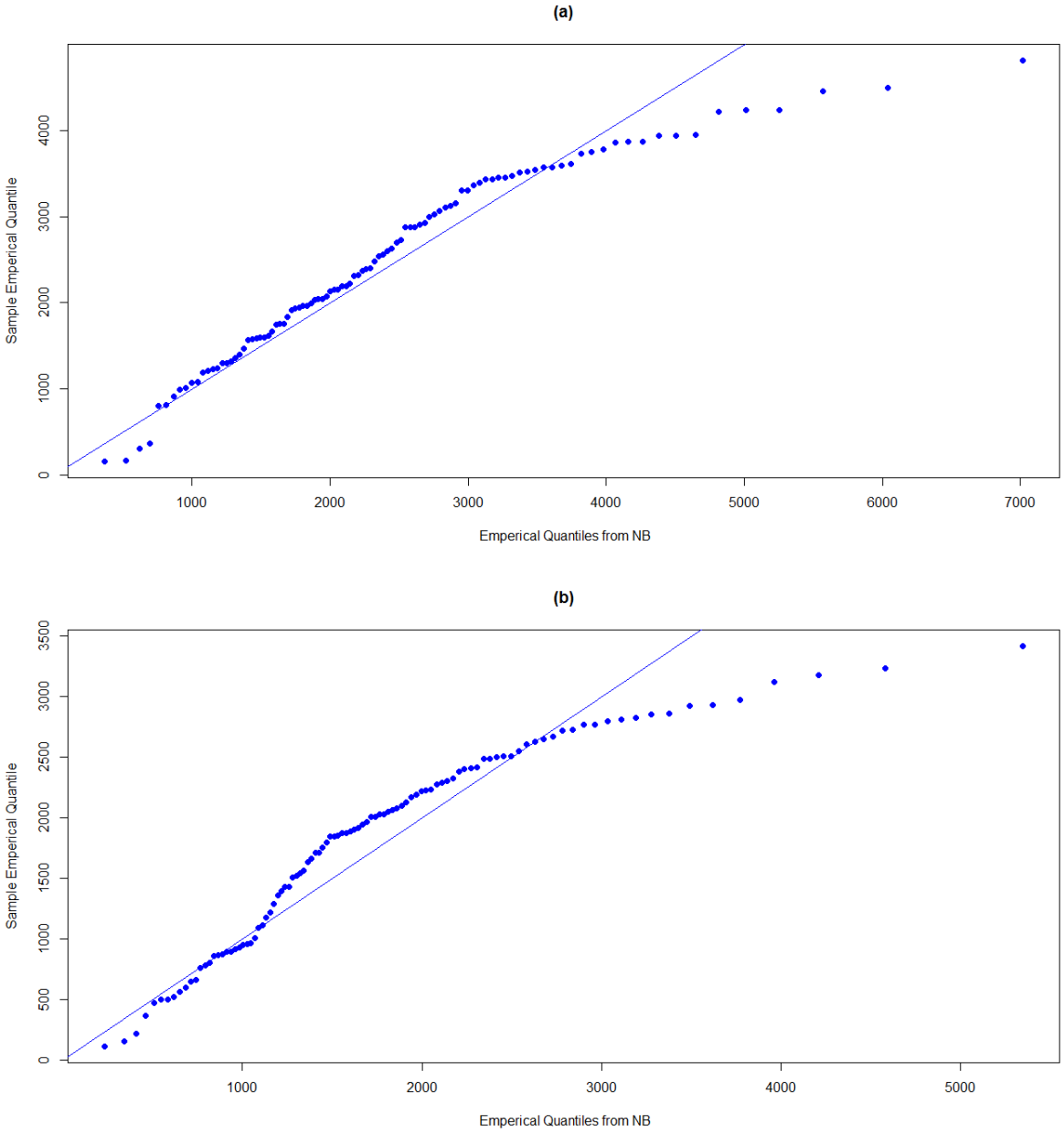

FIGURE 3.7 – ADÉQUATION À LA DISTRIBUTION BINOMIALE NÉGATIVE. *Graphique des quantiles de la loi Binomiale Négative et des quantiles de nos données triées dans un ordre croissant pour la Peugeot 207 (a) et de la Citroen C3 (b).*

Conclusion et perspectives

Le principal propos de cette thèse consiste en l'élaboration de modèles prédictifs des prix et des délais de vente des véhicules d'occasion.

Il ne nous a pas été possible d'atteindre de tels objectifs sans une étude prolongée et fastidieuse sur la qualité des données. C'est pourquoi, il nous a été impératif de consacrer une première partie du travail à l'élaboration des méthodes de détection de valeurs aberrantes.

Synthèse

La motivation de la thèse, le cadre et les objectifs, les contraintes industrielles et la problématique académique ainsi que les enjeux liés à la disposition d'un modèle de qualité ont été présentés dans l'introduction. De plus, dans cette même introduction, nous nous sommes attachés à mieux définir les besoins inhérents à notre sujet au regard de la bibliographie existante sur les sujets de la détection d'outliers et de la prédiction des prix VO. Afin de délimiter et d'identifier les démarches appropriées pour la détection des outliers présents dans la base de données d'Autobiz, le processus d'acquisition des données, leur structure ainsi que la nature des outliers existants ont été présentés. Des considérations générales sur les outliers comprenant les différentes définitions proposées dans la littérature ainsi que les différentes méthodes d'identification usuelles ont été exposées au préalable. Des règles empiriques pour l'identification des outliers ont été proposées pour un traitement préliminaire de la base de données. Sous l'hypothèse que les courbes traduisant les kilométrages observés au cours du temps respectent une allure de référence, une première technique de détection d'outliers consiste en une observation des courbes associées aux différentes quantiles à travers le temps et déclarant comme outliers les courbes dont l'allure dévie significativement de celle des autres. Par ailleurs, cette analyse détaillée nous a permis d'extraire une information pertinente sur les différentes phases d'utilisation d'un véhicule sur toute sa durée de vie.

Le travail sur la détection d'outliers consiste en une élaboration d'une test non paramétrique permettant la détection d'outliers univariés pour des lois non bornées et d'en étudier les propriétés asymptotiques. La statistique de test proposée s'apparente à l'estimateur de Hill et est valable sur une grande variété de lois de probabilité telle que, la valeur absolue d'une gaussienne, la loi gamma, la loi de Weibull et de Student ou les distributions à variations régulières. Les hypothèses pour l'utilisation du test ont été établies en ce qui concerne les lois de probabilité. Dans le but d'illustrer le comportement de la statistique de test proposée, nous avons détaillé plusieurs expériences numériques. Tout d'abord, un effort important s'est porté sur la comparaison de la performance de la statistique de test par rapport aux autres méthodes usuelles connues pour la détection des outliers univariés sur des données simulées par l'approche de Monte Carlo. Il s'est averé qu'à plusieurs reprises, la méthode que nous proposons a une plus grande capacité à ne pas détecter de faux outliers. Ensuite, une application sur des données réelles a été effectuée.

L'étude conduite étant d'une part, d'observer à travers les résidus d'une régression du prix sur le kilométrage et l'âge des véhicules les discordances dans la structure de la relation entre ces variables et d'autre part, d'identifier les outliers parmi les valeurs extrêmes observées pour le kilométrage moyen mensuel des véhicules.

Pour ce qui est de la modélisation et de la prédiction des prix VO, des études préliminaires ont montré les effets sous-jacents des variables existant dans notre base de données sur la variabilité du prix. Nous avons constaté qu'en isolant les effets de certaines variables qualitatives caractéristiques du VO, la quantification de leur impact devient pertinente. Cela nous a incité, pour notre étude, à se placer à une échelle d'analyse correspondant au niveau de maille assez fine. Dans la modélisation, notre choix s'est d'abord porté sur un modèle de régression linéaire dans lequel nous supposons que le prix est une combinaison linéaire des variables explicatives. Les valeurs des coefficients indiquent alors l'influence de la variable explicative sur la variable réponse, et leur signe décrit la nature de cette influence. Plusieurs limites ont été constatées dans son utilisation. La prise en compte du logarithme du prix comme variable dépendante a permis de résoudre certaines problématiques sans répondre complètement aux objectifs industriels. Une légère restriction sur la flexibilité des modèles additifs étudiés en régression non paramétrique nous a permis d'atteindre notre objectif. Nous avons justifié cette approche par la spécificité de notre problème. Ce modèle suppose que la vraie fonction ayant généré les données ne peut pas être approchée convenablement par des fonctions linéaires mais par une somme de fonction des variables explicatives. La structure de ces fonctions a été approchée à partir d'une visualisation graphique de la relation entre le prix et les variables âge et km. Ce modèle a été estimé par les moindres carrés et également par la méthode de régression sur les quantiles. Dans chaque cas, l'algorithme de résolution a été défini de telle sorte qu'il permet de prendre en compte toutes les contraintes imposées par les experts et d'incorporer dans le modèle les connaissances que nous avons du phénomène étudié. Ce modèle a été ensuite estimé par une régression sur les quantiles. L'intérêt non négligeable de ce modèle réside dans son interprétabilité et l'application de la régression sur les quantiles a offert une plus grande précision en terme de prédiction. Ce problème étant spécifique, une approche sur l'évaluation de la qualité de prédiction a été proposée. De leur particularité, les VO de luxe et de prestige ont été exclus de notre étude.

Le problème sur les délais de vente a été décomposé en plusieurs sous-problèmes. Après une agrégation convenable de nos données, nous avons élaboré dans une étape descriptive une méthode permettant de définir un profil de référence afin de quantifier les délais de vente. Nous avons ensuite introduit un modèle de régression linéaire à travers lequel une procédure de sélection de modèles par apprentissage a été présentée pour choisir les paramètres. Un modèle convenable a été trouvé et les résultats sur une base de données servant de test ont été pertinents. Le nombre de ventes de VO dans une période T a été modélisé par une loi de mélange Poisson-Gamma. Dans la validation du modèle, la majorité des ventes de VO observées dans la base de données a pu être reproduite grâce au modèle. Les résultats de la comparaison entre les essais simulés et les données réelles sont très raisonnables.

Bilan et perspectives

Rappelons que dans cette thèse la théorie a été guidée par les applications et la volonté d'obtenir des résultats exploitables.

L'étude de la bibliographie nous a révélé que de nombreuses techniques ont été appliquées

dans l'analyse du marché VO. Malheureusement, il manquait jusqu'à présent des modélisations formelles et exhaustives, aussi bien pour le prix que pour les délais de vente, pouvant servir de base de comparaison. C'était avec beaucoup d'intérêt que nous avons participé à la mathématisation du problème. Les résultats obtenus ont confirmé la pertinence de notre approche et l'adéquation de la méthodologie au problème posé. Toutes ces méthodes ont été validées par **Autobiz** et font l'objet d'une mise en production commerciale. Cependant, nous sommes bien conscients que de nombreux perfectionnements pourront prolonger ce travail. En effet, le cheminement proposé, partant de l'analyse de données, de la détection des outliers vers la modélisation, puis de la modélisation vers les applications, n'est pas exempte de critiques de différents ordres, dont la plupart pourraient se transformer en pistes de recherche pertinentes pour le futur. Nous ne doutons pas que dans un avenir proche nous assisterons à la conception des modèles de plus en plus réalistes tant sur la prédiction des prix que la modélisation des délais de vente des VO.

Dans la mise en œuvre de la méthodologie, des observations manquantes ou atypiques, des erreurs de saisie et/ou de codage, de mauvaises déclarations volontaires ou involontaires ont fait que les données, à plusieurs reprises ont été rebelles à l'analyse. Le repérage de toutes ces imperfections bien qu'il puisse être intéressant du point de vue méthodologique, n'a sans doute pas été l'aspect le plus gratifiant du travail. Les procédures que nous avons développées pour la détection d'outliers ne résolvent pas ces difficultés d'un coup, mais elles peuvent se présenter comme des jumelles à travers lesquelles les praticiens peuvent gérer les observations extrêmes avec une certaine objectivité. Il faut continuer les efforts dans le mode d'acquisition des données pour arriver à de meilleurs résultats. Dans cette optique, la recherche académique peut aider à l'amélioration des données en faisant appel à des problématiques de classification textuelle. L'extension de l'analyse des courbes des kilométrages aux thématiques liées à la detection d'outliers sur les courbes fonctionnelles est envisageable comme nous pouvons voir dans [FGGM08] [Ger09]. La construction de notre statistique de test peut être également proposée en se plaçant toujours dans un cadre univarié tout en avançant une hypothèse supplémentaire permettant d'identifier les outliers parmi les valeurs extrêmes inférieures. Les deux phénomènes liés à la détection des outliers et qui sont le «masking effect» ainsi que le «swampping effect» peuvent également être étudiés pour la statistique de test qui a été proposée.

Le travail sur la modélisation de prix VO peut être facilement enrichi de plusieurs manières. Une des premières possibilités est l'introduction d'autres facteurs, notamment le facteur régional qui peut être pris en compte sous forme de variables indicatrices associées chacune à une zone de vente pouvant être un département ou une région administrative. Il est regrettable que notre modélisation ne puisse tenir compte de l'état du véhicule et de la volumétrie de l'offre de véhicules similaires disponibles sur le marché VO au moment de la vente. Il faudrait pour cela être capable de quantifier le biais systématique du à la sous-estimation des prix finaux. Ces arguments sont nécessaires pour bien prendre en considération le mécanisme réel du marché VO. Nous n'avons pas développé cette question qui n'entre pas dans notre domaine de compétence. Il faut aussi à long terme dans un travail futur tenir compte dans la détermination du prix de la valeur résiduelle du VO. En effet, les caractéristiques des véhicules étant de moins en moins différenciées, les valeurs résiduelles fortes et stables représentent un facteur clé de succès distinguant le produit de la marque de celui de ses compétiteurs. Cette question a déjà été soulevée par Autobiz, mais faute de temps et de moyens, nous n'avons pas pu l'aborder dans le cadre de cette thèse. Cette forme de prédiction est très intéressante pour les acteurs dans la location de courte et longue durée. Nous avons eu beaucoup de difficulté à modéliser le prix des VO récents, notamment ceux de moins d'un an. Nous aurions pu adopter une autre stratégie d'analyse, face

à cette difficulté, et opter pour une approche complètement différente pour les données touchant tous les VO inférieurs à 12 mois. À propos de l'algorithme de recherche de modèle optimal, il serait plus satisfaisant de proposer en même temps la minimisation de l'écart quadratique et la maximisation de la probabilité d'obtenir le seuil α. Orienter donc le travail dans un cadre d'optimisation sous contrainte en probabilité [Ury00]. Le développement d'un modèle à une échelle d'agrégation plus globale peut paraître hors du champ des problèmes de recherche. Pourtant du point de vue de la mise en œuvre industrielle cela représente probablement le défi le plus long, le plus dur mais le plus intéressant à réaliser.

Le travail sur les délais de vente est riche de perspectives. L'hypothèse selon laquelle le véhicule est vendu lorsque l'annonce disparait peut être remise en cause puisque qu'elle peut diminuer de façon importante la précision des estimations des paramètres de régression ainsi que la puissance des tests statistiques pratiqués. Au cours de l'étude des délais de vente, nous n'avons pu occulter le problème que posait l'ampleur de ce biais d'information. Quant à la définition de zone de vente dans l'approche modélisatrice, un cas pratique et plus réaliste associerait les zones de vente à des points de vente d'un même groupe *(ex : garage Renault)*. Par ailleurs, la particularité essentielle que présentait la proposition d'un profil de référence pour les délais de vente, à savoir la quantification de l'hétérogénéité prend tout son sens lors d'une étude au sein de laquelle le professionnel du métier désire connaitre les caractéristiques des segments de véhicules « les plus lents ». Un prolongement naturel de cette étude eût été l'analyse plus particulière de ces groupes « lents », afin de déterminer les causes de ce phénomène de ralentissement de la rotation des délais de vente. Une telle étude pourra faire appel aux méthodes de classification et d'analyse discriminante. La modélisation du délai nécessaire à la vente d'une quantité n de VO à travers K sites de vente semble également constituer une suite naturelle au travail sur la modélisation du nombre de vente.

L'efficacité et la fiabilité du SystemVO, produit phare d'Autobiz, pourraient bénéficier de la suite de ces pistes de réflexion.

Bibliographie

[AB12] J.M. Azaïs et J.M. Bardet : *Le modèle linéaire par l'exemple - 2e éd. : Régression, analyse de la variance et plans d'expérience illustrés avec R et SAS.* Mathématiques. Dunod, 2012.

[Abd07] H. Abdi : *Bonferroni and Sidak corrections for multiple comparisons.* Sage, 2007.

[AFO94] T. W. Anderson, K. Fang et I. Olkin : *Multivariate Analysis and Its Applications.* Institute of Mathematical Statistics : Lecture notes, monograph series. Institute of Mathematical Statistics, 1994.

[AHMftF93] A. Alberini, W. Harrington, V.D. McConnell et Resources for the Future : *Determinants of participation in accelerated vehicle retirement programs.* Discussion paper. Resources for the Future, 1993.

[Baa68] W. Baarda : *A Testing Procedure for Use in Geodetic Networks*, volume 2 de *A testing procedure for use in geodetic networks.* Netherlands Geodetic Commission, 1968.

[Bar76] V. Barnett : The ordering of multivariate data. *Journal of royal staitistical society. Series A(General).*, 139(3):318–355, march 1976.

[Bas56] D. Basu : The concept of asymptotic efficiency. *Sankhyā : The Indian Journal of Statistics (1933-1960)*, 17(2):pp. 193–196, 1956.

[Bas11] S. Baselga : Nonexistence of rigorous tests for multiple outlier detection in least-squares adjustment. *Journal of Surveying Engineering*, 137(3):109–112, 2011.

[BASH06] I. Bhatti, H. Al-Shanfari et Z. Hossain : *Econometric Analysis of Model Selection and Model Testing.* Ashgate, 2006.

[BC83] R. J. Beckman et R. D. Cook : Outlier....s. *Technometrics*, 25(2):119–149, may 1983.

[BCOCSL07] F. Belzunce, A. Castaño, A. Olvera-Cervantes et A. Suárez-Llorens : Quantile curves and dependence structure for bivariate distributions. *Computational Statistics & Data Analysis*, 51(10):5112 – 5129, 2007.

[BD11] D. Birkes et Y. Dodge : *Alternative Methods of Regression.* Wiley Series in Probability and Statistics. Wiley, 2011.

[BDH82] P.J. Bickel, K. Doksum et J.L. Hodges : *The notion of the breakdown point, in A Festschrift For Erich L. Lehmann.* Wadsworth international statistics. Taylor & Francis, 1982.

[Ber85] J. Berkovec : New car sales and used car stocks : A model of the automobile market. *The RAND Journal of Economics*, 16:195–214, 1985.

[BF85] L. Breiman et J. H. Friedman : Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, september 1985.

[BGP96] P. L. Brockett, L. L. Golden et H. H. Panjer : Flexible purchase frequency modeling. *Journal of Marketing Research*, 33(1):94–107, febroary 1996.

[BHT89] A. Bujas, T. Hastie et R. Tibshirani : Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–555, 1989.

[BK96] W. Butler et R. Kavesh : How business economists forecast. *Financial Analysts Journal*, 22(5):144–146, october 1996.

[BL94] V. Barnett et T. Lewis : *Outliers in Statistical Data*. Wiley Series in Probability & Statistics. Wiley, 1994.

[BLW09] D. Bams, T. Lehnert et C. C. P. Wolff : Loss functions in option valuation : A framework for selection. *Management Science*, 55(5):pp. 853–862, 2009.

[Boi56] M. Boiteux : L'amortissement dépréciation des automobiles. *Revue de statistique appliquée*, 4(4):57–73, february 1956.

[BS71] R. C. Blattberg et T. Sargent : Regression with non-gaussian stable disturbances : Some sampling results. *Econometrica*, 39(3):501–510, may 1971.

[Buc94] Moshe Buchinsky : Changes in the U.S. Wage Structure 1963-1987 : Application of Quantile Regression. *Econometrica*, 62(2):405–58, March 1994.

[Car76] G. Carletti : Détection automatique de valeurs anormales, revue de statistique appliquée. *Revue de statistique Appliquées*, 24(3):61–70, 1976.

[CDS94] P. Chaudhuri, K. Doksum et A. Samarov : Regression estimators based on conditional quantiles. *Journal of Nonparametric Statictics*, 3:317–321, 1994.

[Cho05] S. J. Cho : The determinants of used rental car prices. *Journal of economic research*, 2005.

[CI80a] D. R. Cox et V. Isham : *Point Processes*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1980.

[CI80b] D.R. Cox et V. Isham : *Point Processes*. Monographs on applied probability and statistics. Chapman and Hall, 1980.

[CM04] S. Choo et P. L. Mokhtarian : What type of vehicle do people drive ? the role of attitude and lifestyle in influencing vehicle type choice. *Transportation Research Part A : Policy and Practice*, 38(3):201 – 222, 2004.

[Cox55] D. R. Cox : Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):129–164, march 1955.

[CvC88] P. Capéraà et B. van Cutsem : *Méthodes et modèles en statistique non paramétrique : exposé fondamental*. Dunod décision. Presses de l'Université Laval, 1988.

[DG07] P. L. Davies et U. Gathier : The breakdown point - examples and counterexamples. *Revstat*, 5(1):1–17, march 2007.

[DJ00] Y. Dodge et J. Jureckova : *Adaptive Regression*. Springer New York, 2000.

[DR05] D. K. Dey et C. R. Rao : *Handbook of Statistics : Bayesian Thinking, Modeling and Computation*. Handbook of Statistics. Elsevier Science, 2005.

[DR12] S.-F. Dimby et J. Rynkiewicz : Quantile regression with multilayer perceptrons. *In ESANN 2012. The 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Proceedings - Bruges, Belgium from 25 to 27 April 2012*, pages 41–50. ESANN, 2012.

[DS11] J. J. Droesbeke et G. Saporta : *Approches non paramétriques en régression.* Editions Technip, 2011.

[Dub70] S. D. Dubey : Compound gamma, beta and f distributions. *Metrika*, 16(1):27–31, 1970.

[EC07] T. G. Edwards et K. R. Chelst : Purchasing a used car using multiple criteria decision making. *The Mathematics Teacher*, 101(2):126–135, september 2007.

[EE11] S. Eakambaram et R. Elangovan : *Least Absolute Deviation Regression Theory and Methods.* LAP Lambert Acad. Publ., 2011.

[EHS09] M. Engers, M. Hartmann et S. Stern : Annual miles drive used car prices. *Journal of Applied Econometrics*, 24(1):1–33, 2009.

[EM92] S. P. Ellis et S. Morgenthaler : Leverage and breakdown in l1 regression. *Journal of the American Statistical Association*, 87(417):143–148, march 1992.

[ET94] B. Efron et R.J. Tibshirani : *An Introduction to the Bootstrap.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[Eub88] R. L. Eubank : *Spline smoothing and nonparametric regression.* Marcel Dekker Inc, february 1988.

[FGGM08] M. Febrero, P. Galeano et W. González-Manteiga : Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, 19(4):331–345, 2008.

[FM01] Ricardo Fraiman et Graciela Muniz : Trimmed means for functional data. *Test*, 10(2):419–440, 2001.

[Fri80] G. A. Frisbie : Ehrenberg's negative binomial model applied to grocery store trips. *Journal of Marketing Research*, 17(3):385–390, August 1980.

[Gen91] D.J. Genesove : *Adverse Selection in the Wholesale Used Car Market.* 1991.

[Ger09] D. Gervini : Detecting and handling outlying trajectories in irregularly sampled functional datasets. *The Annals of Applied Statistics*, 3(4):1758–1775, dec 2009.

[GK72] R. Gnanadesikan et J. R. Kettenring : Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124, march 1972.

[Gro03] J. Gross : *Linear Regression.* Lecture Notes in Statistics - Springer. Springer Berlin Heidelberg, 2003.

[Gru69] F. E. Grubbs : Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, february 1969.

[GS99] U. Gather et V. Schultze : Robust estimation of scale of an exponential distribution. *Statistica Neerlandica*, 53(3):327–341, 1999.

[Gyö02] L. Györfi : *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer, 2002.

[Här90] W. Härdle : *Applied Nonparametric Regression.* Econometric Society Monographs. Cambridge University Press, 1990.

[Haw80] D. M. Hawkins : *Identification of Outliers*. Monographs on applied probability and statistics. Chapman and Hall, 1980.

[HB03] M. Hubert et K. V. Branden : Robust methods for partial least squares regression. *Journal of Chemometrics*, 17(10):537–549, 2003.

[HB10] M. A. Haan et H.-W. Boer : Has the internet eliminated regional price differences ? evidence from the used car market. *De Economist*, 158(4):373–386, 2010.

[HBK84] D. M. Hawkins, D. Bradu et G. V. Kass : Location of several outliers in multiple-regression data using elemental sets. *Technometrics.*, 26(3):197–208, august 1984.

[HL02] J. L. Horowitz et S. Lee : Semiparametric methods in applied econometrics : do the model fit the data ? *Statistical Modelling*, 2(1):3–22, april 2002.

[Hoc05] R. R. Hocking : *Methods and Applications of Linear Models : Regression and the Analysis of Variance*. Wiley Series in Probability and Statistics. Wiley, 2005.

[HR11] P.J. Huber et E.M. Ronchetti : *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 2011.

[HT86] T. Hastie et R. Tibshirani : Generalized additive models. *Statistical Science*, 1986.

[HTF09] T. Hastie, R. Tibshirani et J.H. Friedman : *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer series in statistics. Springer, 2009.

[Jan04] A. Janczak : *Identification of Nonlinear Systems Using Neural Networks and Polynomial Models : A Block-Oriented Approach*. Lecture Notes in Control and Information Sciences. Springer, 2004.

[Jer08] A. Jerenz : *Revenue Management and Survival Analysis in the Automobile Industry*. Gabler Edition Wissenschaft. Betriebswirtschaftlicher Verlag Gabler, 2008.

[JKB94] N .L. Johnson, S. Kotz et N. Balakrishnan : *Continuous Univariate Distributions*, volume 1. New York, 2 édition, 1994.

[JKB97] N. L. Johnson, S. Kotz et N. Balakrishnan : *Discrete Multivariate Distributions*. Wiley Series in Probability and Statistics. Wiley, 1997.

[JKK92] N. L. Johnson, S. Kotz et A. W. Kemp : *Univariate discrete distributions*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. John Wiley & Sons, 1992.

[JKK93] N. L. Johnson, S. Kotz et A. W. Kemp : *Univariate discrete distributions*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. John Wiley & Sons, 2 édition, 1993.

[JKP01] J. Jureéckovâ, R. Koenker et S. Portnoy : Tail behavior of the least-squares estimator. *Statistics & Probability Letters*, 55(4):377 – 384, 2001.

[JWK05] N. L. Johnson, Kemp A. W. et S. Kotz : *Univariate discrete Distributions*, volume 1. Hoboken, New York, 2 édition, 2005.

[Kad11] J. B. Kadane : *Principles of Uncertainty*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2011.

[KB78] R. Koenker et G. Bassett : Regression quantiles. *Econometrica*, 46:33–50, 1978.

[Kel88] J. H. Kellar : New methodology reduces importance of used cars in the revised cpi. *Monthly Labor Review*, 111(12):34–36, December 1988.

[KH06] P. Kooreman et M. A. Haan : Price anomalies in the used car market. *De Economist*, 154(1):41–62, 2006.

[KHM05] Masha Kocherginsky, Xuming He et Yunming Mu : Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14(1):41–55, 2005.

[Kim82] A. C. Kimber : Tests for many outliers in an exponential sample. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):263–271, april 1982.

[KNT00] Edwin M. Knorr, Raymond T. Ng et Vladimir Tucakov : Distance-based outliers : algorithms and applications. *The VLDB Journal*, 8(3-4):237–253, 2000.

[Koe05] R. Koenker : *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

[Kol06] S. Kollias : *Artificial Neural Networks - ICANN 2006 : 16th International Conference, Athens, Greece, September 10-14, 2006, Proceedings*. Artificial Neural Networks : ICANN 2006 : 16th International Conference, Athens, Greece, September 10-14, 2006 : Proceedings. Springer, 2006.

[Kre59] M. E. Kreinin : Analysis of used car purchases. *The Review of Economics and Statistics*, 41(4):419–425, november 1959.

[Kri96] K. Krickeberg : *Petit Cours de Statistique*. Springer, 1996.

[KS81] A. C. Kimber et H. J. Stevens : The null distribution of a test for two upper outliers in an exponential sample. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30(2):153–157, april 1981.

[Law80] R. J. Lawrence : The lognormal distribution of buying frequency rates. *Journal of Marketing Research*, 17(2):212–220, may 1980.

[LH96] O. B. Linton et W. Härdle : Estimation of additive regression models with known links. *Biometrika*, 83(3):529–540, 1996.

[Man86] E. B. Manoukian : *Guide de statistique appliquée*. Collection Méthodes. Editions Hermann, 1986.

[MGMRPA90] J. Munõz-Garcia, J. L. Moreno-Rebollo et A. Pascual-Acosta : Outliers : A formal approach. *International Statistical Review*, 58(3):215–226, december 1990.

[MK85] A. F. S. Mitchell et W. J. Krzanowski : The mahalanobis distance and elliptic distributions. *Biometrika*, 72(2):464–467, august 1985.

[Mon67] Seasonal demand and used car prices. *Monthly Labor Review*, 90(10):12–16, october 1967.

[MS99] J. Murray et N. Sarantis : Quality, user cost, forward-looking behavior, and the demand for cars in the uk. *Journal of Economics and Business*, 51(3):237–258, 1999.

[MT12] R. B. McKenzie et G. Tullock : Pricing lemons, views, and university housing. *In The New World of Economics*, pages 69–90. Springer Berlin Heidelberg, 2012.

[NdMM05] N. Nedjah et L. de Macedo Mourelle : *Evolvable Machines : Theory & Practice*. Studies in Fuzziness and Soft Computing. Springer, 2005.

[NW77] S. C. Narula et J. F. Wellington : Prediction, linear regression and the minimum sum of relative errors. *Technometrics*, 19(2):185–190, may 1977.

[NW82] S. C. Narula et J. F. Wellington : The minimum sum of absolute errrors regression : A state of the art survey. *International Statistical Review*, 50(3):317–326, december 1982.

[NW96] J. Neter et W. Wasserman : *Applied Linear Statistical Models*. The Irwin series in statistics. WCB/McGraw-Hill, 1996.

[NW02] S. C. Narula et J. F. Wellington : Sensitivity analysis for predictor variables in the msae regression. *Comput. Stat. Data Anal.*, 40(2):355–373, août 2002.

[Ops00] J. D. Opsomer : Nonparametric regression in environmental statistics, 2000.

[Pas13] B. P. Pashigian : *The Used Car Price Index : A Checkup and Suggested Repairs*. BiblioGov, 2013.

[Pen96] K. I. Penny : Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *Journal of royal staitistical society. Series C (Applied Statistics).*, 45(1):73–81, 1996.

[Pre78] P. Prescott : Examination of the behaviour of tests for outliers when more than one outlier is present. *Journal of the Royal Statistical Society. Series C(Applied Sytatistics)*, 27(1):10–25, 1978.

[Pur92] D. Purohit : Exploring the relationship between the markets for new and used durable goods : The case of automobiles. *Marketing Science*, 11(2):154–167, 1992.

[Que49] M. H. Quenouille : A relation between the logarithmic, poisson, and negative binomial series., 1949.

[Ray07] S. Rayer : Population forecast accuracy : Does the choice of summary measure of error matter ? *Population Research and Policy Review*, 26(2):pp. 163–184, 2007.

[RHW88] D. E. Rumelhart, G. E. Hinton et R. J. Williams : Neurocomputing : Foundations of research. chapitre Learning Internal Representations by Error Propagation, pages 673–695. MIT Press, Cambridge, MA, USA, 1988.

[RL05] P. J. Rousseeuw et A. M. Leroy : *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics. Wiley, 2005.

[Ros75] B. Rosner : On the detection of many outliers. *Technometrics*, 17(2):221–227, may 1975.

[RPD98] J. O. Rawlings, S. G. Pantula et D. A. Dickey : *Applied Regression Analysis : A Research Tool*. Springer Texts in Statistics. Wadsworth & Brooks, 1998.

[RS06] J. Ramsay et B.W. Silverman : *Functional Data Analysis*. Springer Series in Statistics. Springer, 2006.

[Ryn08] J. Rynkiewicz : Consistent estimation of the architecture of multilayer perceptrons. *ArXiv e-prints*, février 2008.

[Ryn12] J. Rynkiewicz : General bound of overfitting for mlp regression models. *ArXiv e-prints*, janvier 2012.

[RZ08] P. Radaelli et M. Zenga : Quantity quantiles linear regression. *Statistical Methods and Applications*, 17(4):455–469, 2008.

[SD70] M. Snowbarger et B. Dunkelberg : A cross-sectional study of used car prices : 1962-64. *Journal of Marketing Research*, 7(4):493–497, november 1970.

[SL92] P. Sprent et J. P. Ley : *Pratique des statistiques nonparamétriques*. Techniques et pratiques. Institut national de la recherche agronomique, 1992.

[Sol06] B. Solaiman : *Processus stochastiques pour l'ingénieur*. Collection technique et scientifique des télécommunications. Presses polytechniques et universitaires romandes, 2006.

[Sto85] C. J. Stone : Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, june 1985.

[SW07] M. Somers et J. Whittaker : Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, 183(3):1477 – 1487, 2007.

[TB91] Y. K. Tse et U. Balasooriya : Tests for multiple outliers in an exponential sample. *The Indian Journal of Statistics, Series B*, 53(1):56–63, april 1991.

[TM72] G. L. Tietjen et R. H. Moore : Some grubbs-type statistics for the detection of several outliers. *Technometrics*, 14(3):583–597, 1972.

[TSB99] J. Tayman, D. A. Swanson et C. F. Barr : In search of the ideal measure of accuracy for subnational demographic forecasts. *Population Research and Policy Review*, 18(5):pp. 387–409, 1999.

[Tsy08] A. B. Tsybakov : *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2008.

[Tuk70] J. W. Tukey : *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1970.

[Ury00] S. Uryasev : *Probabilistic Constrained Optimization : Methodology and Applications*. Nonconvex Optimization and Its Applications. Springer, 2000.

[vdV00] A. W. van der Vaart : *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[WA85] A. A. Weiss et A. P. Andersen : Estimating time series models using the relevant forecast evaluation criterion, journal of the royal statistical society (a) 147 (1984), pp. 484-487. *International Journal of Forecasting*, 1(4):314–314, 1985.

[Wei91] Andrew A. Weiss : Estimating nonlinear dynamic models using least absolute error estimation. *Econometric Theory*, 7:46–68, 3 1991.

[Whi92] H. White : *Artificial Neural Networks : Approximation and Learning Theory*. Library of American Biography. Blackwell, 1992.

[Wil12] R. R. Wilcox : *Introduction to Robust Estimation and Hypothesis Testing*. Statistical modeling and decision science. Academic Press, 2012.

[WJ03] G. C. S. Wang et C. L. Jain : *Regression Analysis : Modeling & Forecasting*. Graceway Pub., 2003.

[ZHDC12] M. Zhou, L. Hannah, D. B. Dunson et L. Carin : Beta-negative binomial process and poisson factor analysis. *Journal of Machine Learning Research - Proceedings Track*, 22:1462–1471, 2012.

C'est absurde, repartit Ulrich avec force. J'ai dit que ce qui comptait, ce n'était pas un faux pas, mais le pas qui suit ce faux pas. Mais qu'est-ce qui compte après le pas suivant ? Sans doute, bien sûr, celui qui suit ? Et après le énième pas, le pas n+1 ? Cet homme devrait donc vivre privé de fin et de décision, privé même, somme toute, de réalité. Pourtant il est bien vrai que c'est toujours le pas suivant qui compte. La vérité est que nous ne disposons d'aucune méthode pour traiter comme il faudrait cette série infatigable.
Extrait de ***"L'homme sans qualités"****, Robert Musil.*

Résumé : La société Autobiz édite et diffuse de l'information sur le secteur automobile. Cette thèse contribue à l'enrichissement de cette information et à une meilleure compréhension du marché de l'occasion par l'élaboration des modèles de prédiction du prix des véhicules et du délai de vente qui leur est associé. Nous avons eu à notre disposition une base de données réelles constituée d'annonces de sources diverses induisant une nombre considérable d'outliers. Ainsi, la première partie de travail s'est consacrée à la construction de méthodes de détection d'outliers incluant aussi bien de simples règles empiriques qu'un test statistique dont les propriétés asymptotiques ont été étudiées. Partant d'un état de l'art sur la prédiction des prix des véhicules d'occasion, il est apparu que les études existantes soulèvent le besoin de fonder une méthodologie d'analyse plus rigoureuse. Cette méthodologie a été développée dans un objectif de proposer des solutions automatisables et adaptées aux contraintes imposées par les experts. Nous faisons alors l'hypothèse que les prix des véhicules d'une même version se déprécient en fonction de l'âge et du kilométrage selon une forme qui lui est propre. La dernière partie du travail est dédiée à l'analyse des délais de vente. Dans un premier temps, nous caractérisons la variable associée aux délais de vente. Ensuite nous proposons une modélisation de cette variable par une régression à l'échelle d'un segment correspondant à l'arborescence marque-modèle-carrosserie-énergie en fonction des variables liées au kilométrage, au prix et à l'âge. Enfin, nous discutons de la possibilité de modéliser le nombre de véhicules vendus dans une période donnée selon une loi binomiale négative.

Mots-clefs : outliers, modèle de régression, critère de prédiction, prédiction de prix, délais de vente, véhicules d'occasion, nombre de vente.

Abstract : Autobiz publishes information on the automotive sector. The subject of this thesis is to give more tools for best understanding the used cars market by proposing modeling the price and the sale duration of vehicles. In our disposal we have a dataset consisted of used car advertisements automatically collected from the most popular website in France. Such data records often include outlying values. So, we need to start our analysis by considering outliers problem and we propose an outliers detector for univariate case for which we study asymptotic properties. Next, we develop a predicting model for used cars price. Although enumerable amount of works are stored in the literature we see that each of them lacks rigorous statistical foundations. We investigate the relationships between the price, the mileage, the age and others vehicle characteristics. More precisely we discuss how incorporate these variables in a model and compare different modeling approaches with the object to find the one best fitting the dataset and easy to implement. Expert's opinions are minded at different stages of the model-building process. Next, we identify variables and how they affect the probability of a used vehicle's sale from a list of explanatory variables related to price, mileage and age. In the sequel, we build a model allowing predicting the sale duration. Finally, we discuss about modeling sales of used cars by using the negative binomial distribution.

Keywords : outliers, modelling, prediction, used cars price, policy price for used cars, sale duration.