Lien entre comportement cinématique des véhicules routiers et la survenue des accidents

Yannick Saas

Master 1 Mathématiques et Applications

Spécialité Statistique

Université de Strasbourg

Lien entre comportement cinématique des véhicules routiers et la survenue des accidents

Rapport de stage

Auteur : Yannick SAAS

Responsable du Master : Mme Armelle GUILLOU

Maîtres de stage : Mr Maurice ARON & Mme Régine SEIDOWSKY

Réalisé à l'IFSTTAR / GRETTIA **Juin – Août 2011**

IFSTTAR : Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux

http://www.ifsttar.fr

Laboratoire GRETTIA (Génie des Réseaux de Transports Terrestres et Informatique Avancée)

Site de Marne-la-Vallée

2, rue de la Butte-Verte – 93166 Noisy-le-Grand CEDEX

Université de Strasbourg

UFR de Mathématique et d'Informatique

7, rue René Descartes – 67084 Strasbourg CEDEX

03 68 85 50 00

http://mathinfo.unistra.fr

Remerciements

Je tiens, tout d'abord, à remercier Neïla BHOURI, chargée de recherche au GRETTIA, qui m'a mis en contact avec ses collègues en recherche d'un stagiaire, ainsi que Jean-Patrick LEBACQUE, directeur du GRETTIA, de m'avoir permis de réaliser mon stage au sein de son unité de recherche.

Je remercie ensuite naturellement mes deux maîtres de stage, Maurice ARON et Régine SEIDOWSKY, pour leur disponibilité et leur ouverture. Ils m'ont accompagné tout au long du stage et ont toujours été à l'écoute de mes difficultés.

Merci également à Mustapha TENDJAOUI, ingénieur de recherche au GRETTIA, pour son aide précieuse et ses conseils lors de la résolution de problèmes d'ordre informatique. Je remercie aussi Simon COHEN, directeur de recherche au GRETTIA, pour les éclaircissements qu'il m'a apportés en théorie du trafic.

Je remercie enfin toute l'équipe du GRETTIA pour son accueil et sa convivialité.

Résumé

Ce rapport de stage de fin de Master 1 est le fruit d'une expérience de trois mois, de juin à août 2011, qui m'a fait intégrer le GRETTIA, un des laboratoires de recherche de l'IFSTTAR sur les transports terrestres. J'ai effectué ce stage sous la tutelle de Maurice ARON et de Régine SEIDOWSKY, chargés de recherche au GRETTIA.

Le but de ce stage est de faire le lien entre la survenue des accidents et les caractéristiques cinématiques des véhicules, telles que les vitesses ou les temps intervéhiculaires.

Pour répondre à ce problème, nous pouvons considérer l'approche suivante : nous élaborons, à l'aide des principales variables de trafic, des variables dont on pense, a priori, qu'elles pourraient être des indicatrices pertinentes pour quantifier le risque d'accident. Etant donné que nous ne disposons pas des « boîtes noires » des véhicules accidentés, nous considérons la situation cinématique des véhicules correspondant à la situation d'avant-accident. Pour valider la pertinence des indicateurs de risque construits, il s'agit alors de comparer les valeurs qu'ils prennent en situations d'avant-accident aux valeurs prises en situations d'absence d'accident.

Cette procédure d'évaluation de la qualité des indicateurs de risque est principalement basée sur l'évaluation des mesures de performances, telle que l'Odd Ratio, associées à un modèle de régression logistique sous-jacent. Le cœur de l'étude consiste donc à mettre au point des méthodes statistiques permettant de valider ou d'infirmer la pertinence des indicateurs de risque.

Tables des matières

Remerciements **3**

Résumé **4**

1. Introduction **7**
1.1. L'IFSTTAR et le GRETTIA 7
1.2. Présentation de la problématique du stage 9
1.3. Présentation des données 9
1.3.1. Les données de trafic 9
1.3.2. Les données d'accidents 11
1.4. Stratégie d'attaque 12

2. Les indicateurs de risque **13**
2.1. Introduction 13
2.2. Les variables microscopiques de trafic 13
2.2.1. Les variables microscopiques individuelles 13
2.2.2. Les variables microscopiques agrégées 16
2.2.3. Corrélations entre variables microscopiques 18
2.3. Les variables macroscopiques de trafic 21
2.3.1. Construction 21
2.3.2. Corrélations entre variables macroscopiques 22
2.4. Corrélations entre variables microscopiques et variables macroscopiques 23
2.4.1. Introduction 23
2.4.2. Lien entre débit moyen et temps intervéhiculaire moyen 23
2.4.3. Autres corrélations 25

3. Traitement préparatoire des données **27**
3.1. Introduction 27
3.2. Traitement préparatoire des données d'accidents 27
3.3. Traitement préparatoire des données de trafic 28
3.4. Traitement du lien entre données d'accidents et données de trafic 30
3.5. Agrégation finale des fichiers de données 31

4. Les outils de la régression logistique **32**
4.1. Introduction 32
4.2. Principe générale de la régression logistique 33
4.2.1. Introduction 33
4.2.2. La transformation logistique 34
4.2.3. Estimation des paramètres et influence sur la réponse 35
4.3. Mesures de performances du modèle 37
4.3.1. Test de Wald 38
4.3.2. Mesures de performances intuitives 40

5. Évaluation de la pertinence des indicateurs de risque 42
5.1. Introduction 42
5.2. Méthodes d'évaluation 43
5.2.1. Comparaison de deux moyennes 43
5.2.2. Optimisation du choix du seuil de coupure 44
5.3. Résultats 48
5.3.1. Étude des accidents isolés 49
5.3.2. Étude des accidents non-isolés 52
5.3.3. Étude croisée de deux indicateurs 54

6. Conclusion 57

Table des annexes 60

Annexes 61

Références 114

1. Introduction

1.1. L'IFSTTAR et le GRETTIA

L'IFSTTAR, Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux, est un institut de recherche public à caractère scientifique et technologique, qui intervient principalement dans la recherche au niveau des transports terrestres.

C'est un institut qui se situe au premier plan au niveau national ; il est impliqué de manière déterminante dans l'ensemble des décisions et des actions du PREDIT (programme de recherche et d'innovation dans les transports terrestres), qui est l'instance qui structure et synthétise l'ensemble de la recherche sur le sujet en France. Par ailleurs, l'IFSTTAR mène un grand nombre de projets de recherche en partenariat avec de grandes agences telles que l'ADEME (Agence De l'Environnement Et de la Maîtrise de l'Énergie) ou l'ANR (Agence Nationale de la Recherche).

L'IFSTTAR jouit aussi d'une stature internationale, dans la mesure où il participe à plusieurs instances et associations européennes et internationales telles que l'ECTRI (conférence européenne des instituts de recherche sur les transports), le FEHRL (forum européen des laboratoires de recherche routière), ou l'AIPCR (Association mondiale de la route). La majorité des projets auxquelles l'IFSTTAR est intégré est financée par des fonds européens.

Une des originalités de l'IFSTTAR est la multidisciplinarité de ses équipes et de ses 26 laboratoires de recherche. Nous y trouvons associées les disciplines suivantes : mathématiques-statistiques, informatique, physique des matériaux, biomécanique, économie, écologie ainsi que des thématiques liées aux sciences humaines : sociologie des transports, psychologie de la conduite etc. Cette multidisciplinarité lui permet d'aborder de manière globale les principales problématiques liées au transport.

L'IFSTTAR mène ses activités de recherche à travers trois grands axes :

- Mobilité, énergie et environnement

  La recherche actuelle sur ce thème vise en premier lieu à mieux évaluer et à mieux comprendre les coûts en énergie des systèmes de transports, ainsi que leur impact sur l'environnement. Par ailleurs, des travaux sont mis en œuvre pour étudier des problématiques spécifiques liées à la mobilité, telles que le trafic en milieu urbain ou le transport des marchandises. Une partie des travaux consiste aussi à évaluer l'impact des décisions publiques en matière de gestion des transports sur les bilans énergétiques, environnementaux et sociaux, dans une logique de développement durable.

- Qualité, sécurité et optimisation des systèmes de transport

  À partir d'une approche globale prenant en compte à la fois les enjeux de sécurité routière et les enjeux environnementaux du trafic, la recherche dans ce domaine tente d'évaluer la qualité des systèmes de transports en termes de service rendu, de sécurité des personnes, de rendement énergétique et d'impact sur l'environnement.

- Transport et santé

  Ce domaine d'étude est principalement celui de l'accidentologie ; il s'agit d'étudier les accidents et leurs causes pour apporter des solutions au problème de la sécurité routière en France, qui est un problème majeur de santé publique depuis des décennies. Une autre partie de la recherche est consacrée à l'étude de l'impact sur la santé de la pollution générée par les systèmes des transports (principalement les voitures).

L'IFSTTAR est implémenté sur sept sites différents : Lille-Villeneuve d'Ascq, Paris, Marne-la-Vallée, Versailles-Satory, Nantes, Lyon-Bron, Marseille-Salon de Provence. Le siège actuel se trouve à Paris, mais une nouvelle infrastructure est actuellement en construction à Marne-la-Vallée, qui regroupera sur un même site l'ensemble des activités de l'IFSTTAR.

Au site de Marne-la-Vallée où j'ai réalisé mon stage, j'ai intégré le laboratoire de recherche « GRETTIA » (Génie des Réseaux de Transport Terrestres et Informatique Avancée). Il s'agit d'une unité de recherché créée en décembre 2010, dont les principales activités se concentrent sur l'étude du domaine routier, du transport collectif et des transports guidés. Les équipes de recherches développent par exemple des modèles de trafic permettant de modéliser l'ensemble des phénomènes routiers tels que la circulation au niveau d'un carrefour ou les situations de congestion (bouchons). Leur mission est de concevoir des systèmes intelligents de transports, de les gérer et de les optimiser.

Au cours de mon séjour à l'IFSTTAR, j'ai pu assister à plusieurs conférences de chercheurs. La première d'entre elle concernait l'élaboration d'un programme destiné aux usagers de la route et intégrable sur les smartphones, qui calcule le temps de parcours d'un trajet donné grâce à des données de trafic en temps réel. La seconde évoquait les problématiques de sécurité dans le RER et le métro face aux attaques terroristes.

1.2. Présentation de la problématique du stage

Le but de ce stage est de tenter de mettre en évidence le lien entre la survenue d'accidents et les caractéristiques cinématiques de véhicules routiers.

Le risque d'accident est lié à la fois à la situation globale du trafic, et à la situation cinématique individuelle du conducteur, qui résulte de son comportement au volant. Pour modéliser le risque d'accident, nous construirons des variables appelées « indicatrices de risque » qui permettent d'évaluer un risque potentiel d'accident pour un groupe de véhicules donné, à partir des caractéristiques cinématiques des véhicules tels que les temps intervéhiculaires ou les vitesses. Le cœur de ce stage est de mettre au point des procédures statistiques permettant de valider ou d'infirmer la pertinence de ces indicateurs. Ces méthodes de validation s'appuient principalement sur un puissant outil statistique, à savoir la régression logistique, que nous introduirons de manière assez générale.

Ce type d'étude pourrait typiquement aboutir à la mise en œuvre de politiques de sécurité routière. En effet, si certains de ces indicateurs se révélaient être pertinents pour discriminer de manière significative la survenue des accidents ou la survenue d'un certain type d'accident, cela permettrait d'informer de manière précise les usagers de la route des risques qu'ils encourent lors de leurs déplacements, et ainsi de faire de la prévention. Par exemple, si nous parvenons à détecter une corrélation entre des temps intervéhiculaires courts et la survenue d'accidents, cela justifiera des mesures de prévention du type : « Vous roulez trop près du véhicule précédent ».

1.3. Présentation des données

Nous disposons de deux types de fichiers de données de base : les données de trafic et les données d'accident.

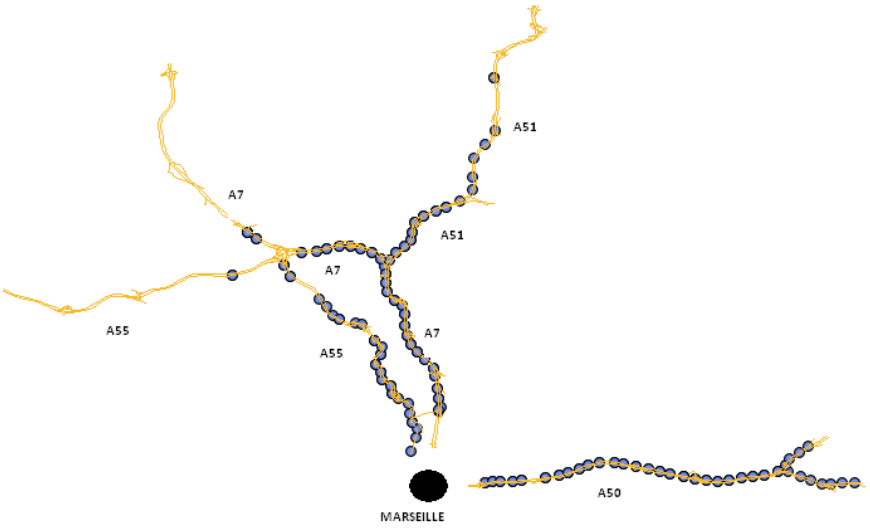
1.3.1. Les données de trafic

Les données de trafic fournissent une série de mesures individuelles (dites *microscopiques*) de trafic. Il s'agit des mesures de la longueur du véhicule, de sa vitesse, et de son temps de passage au niveau de l'appareil de mesure.

Sur les autoroutes, les principaux appareils de mesure sont les boucles magnétiques ; il s'agit de capteurs électriques implémentés dans la chaussée qui sont sensibles à la variation du champ magnétique produite par le passage d'un véhicule. Ces boucles magnétiques détectent donc le passage des véhicules, et enregistrent les données.

Boucle magnétique

Les données de trafic dont nous disposons proviennent des mesures de 180 capteurs de ce type pendant une durée d'environ un an, du 17 mai 2009 au 31 mai 2010, sur une partie du réseau autoroutier « Marius » autour de Marseille, qui réunit quelques sections des autoroutes A7, A50, A51 et A55. Sur le graphe ci-dessous, les capteurs sont représentés par des points bleus.



Les données sont classées, pour un jour, une heure et une minute fixés, selon le temps de passage des véhicules devant un capteur quelconque, autoroute, voies et sens de circulation confondus. Les fichiers de données sont hiérarchisés de la manière suivante : le dossier source « Marius » contient deux sous-dossiers « 2009 » et « 2010 » correspondant aux années 2009 et 2010. Le dossier « 2009 », par exemple, contient une liste de sous-dossiers de « 06 » à « 12 » correspondant aux mois. Puis, le dossier « 07 », par exemple, contient des milliers de fichiers du type « 031454.dat », qui stocke l'ensemble des données individuelles des véhicules qui ont été détectés le 3 juillet entre 14h54 et 14h55 par un capteur quelconque, sur une autoroute quelconque, dans un sens de circulation quelconque et sur une voie de circulation quelconque (voie de droite, voie centrale ou voie de gauche). Nous avons fait figurer dans l'annexe A.1. un extrait du fichier « 031454.dat ».

Cette base de données est très volumineuse ; elle correspond aux données individuelles d'environ 100 millions de véhicules, ce qui représente en termes de volume de stockage, environ 150 giga-octets de données de trafic. C'est la première fois, au GRETTIA, qu'une étude de ce type, basée sur des données microscopiques, est mise en place. Généralement, les données utilisées sont de type macroscopique, c'est-à-dire qu'elles ne fournissent pas les valeurs des variables de trafic véhicules par véhicules, mais des valeurs moyennes, sur des périodes d'une à six minutes, associées à des groupes de véhicules. Le traitement de données microscopiques pose des problèmes spécifiques, comme nous le verrons dans le chapitre 3.

1.3.2. Les données d'accidents

Les données d'accidents sont issues des BAAC (Bulletins d'Analyse des Accidents Corporels), constitués par les forces de l'ordre lors des accidents.

Les données se composent de quatre fichiers Excel distincts, qui fournissent toutes les informations disponibles liées aux 541 accidents qui se sont produits entre janvier 2009 et novembre 2010 sur les sections d'autoroute que nous étudions. Le nombre d'accidents correspondant aux données de trafic étudiées (entre mai 2009 et juin 2010) s'élève à 289. Des extraits de ces quatre fichiers ont été placés dans l'annexe A.2.

Les deux premiers fichiers sont composés d'autant de lignes que d'accidents, et stockent les informations élémentaires telles que la date de l'accident, l'autoroute sur laquelle il a eu lieu et le point routier correspondant. Il y a une quantité d'autres informations disponibles, telles que la luminosité, les conditions météorologiques, l'état de la surface de la route etc.

Le troisième fichier d'accident fournit, pour chaque accident, des informations propres à chacun des véhicules impliqués dans l'accident. Ainsi, nous avons accès au type de véhicule impliqué (voiture, deux-roues, poids-lourds, etc.), à l'ancienneté du véhicule etc. Ce qui est plus intéressant est la présence de plusieurs variables nous renseignant sur le contexte dans lequel l'accident a eu lieu : la manœuvre principale avant l'accident (changement de voie, insertion à partir d'une bretelle, etc.), le point de choc initial, la présence ou non d'un obstacle mobile ou fixe heurté etc. Ce troisième fichier contient aussi une information essentielle à la localisation de l'accident par rapport à nos données de trafic, à savoir le sens de circulation (selon le sens des points routiers croissants ou décroissants).

Enfin, le quatrième fichier d'accident nous renseigne, pour un accident donné et un véhicule concerné fixé, sur les personnes blessées (blessés légers, blessés grave ou morts) présentes dans le véhicule. Nous disposons alors de données civiles (âge, sexe, etc.) rendues anonymes, de données d'alcoolémie, de la gravité des blessures et, entre autres, d'une variable appelé « facteur lié à l'usager » qui précise si le conducteur a, par exemple, eu un malaise, ou si son attention a été perturbée.

Ainsi, à partir de ces informations, nous pouvons donc à la fois localiser précisément l'accident dans le temps et dans l'espace, et aussi repérer les accidents dont il apparait clairement que l'origine n'est pas liée à des caractéristiques cinématiques particulières. Par exemple, le conducteur a pu subir un malaise et générer un accident, sans que son comportement au volant ne soit particulièrement accidentogène. Ce type d'accident est absolument impossible à expliquer et à prédire de manière statistique, et ils risquent de fausser la significativité de nos indicateurs de risque ; ainsi, pour ne pas fausser l'étude, il serait bon de ne pas inclure ces accidents dans notre étude.

Un accident est un phénomène très rare : la probabilité qu'un accident se produise est de l'ordre de 10^{-8}. Ainsi, sur la portion d'autoroute étudiée en une année et correspondant à nos données de trafic, nous ne retiendrons que 289 accidents.

1.4. Stratégie d'attaque

Notre but est de tenter de mettre en évidence le lien entre la survenue des accidents et les caractéristiques cinématiques des véhicules impliqués. Or, nous ne disposons pas des boîtes noires des accidents ; nous n'avons pas donc accès aux caractéristiques cinématiques individuelles des véhicules impliqués. L'idée est alors de regarder, pour chaque accident, la situation cinématique « moyenne » des véhicules avant l'accident.

Afin de modéliser le risque d'accident, nous allons construire dans le chapitre 2. des variables appelées « indicateurs de risque », calculables à partir des variables individuelles de trafic que nous introduirons.

Concrètement, nous allons diviser chaque journée en 240 cycles de 6 minutes, et nous allons calculer la valeur de ces indicateurs de risque sur chaque cycle de 6 minutes. Certains de ces cycles correspondront à des situations d'avant-accident. Le cœur de notre étude, qui sera l'objet du chapitre 5., consistera à mettre au point des méthodes de validation de la pertinence de ces indicateurs de risque, en regardant s'ils prennent des valeurs significativement différentes, c'est-à-dire significativement « à risque », en cas d'accident. Ces méthodes statistiques de validation s'appuient principalement sur les outils de la régression logistique, présentée dans le chapitre 4.

Nous serons rapidement confrontés à des problèmes informatiques de temps de calcul, qui nous obligeront à ne traiter qu'une petite partie des données de trafic et d'accidents. Nous ne parviendrons donc pas à approfondir notre étude comme nous le souhaitions au départ ; des problèmes de significativité statistique vont apparaître, à cause du trop faible nombre d'accidents inclus dans l'étude. Pour l'étude des accidents en eux-mêmes, nous pourrons difficilement remédier à ce problème. Cependant, pour illustrer de manière satisfaisante nos méthodes d'optimisation, nous ne considérerons plus les situations réelles d'accidents, dont le nombre est trop faible, mais les situations « à risque d'accidents » par rapport à un indicateur donné. Ces notions seront précisées à la fin du chapitre 5.

2. Les indicateurs de risque

2.1. Introduction

Nous allons définir deux types de variables : les variables microscopiques, que l'on peut construire à partir des données individuelles de chaque véhicule et qui témoignent du comportement individuel du conducteur et des caractéristiques cinématiques locales du véhicule, et les variables macroscopiques, qui permettent de rendre compte des caractéristiques globales du trafic sur une période donnée. Nous élaborerons alors des indicateurs de risque pertinents à partir de ces diverses variables de trafic.

Par ailleurs, il s'agit de distinguer deux grandes catégories d'accidents :

- Les accidents isolés, c'est-à-dire n'impliquant qu'un seul véhicule : ce type d'accident a souvent été étudié, ils se produisent généralement la nuit, à vitesse élevée. Seule la variable « vitesse moyenne » permettrait ainsi de discriminer ces situations d'accident. Dans la suite, pour ce type d'accident, la variable indicatrice du type d'accident codera la modalité « Seul ».

- Les accidents non-isolés, c'est-à-dire impliquant plusieurs véhicules insérés dans un peloton de véhicules. Les principaux indicateurs que nous développerons dans la suite sont destinés à mesurer le risque de survenue de ce type d'accident. La variable indicatrice du type d'accident codera alors la modalité « PasSeul ».

2.2. Les variables microscopiques de trafic

Nous construirons deux types de variables microscopiques de trafic : les variables microscopiques individuelles, associées aux véhicules et aux mesures individuelles, et les variables microscopiques agrégées, construites à partir des précédentes, qui seront associées aux cycles de 6 minutes.

2.2.1. Les variables microscopiques individuelles

Tout d'abord, présentons la construction des principales variables microscopiques individuelles dont nous allons nous servir. Ces variables microscopiques sont élaborées à partir des données individuelles de trafic ; nous pouvons donc associer à chaque véhicule une série de variables microscopiques individuelles, qui vont quantifier le risque individuel du véhicule.

Fixons les notations suivantes :

- Le véhicule $i-1$ est le véhicule de devant, le véhicule i celui qui le suit
- t_{i-1} et t_i : les temps de passage respectifs, en secondes, des véhicule $i-1$ et i au niveau du capteur
- v_{i-1} et v_i : les vitesses respectives, en m/s, des véhicule $i-1$ et i au niveau du capteur
- L_{i-1} et L_i : les longueurs respectives, en m, des véhicule $i-1$ et i

Les variables microscopiques individuelles sont les suivantes :

- Le Temps InterVéhiculaire (TIV) : il s'agit de la différence entre les temps de passage au niveau du capteur de deux véhicules consécutifs circulant sur la même voie. C'est la mesure du temps séparant les deux véhicules.

$$TIV_i = t_i - t_{i-1}$$

  ⇨ Le TIV est a priori indicateur d'un risque individuel lorsqu'il prend des valeurs faibles.

- La Vitesse Relative (VR) : il s'agit de la différence des vitesses entre le premier et le second véhicule.

$$VR_i = v_i - v_{i-1}$$

  ⇨ A priori, plus la vitesse relative est élevée, plus le risque d'accident est important.

- Le Temps d'Arrêt (TA) : c'est le temps nécessaire à l'arrêt du véhicule. Il se calcule de manière simple en sommant le temps de réaction du conducteur au temps de freinage. Nous supposons que le freinage se réalise à décélération constante $\gamma = 6.25\ m.s^{-2}$. Le temps de freinage vaut alors : $T_f = {}^{v_i}/_{\gamma}$. En considérant que le temps de réaction T_r vaut 1 seconde, nous obtenons la formule suivante :

$$TA_i = 1 + {}^{v_i}/_{\gamma}$$

  Par exemple : $TA_i = 2.8s$ si $v_i = 40\ km/h$, $TA_i = 5s$ si $v_i = 90\ km/h$

  ⇨ Le risque d'accident a priori est d'autant plus important que le temps d'arrêt est grand.

- Le Temps de collision (TTC, de l'anglais Time To Collision) : lorsqu'un véhicule roule à une vitesse supérieure à celle du véhicule devant lui, si aucun des deux véhicules ne change de vitesse, une collision se produit un bout d'un temps appelé « Temps de collision ». La démonstration de la formule ci-dessous est placée dans l'annexe B.1. :

$$TTC_i = \frac{v_{i-1} * TIV_i}{VR_i}$$

  Par exemple, $TTC_i = 6.7s$ si $v_{i-1} = 100\ km/h$, $VR_i = 30\ km/h$ et $TIV_i = 2s$

  ⇨ Le risque d'accident a priori est d'autant plus important que le temps de collision est faible.

- Le PICUD (de l'anglais Potential Index for Collision with Urgent Deceleration) : il s'agit d'une variable microscopique fondée sur la différence des positions entre les deux véhicules, dans le cas où le premier freine avec une décélération constante $\gamma = 6.25\ m.s^{-2}$, et le second freine avec la même décélération après un temps de réaction T_r, estimé à 1 seconde.
  L'idée est de dire qu'une collision se produit si la distance d'arrêt $D_i^{arrêt}$ du second véhicule est supérieure à la somme de la distance d'arrêt $D_{i-1}^{arrêt}$ du premier véhicule et de la distance entre l'avant du second véhicule et l'arrière du premier véhicule. Le PICUD se calcule par la formule suivante, dont la démonstration se trouve en annexe B.2. :

$$PICUD_i = \frac{v_{i-1}^2 - v_i^2}{2 * \gamma} + TIV_i * v_{i-1} - v_i * T_r - L_{i-1}$$

  Si le PICUD est positif, cela signifie que le véhicule de derrière est suffisamment éloigné du premier pour pouvoir freiner à temps et éviter la collision avec le premier véhicule, si celui-ci vient à freiner de façon brutale.
  ⇨ Un PICUD négatif est ainsi révélateur d'un risque d'accident.

- Le PICUDBIS : Lorsque les vitesses v_{i-1} et v_i sont élevées et proches l'une de l'autre et que TIV_i est très faible, les termes $\frac{v_{i-1}^2 - v_i^2}{2 * \gamma}$ et $TIV_i * v_{i-1}$ sont négligeables devant $-v_i * T_r$, qui prend alors des valeurs très négatives.
  Nous proposons de modifier artificiellement la formule du $PICUD_i$ en supprimant ce terme, afin que le variable ne prenne pas trop facilement des valeurs très négatives, c'est-à-dire très « à risque ». Nous construisons alors la variable $PICUDBIS_i$ par la formule :

$$PICUDBIS_i = \frac{v_{i-1}^2 - v_i^2}{2 * \gamma} + TIV * v_{i-1} - L_{i-1}$$

  ⇨ Un PICUDBIS négatif est révélateur d'un risque d'accident.

2.2.2. Les variables microscopiques agrégées

Comme nous l'avons déjà expliqué, nous allons diviser la journée en 240 cycles de 6 minutes. Pour modéliser le risque d'accident « global » pour chacun des cycles de 6 minutes, nous allons introduire des variables microscopiques « agrégées », construites à partir des variables microscopiques individuelles vues précédemment. Concrètement, pour chaque cycle de 6 minutes, nous allons alors calculer la valeur de chacune des variables microscopiques agrégées, qui vont quantifier le risque d'accident au niveau du cycle.

Pour une variable microscopique individuelle fixée, nous allons considérer deux types de variables microscopiques agrégées, construites à partir de cette variable microscopique. Le 1er type est construit à partir d'une moyenne, le 2nd à partir d'une proportion.

Le 1er type de variable microscopique agrégée est construit à partir de la moyenne des valeurs de la variable microscopique individuelle concernées par le cycle de 6 minutes. Notre but étant de quantifier le niveau de risque à l'aide de cette variable agrégée, nous prendrons uniquement en compte, dans le calcul de la moyenne, les valeurs « à risque » de la variable microscopique individuelle. Par exemple, si l'on considère la variable $PICUD_i$ (où i décrit la série des véhicules), pour calculer la variable microscopique agrégée de 1er type correspondante, nous ferons la moyenne des valeurs $PICUD_i$ négatives, puisque seules les valeurs négatives sont indicatrices d'un risque potentiel d'accident. Nous obtenons alors l'indicateur microscopique agrégé $PICUD_j^{moy}$ (où j décrit les cycles de 6 minutes, du 1er au 240ème) qui se calcule de la manière suivante :

$$PICUD_j^{moy} = \frac{\sum_{\substack{i \in cycle(j) \\ tq PICUD_i < 0}} |PICUD_i|}{NbreVeh_j}$$

où $NbreVeh_j$ est le nombre de véhicules appartenant au cycle j.

Notons bien que nous avons mis une valeur absolue dans la formule, pour obtenir un résultat positif.

Cette variable agrégée va alors être indicatrice du risque d'accident global du cycle lorsqu'elle prend des valeurs élevées.

Le 2nd type de variable agrégée se construit en utilisant la proportion de véhicules appartenant au cycle et présentant des valeurs individuelles à risque. Par exemple, en reprenant l'exemple du $PICUD_i$, nous allons regarder, pour un cycle j donné, la proportion de véhicules dans ce cycle dont le $PICUD_i$ est négatif. Nous obtenons ainsi l'indicateur microscopique agrégé $pPICUD0_j$ qui se calcule de la manière suivante :

$$pPICUD0_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{PICUD_i < 0}}{NbreVeh_j}$$

Plus il y aura de véhicules dans le cycle dont le $PICUD_i$ est négatif, c'est-à-dire « à risque », plus le risque d'accident global associé au cycle sera élevée. La variable $pPICUD0_j$ sera donc indicatrice de risque lorsqu'elle prendra des valeurs élevées.

Les variables microscopiques agrégées que nous avons créées sont représentées dans le tableau suivant. L'indice i décrit l'ensemble véhicules, l'indice j l'ensemble des cycles. $NbreVeh_j$ représente le nombre de véhicules appartenant au cycle j.

Variable microscopique individuelle	Variable microscopique agrégée de 1er type (moyenne)	Variable microscopique agrégée de 2nd type (proportion)
TIV_i	$TIV_j^{moy} = \frac{\sum_{i \in cycle(j)} TIV_i}{NbreVeh_j}$	$\begin{cases} pTIV0.5_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{TIV_i<0.5}}{NbreVeh_j} \\ pTIV1_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{TIV_i<1}}{NbreVeh_j} \\ pTIV2_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{TIV_i<2}}{NbreVeh_j} \end{cases}$
VR_i	$VR_j^{moy} = \frac{\sum_{i \in cycle(j)} VR_i}{\substack{tq\ VR_i>0 \\ NbreVeh_j}}$	
$PICUD_i$	$PICUD_j^{moy} = \frac{\sum_{\substack{i \in cycle(j) \\ tqPICUD_i<0}} \lvert PICUD_i \rvert}{NbreVeh_j}$	$\begin{cases} pPICUD0_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{PICUD_i<0}}{NbreVeh_j} \\ pPICUD10_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{PICUD_i<-10}}{NbreVeh_j} \\ pPICUD20_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{PICUD_i<-20}}{NbreVeh_j} \end{cases}$
$PICUDBIS_i$	$PICUDBIS_j^{moy} = \frac{\sum_{\substack{i \in cycle(j) \\ tqPICUDBIS_i<0}} \lvert PICUDBIS_i \rvert}{NbreVeh_j}$	$pPICUDBIS0_j = \frac{\sum_{i \in cycle(j)} \mathbb{1}_{PICUDBIS_i<0}}{NbreVeh_j}$

Sur les conseils de mon maître de stage, devant les difficultés informatique en termes de temps de calcul, nous avons décidé de ne pas prendre en compte les variables « Temps d'arrêt » (TA) et « Temps de collision » (TTC) qui modélisent a priori moins bien le risque d'accident que le PICUD, qui est une variable plus précise et mieux élaborée.

Notons par ailleurs que, dans la construction de la variable microscopique agrégée de 1er type relative à la vitesse relative VR_i, nous faisons la moyenne sur les VR_i positives qui sont les seules qui nous intéressent. En effet, une VR_i négative ne correspond pas à une situation de risque.

Nous avons fait le choix de prendre trois seuils différents pour la construction des variables agrégées de 2nd type relatives à TIV_i et $PICUD_i$. En effet, nous pensons a priori que les accidents seraient corrélés aux valeurs extrêmes de nos indicateurs de risque, c'est-à-dire aux queues de distributions. Dans cette optique, nous avons construit les variables $pTIV0.5_j$, $pTIV1_j$, $pPICUD10$ et $pPICUD20$, qui vont prendre des valeurs bien plus faibles, c'est-à-dire moins facilement « à risque » que, par exemple, $pTIV2_j$ et $pPICUD0_j$.

Les 11 variables microscopiques agrégées obtenues sont alors candidates au statut d' « indicatrices de risque » ; il s'agira, dans le chapitre 5., de valider ou d'infirmer, à l'aide de méthodes statistiques adéquates, la pertinence de la discrimination qu'elles réalisent entre les situations d'accident et les situations de non-accident.

2.2.3. Corrélations entre variables microscopiques

Les graphes des corrélations entre les variables microscopiques agrégées sont disponibles dans l'annexe B.3. Ces graphes ont été réalisés à partir des données correspondant à une journée de trafic, issues du fichier « M1as1000430C20090711CYCLES.txt ». Nous préciserons la construction de ce type de fichiers de données dans le chapitre 3. Nous disposons donc de 240 observations indépendantes.

Nous pouvons faire plusieurs remarques :

- Etant donné que ces variables microscopiques agrégées sont censées mesurer un même risque d'accident (dans le cas où elles ont toutes la même qualité en termes d'indicateurs de risque), elles devraient évoluer l'une par rapport à l'autre de manière globalement équivalente. On constate, en effet, que cette propriété est vérifiée sur l'ensemble des graphes, hormis ceux incluant les variables $PICUDBIS_j^{moy}$ ou $pPICUD20_j$. Le comportement de ces deux variables semble particulier, mais cela ne préjuge pas pour autant de leur qualité dans la discrimination des accidents.
- Par ailleurs, il apparaît clairement que certaines de ces relations sont linéaires. Si l'on parvenait à ajuster une régression linéaire performante entre deux variables, cela signifierait que l'information contenue dans l'une est globalement contenue dans l'autre. Cela ne signifie pourtant pas que l'on pourrait se contenter d'étudier ultérieurement la qualité de la discrimination d'une seule de ces deux variables.

Tentons, par exemple, d'ajuster une régression linéaire simple entre les variables $pPICUD0_j$ et $pTIV2_j$:

$$pPICUD0_j = \beta_0 + \beta_1 * pTIV2_j + \varepsilon_j$$

en notant :

- n le nombre d'observations, ici n = 240, jϵ [1,..,240]
- $\beta_0 \; et \; \beta_1$ les coefficients de la régression
- ε_j les erreurs

Les hypothèses à vérifier pour valider ce modèle sont les suivantes :

- Espérance nulle : Les ε_j sont centrées
- Normalité : Le vecteur aléatoire formé des ε_j est gaussien
- Indépendance : Les ε_j sont indépendantes
- Homoscédasticité : Les variances des ε_j sont égales

Nous allons donc vérifier une à une ces hypothèses pour pouvoir continuer notre étude. Ces hypothèses concernant les erreurs doivent être vérifiées à partir de leurs observations, c'est-à-dire à partir des résidus. En effet, les erreurs elles-mêmes sont inconnues, et peuvent être approchées par les résidus. Les sorties R sont disponibles dans l'annexe B.4.

- Espérance nulle : par construction du modèle linéaire (puisque la constante fait partie du modèle), les résidus sont d'espérance nulle.
- Normalité : notons que, comme il y a plus de 30 observations, le test de Shapiro-Wilk peut être mis en œuvre. Le test ne permet pas de déceler un problème de normalité des données ; en effet, la p-value étant strictement supérieure à α=5%, le test conserve l'hypothèse nulle de normalité des données. Cette décision est prise avec le risque de seconde espèce β qu'il faudrait évaluer à l'aide d'une étude de la puissance. Par ailleurs, le tracé du QQ-plot montre un assez bon ajustement des résidus à une loi normale.
- Indépendance : nous faisons l'hypothèse que les cycles de 6 minutes sont indépendants deux à deux. Les mesures associées sont donc supposées indépendantes.
- Homoscédasticité : nous constatons sur le graphe des résidus que les points sont globalement répartis de manière homogène autour d'une valeur centrale, ce qui est caractéristique de l'homoscédasticité. L'hypothèse d'homoscédasticité est donc considérée comme valide.

Toutes les hypothèses du modèle linéaire sont donc vérifiées.

Nous voyons, dans le « summary » que R renvoie, que la régression linéaire s'ajuste parfaitement : le coefficient R^2 vaut environ 94%, ce qui signifie que presque toute la variation de $pPICUD0_j$ est expliquée par $pTIV2_j$.

Par ailleurs, le test usuel de Student permettant de juger de l'influence de $pTIV2_j$ est significatif au seuil α=5%. Nous décidons donc, avec un risque d'erreur de première espèce de α=5%, que le coefficient de régression associé à $pTIV2_j$ est non-nul. L'estimation du paramètre β_1 vaut environ 0.59.

Le test de Student concernant la constante est non-significatif au seuil α=5%. Nous décidons donc que $\beta_0 = 0$. Le risque d'erreur associé à cette décision est un risque de seconde espèce, qu'il faudrait évaluer à l'aide d'une étude de la puissance.

On obtient ainsi la relation de régression suivante :

$$pPICUD0_j = 0.59 * pTIV2_j + \varepsilon_j$$

2.3. Les variables macroscopiques de trafic

2.3.1. Construction

A présent, présentons les variables macroscopiques de trafic que nous allons étudier. Elles témoignent de l'état ambiant du trafic sur les périodes de 6 minutes. Introduisons les variables macroscopiques suivantes :

- Le débit moyen : c'est le nombre de véhicules appartenant au cycle que l'on divise par l'unité de temps, c'est-à-dire 6 minutes (360 secondes). Cette variable témoigne donc de la répartition des véhicules dans le temps. Le débit moyen s'écrit alors :

$$débit_j^{moy} = \frac{NbreVeh_j}{360}$$

En pratique, la grandeur considérée est le débit moyen horaire ; il s'agit simplement de ramener le débit moyen sur une heure (3600 secondes). La formule du débit horaire moyen est alors :

$$débit_j^{moy} = \frac{NbreVeh_j}{\frac{360}{3600}} = NbreVeh_j * 10$$

- Le taux d'occupation : c'est une variable qui mesure, pour un cycle de 6 minutes donné, la proportion de temps durant laquelle la boucle magnétique est occupée. La formule est donc la suivante :

$$TauxOcc_j = \frac{\sum_{i \in cycle(j)} TP_i}{360}$$

où TP_i représente la durée de passage du véhicule i sur la boucle. Si l'on note respectivement L_i et v_i la longueur et la vitesse du véhicule i au niveau de la boucle, le temps de passage TP_i du véhicule i sur la boucle est : $TP_i = \frac{L_i + l}{v_i}$, où $l = 1m$ est la longueur de la boucle. Ainsi, la formule pour calculer le taux d'occupation à partir des données de trafic est :

$$TauxOcc_j = \frac{\sum_{i \in cycle(j)} \frac{L_i + l}{v_i}}{360}$$

Le taux d'occupation permet de mesurer la congestion du trafic ; on considère que, au-delà d'un taux d'occupation de 15%, le trafic est congestionné.

- La vitesse moyenne : elle se calcule simplement par la formule suivante :

$$vitesse_j^{moy} = \frac{\sum_{i \in cycle(j)} v_i}{NbreVeh_j}$$

Le lecteur trouvera dans l'annexe B.5. les graphes, sur une journée de données de trafic, des variables macroscopiques introduites.

2.3.2. Corrélations entre variables macroscopiques

Les trois graphes, disponibles dans l'annexe B.6., représentent les corrélations entre les trois variables macroscopiques débit moyen, taux d'occupation et vitesse moyenne. Pour réaliser ces graphes, nous nous sommes servis des données d'un capteur sur 15 jours, comprenant beaucoup de situations de congestion. Celles-ci sont repérées par des valeurs de taux d'occupation supérieures à 15%.

Nous pouvons identifier deux régimes de trafic : le régime de trafic libre et le régime de trafic congestionné. Voici une tentative d'explication de la dynamique du trafic, réalisée à l'aide de ces trois graphes.

Lorsque le taux d'occupation est faible (inférieur à 15%) et le débit peu élevé (inférieur à 1300, c'est-à-dire à 130 véhicules par cycles de 6 minutes), le trafic est fluide. La vitesse moyenne des véhicules, appelée alors vitesse libre, est élevée. Sur autoroute, la vitesse libre est généralement de l'ordre de 100 km/h, mais sur nos graphes, elle n'est que de 75-80 km/h, car la vitesse sur la portion de route étudiée est sans doute limitée à 90 km/h. Lorsque le trafic est fluide et que le débit augmente, le taux d'occupation augmente de façon linéaire. Quant à la vitesse moyenne, elle baisse sensiblement mais reste quasiment constante, égale à la vitesse libre. Lorsque le débit atteint environ 130 véhicules par cycles de 6 minutes, et le taux d'occupation une valeur de 15%, les véhicules commencent à se gêner et les conducteurs doivent adapter leur conduite en fonction des autres ; il s'agit d'un début de situation de congestion. Les vitesses moyennes chutent alors rapidement (sur nos graphes, elles passent de 75 km/h à 40km/h), ce qui a pour conséquence de faire stagner le débit. Si la congestion s'intensifie, le taux d'occupation augmente, les vitesses moyennes chutent encore plus et, en conséquence, le débit diminue. Il s'agit alors de la situation de congestion intense, où la vitesse moyenne des véhicules est très faible (de l'ordre de 15 km/h) et où le taux d'occupation, très élevé (supérieur à 40%), est associé à de faibles valeurs du débit.

2.4. Corrélations entre variables microscopiques et variables macroscopiques

2.4.1. Introduction

La grande majorité des études de trafic est réalisée à partir de données macroscopiques, car il est rare de disposer de données microscopiques. Et, ainsi que nous le verrons dans le chapitre 3., leur traitement informatique est long et délicat. Cependant, les données microscopiques contiennent, par nature, plus d'informations que les données macroscopiques. Il serait donc intéressant de vérifier sur nos données si les variables microscopiques dont nous disposons contiennent en effet plus d'informations que les variables macroscopiques, en termes de mesures de risque d'accident. Avant cela, nous tenterons de mettre en évidence, de manière empirique, la validité d'une formule théorique bien connue en théorie du trafic entre débit moyen et temps intervéhiculaire moyen.

Ces études sont basées sur le fichier de données de trafic « M1as1000430C20090711CYCLES.txt » correspondant à une journée de données, pour un capteur, une voie et un sens de circulation fixés.

2.4.2. Lien entre débit moyen et temps intervéhiculaire moyen

Relation théorique

Il existe une relation théorique non aléatoire très simple entre le débit moyen $débit^{moy}$ et la moyenne des temps intervéhiculaires TIV^{moy}. Cette formule établit que, sur une période j de durée T donnée, le débit moyen et le temps intervéhiculaire moyen sont inverses l'un de l'autre :

$$débit_j^{moy} = \frac{1}{TIV_j^{moy}}$$

La démonstration de cette relation est assez aisée. Par définition du débit moyen, nous avons : $débit_j^{moy} = \frac{N}{T}$, où N est le nombre de véhicules appartenant à la période T.

Par ailleurs, en notant t_k, k∈ [1, …, N], les temps de passage respectifs des véhicules devant le capteur, le temps intervéhiculaire moyen s'écrit :

$$TIV_j^{moy} = \frac{\sum_{i \in cycle(j)} TIV_i}{NbreVeh_j} = \frac{(t_2 - t_1) + (t_3 - t_2) + \cdots + (t_N - t_{N-1})}{N} = \frac{t_N - t_1}{N}$$

Si la période T commence au passage du 1er véhicule, et finit au passage du dernier véhicule, nous avons exactement : $T = t_N - t_1$.

Et donc : $TIV_j^{moy} = \frac{T}{N}$. Nous obtenons donc bien la formule annoncée.

En pratique, comme nous le verrons juste après, étant donné que les périodes sont définies de manière fixe par intervalles de 6 minutes, le début et la fin de la période ne coïncident pas nécessairement avec le passage d'un véhicule devant le capteur. Ceci génère d'infimes variations entre $débit_j^{moy}$ et $\frac{1}{TIV_j^{moy}}$.

Vérification empirique

Nous nous proposons de vérifier l'exactitude empirique de cette formule, grâce aux données de trafic agrégées par cycles présentes dans le fichier « M1as1000430C20090711CYCLES.txt » (le chapitre 3. détaille la construction de ce type de fichiers de données). Le lecteur trouvera dans l'annexe B.7. les graphes représentant débit contre temps intervéhiculaires moyens.

En notant $InvTIV_j^{moy} = \frac{1}{TIV_j^{moy}}$, la formule se réécrit de la manière suivante :

$$débit_j^{moy} = InvTIV_j^{moy}$$

Pour valider cette formule, nous allons tenter d'ajuster la régression linéaire simple suivante :

$$InvTIV_j^{moy} = \beta_0 + \beta_1 * débit_j^{moy} + \varepsilon_j$$

en notant :

- n le nombre d'observations, ici n = 240, jϵ [1,..,240]
- β_0 et β_1 les coefficients de la régression
- ε_j les erreurs

Nous allons, tout d'abord, vérifier les hypothèses d'application du modèle linéaire. Les sorties R correspondantes sont disponibles dans l'annexe B.8.

- Espérance nulle : par construction du modèle linéaire (puisque la constante fait partie du modèle), les résidus sont d'espérance nulle.
- Normalité : notons que, comme il y a plus de 30 observations, le test de Shapiro-Wilk peut être mis en œuvre. La p-value étant largement inférieure à α=5%, le test rejette l'hypothèse nulle de normalité. Cette décision est prise avec le risque de première espèce $\alpha = 5\%$. Par ailleurs, le tracé du QQ-plot montre un mauvais ajustement des données à une loi normale. La normalité des données n'est donc pas validée.
- Indépendance : nous faisons l'hypothèse que les cycles de 6 minutes sont indépendants deux à deux. Les mesures associées sont donc supposées indépendantes.
- Homoscédasticité : Nous constatons que, sur le graphe des résidus, les points sont globalement répartis de manière homogène autour d'une valeur centrale, ce qui est caractéristique de l'homoscédasticité. L'hypothèse d'homoscédasticité est donc considérée comme valide.

Ainsi, toutes les hypothèses d'application du modèle linéaire, sauf l'hypothèse de normalité, sont donc considérées comme valides. Nous sommes bien conscients du fait que la validité de l'hypothèse de normalité est centrale pour l'exploitation des résultats fournis par la régression. Nous continuons cependant l'étude.

Le « summary » renvoyé par R indique que la régression linéaire s'ajuste parfaitement : le coefficient R^2 vaut 1, ce qui signifie que toute la variation de l'inverse des TIV moyens est expliquée par le débit moyen. Ceci indique que la relation reliant ces deux variables est déterministe, comme nous l'avons annoncé. Par ailleurs, les tests usuels de Student montrent que la constante du modèle n'a pas d'influence significative au seuil $\propto = 5\%$ sur la réponse, et que la variable $débit^{moy}$ quant à elle est bien significative. L'estimation du paramètre β_1 vaut 0.997, c'est-à-dire quasiment 1.

Ainsi, la relation de régression est la suivante :

$$InvTIV_j^{moy} = débit_j^{moy} + \varepsilon_j$$

Bien que toutes les hypothèses du modèle ne soient pas vérifiées, cette relation de régression indique clairement que la formule théorique énoncée précédemment est bien vérifiée. Un résultat contraire susciterait des doutes sur la validité des données ou sur leur utilisation.

2.4.3. Autres corrélations

Nous avons placé dans l'annexe B.9. les graphes représentant les relations entre les variables microscopiques agrégées et le débit moyen. Les graphes liant variables microscopiques et taux d'occupation ayant exactement la même allure, nous ne les avons pas fait apparaître.

Le but des variables microscopiques agrégées est de quantifier le risque que se produisent les accidents non-isolés. Le bon sens nous dit que ce risque devrait augmenter avec le débit : plus le nombre de véhicules circulant par cycles de 6 minutes est élevé, plus le risque de collision devrait être important. Si l'on analyse les graphes, on constate que les variables qui nous intéressent en premier abord, à savoir le TIV et le PICUD, confirment cette idée : $PICUD_j^{moy}$, $pPICUD0_j$, $pPICUD10_j$, $pTIV2_j$ et $pTIV1_j$ croissent avec le débit, et ce de manière globalement linéaire.

Les variables construites à partir du PICUDBIS ($PICUDBIS_j^{moy}$ et $pPICUDBIS0_j$), VR_j^{moy} ainsi que les variables correspondant aux queues de distribution et prenant de très faibles valeurs ($pPICUD20_j$ et $pTIV0.5_j$) n'ont pas ce type de comportement : lorsque le débit augmente, ces variables tendent à prendre des valeurs plus faibles. Il serait bon de les étudier plus en détails : leur comportement est étrange, et nous noterons, sur leurs graphes, la présence de lignes de niveaux témoignant peut-être de différents profils de conduite.

Nous pourrions tenter, encore une fois, d'ajuster des régressions linéaires performantes entre ces différentes variables et le débit. Par exemple, une régression linéaire entre $pTIV2_j$ et le débit moyen s'ajuste très bien, avec un coefficient R^2 valant 92.4%. Ainsi, que l'on évalue la pertinence, en termes de mesures de risque d'accident, du débit ou de la variable $pTIV2_j$, on obtiendrait globalement les mêmes résultats. Ainsi, cette variable microscopique ne semble pas

apporter, de manière générale, plus d'informations que la variable macroscopique débit. Néanmoins, il est possible que, en situations d'accidents, $pTIV2_j$ prenne des valeurs plus élevées que celle produite par la régression au niveau du débit correspondant à celui de l'accident.

Cependant, il faut noter que les données sur lesquelles nous nous sommes appuyés pour réaliser ces graphes ne comprenaient pas de cycles en régime de congestion. En situation de congestion, le comportement des variables qui nous intéresse est différent. En effet, comme nous l'apercevons sur les graphes présents dans l'annexe B.10., les variables $pTIV1_j$ et $pPICUD10_j$, par exemple, prennent des valeurs bien plus faibles qu'en situation ambiante. En situation de congestion, les vitesses et les vitesses relatives sont très peu élevées, ce qui explique les faibles valeurs de $pPICUD10_j$. Par ailleurs, dans ce contexte, les conducteurs régulent peut-être plus leur comportement en distance qu'en temps : pour une même distance intervéhiculaire, lorsque la vitesse diminue (à cause de la congestion), le temps intervéhiculaire augmente mécaniquement. Cela explique les faibles valeurs de $pTIV1_j$ en situation de congestion. Dans ce contexte, la corrélation entre variables microscopiques et débit est ainsi perdue. Ces variables microscopiques contiennent alors de manière générale, dans ce cas, plus d'informations que le débit seul.

C'est là que réside en effet tout l'intérêt des variables microscopiques construites. A la différence des variables macroscopiques, elles permettent de modéliser le comportement individuel des véhicules. En situation de congestion, les conducteurs sont plus attentifs, les vitesses et les vitesses relatives sont plus faibles et les conducteurs anticipent donc mieux. Le risque d'accident semble ainsi, a priori, diminué. On sait, en effet, que les accidents se produisent beaucoup plus fréquemment en début et en sortie de congestion qu'en situation de congestion intense.

Ainsi, les variables microscopiques présentent un double intérêt : en plus de contenir les informations de type macroscopique du trafic, elles modélisent le comportement individuel des véhicules, ce qui fait d'elles des candidates plus sérieuses que les variables macroscopiques au statut d'indicatrices de risque.

3. Traitement préparatoire des données

3.1. Introduction

Il s'agit, dans ce chapitre, de présenter les principales transformations préparatoires que nous avons réalisées sur les données. Ce prétraitement se fait en quatre étapes. Nous traitons, en premier lieu, à l'aide du logiciel SAS, les données d'accident afin d'en extraire les informations utiles à notre étude. Nous transformons ensuite les données de trafic, à l'aide de R notamment, et nous créons les variables de trafic et les indicateurs que nous allons étudier. Puis, nous établissons le lien entre données de trafic et données d'accidents. Enfin, nous agrégeons les milliers de fichiers de données obtenus en trois fichiers distincts, sur lesquels sera basée l'étude de l'évaluation de la pertinence des indicateurs de risque, présentée dans le chapitre 5.

Le traitement de ce type de données pose des problèmes spécifiques. En effet, en raison du volume très important des données de trafic (150 giga-octets), toutes les procédures de gestion ou de modifications des données doivent être automatisées, ce qui représente un important investissement en temps et en énergie dans l'élaboration de programmes informatiques. Par ailleurs, cela pose de considérables problèmes en termes de temps de calcul, comme nous le préciserons par la suite. Nous serons ainsi contraints de ne traiter qu'un mois de données de trafic : toute l'étude que nous ferons par la suite dans le chapitre 5. sera basée uniquement sur les données du mois de juillet 2009.

3.2. Traitement préparatoire des données d'accidents

Comme nous l'avons brièvement introduit lors de la présentation des données d'accidents, celles-ci se constituent en quatre fichiers Excel, à partir desquels nous allons créer à l'aide de SAS une table qui regroupe, pour chacun des 289 accidents qui ont eu lieu durant notre période d'étude, toutes les informations dont nous aurons besoin pour mener à bien notre étude statistique.

Parmi toutes les informations présentes dans les fichiers, celles que nous gardons ou créons sont de deux types :

- Celles qui permettent de localiser l'accident dans le temps et dans l'espace : la date, l'heure, l'autoroute, le sens de circulation et le point routier correspondant.
- Celles qui apportent des informations supplémentaires sur le contexte de l'accident : la luminosité, les conditions météorologiques, le tracé en plan (à savoir si l'accident s'est produit sur une partie rectiligne de l'autoroute ou au niveau d'une courbe), l'état de la surface de la route, la présence ou non d'infrastructures d'aménagements (bretelle, souterrain ou pont), le type d'accident (à savoir s'il s'agit d'un accident isolé ou d'un accident impliquant d'autres véhicules) et le type de véhicule (à savoir si le véhicule est un deux-roues ou non).

Les informations de localisation seront essentielles dans la suite, lorsque nous ferons le lien entre les accidents et les données de trafic. Et, parmi toutes les variables supplémentaires nous renseignant sur le contexte de l'accident, seule la variable « Type d'accident » sera prise en compte dans la suite de l'étude. Les accidents isolés et non-isolés étant de nature très différentes, il sera nécessaire de les traiter de manière séparée.

Comme nous l'avions suggéré lors de la présentation des données d'accidents, il aurait été bon et facile à mettre en œuvre de retirer de l'étude les accidents dont la cause est lié à des phénomènes ne faisant pas intervenir les caractéristiques cinématiques des véhicules : malaise au volant, somnolence, etc. Ces accidents ne devraient pas être pris en compte, afin de ne pas fausser l'étude, puisqu'ils ne sont pas liés à des caractéristiques cinématiques particulières. Mais, étant donné que nous ne considérerons que les données de trafic du mois de juillet 2009, le nombre d'accidents inclus dans l'étude, à savoir 21, est très faible. Nous avons alors préféré n'en supprimer aucun.

Nous aboutissons ainsi à une table SAS, dont nous exposons un extrait au lecteur dans l'annexe C.1. Le code SAS qui réalise cette table de synthèse des accidents est placé dans l'annexe C.2.

3.3. Traitement préparatoire des données de trafic

1ère étape : Tri des données de trafic

Comme nous l'avons présenté lors de l'introduction des données de trafic, les données « Marius » classent les véhicules par temps de passage croissants, en mélangeant les capteurs, les autoroutes, les sens de circulation et les voies de circulation.

La première étape de la transformation des données de trafic consiste donc à trier les données selon le capteur qui a pris la mesure, l'autoroute, la voie de circulation, le sens de circulation, et finalement les temps de passage. Un programme FORTRAN, crée par Maurice Aron, réalise ce tri et aboutit à la création, pour chaque mois, d'environ 14000 fichiers du type : « M1as1000430C20090712.txt ». Le nom du fichier contient les informations de localisation des données de trafic contenues à l'intérieur du fichier :

- « M1a » correspond à l'autoroute A51. Un « m » minuscule signifie que le capteur est situé au niveau d'une bretelle, alors qu'un « M » majuscule signifie qu'il est situé en section courante. Le chiffre 1 correspond à l'autoroute A51, les chiffres 2,3 et 4 à l'autoroute A50, les chiffres 5 et 6 à l'autoroute A55 et les chiffres 7 et 8 à l'autoroute A7.
- « s1 » signifie qu'il s'agit des données de trafic dans le sens de circulation des points routiers décroissants. « s2 » symbolise le sens de circulation des points routiers croissants.
- « 000430 » (0,430 km) est la valeur du point routier correspondant au capteur qui a enregistré les mesures.

- « C » représente la voie centrale, « D » celle de droite et « G » celle de gauche.
- « 20090712 » correspond à la date du 12 juillet 2009.

Ce fichier stocke ainsi les données individuelles relatives à chacun des véhicules détectés le 12 juillet 2009 par le capteur en question, sur la voie de droite de l'autoroute A51, dans le sens des points routiers décroissants. Le lecteur trouvera un extrait du fichier « M1as1000430C20090712.txt » dans l'annexe C.3.

Pour le mois de juillet 2009, le temps de calcul nécessaire à ce tri a été d'environ 60 heures. La taille maximale des fichiers obtenus est de l'ordre de 2 méga-octets.

2nde étape : Création des variables et agrégation par cycles de six minutes

Il s'agit ensuite, dans un premier temps, de créer les variables microscopiques individuelles associées aux véhicules, puis, dans un second temps, d'agréger l'ensemble de ces véhicules par cycles de 6 minutes, et de calculer la valeur de nos variables macroscopiques et microscopiques agrégées pour chacun des cycles. À partir de la liste des 14000 fichiers issus du traitement précédent, nous allons calculer, pour un fichier donné, la valeur de nos variables macroscopiques et microscopiques agrégées sur chacun des 240 cycles de la journée. Ainsi, à partir de chaque fichier du type « M1as1000430C20090712.txt » qui stocke les données individuelles des véhicules, nous créons un fichier du type « M1as1000430C20090712CYCLES.txt ».

En plus des variables agrégées de trafic, nous créons deux variables qualitatives binaires : un indicateur d'accident « IndAcc » et un indicateur du type d'accident « TypeAcc ». Pour l'instant, la variable « IndAcc » est initialisée à la valeur 0, et « TypeAcc » à la valeur « NA ». L'indicateur d'accident devra prendre la valeur 1 lorsque le cycle concerné correspond à une situation d'avant-accident, et 0 sinon. Et, pour un cycle donné, si « IndAcc » vaut 1, la variable « TypeAcc » précisera la nature de l'accident en question : « TypeAcc » prendra la valeur « Seul » s'il s'agit d'un accident isolé et la valeur « PasSeul » sinon. Le traitement informatique présenté dans la section 3.4. réalisera ces assignations de valeurs, en faisant le lien entre données de trafic et données d'accidents.

La taille maximale des fichiers obtenus est de l'ordre de 50 kilo-octets, ce qui est bien inférieur à celle des fichiers source, à savoir 2 méga-octets. Les données obtenues, agrégées en cycles, sont donc bien plus compactes que les données individuelles.

Le lecteur trouvera dans l'annexe C.4. un extrait du fichier « M1as1000430C20090712CYCLES.txt », et dans l'annexe C.5. le code du programme R qui réalise cette tâche ainsi que les détails de la réalisation. Nous avons dû, notamment, filtrer les données aberrantes issues principalement de pannes ou d'erreurs de mesure des capteurs, et les traiter de manière astucieuse comme des données manquantes. Le temps de calcul nécessaire à ce traitement informatique, pour le mois de juillet 2009, a été d'environ 60 heures.

3.4. Traitement du lien entre données d'accidents et données de trafic

Il s'agit à présent de faire le lien entre les données d'accidents et les données de trafic. Nos milliers de fichiers de données de trafic agrégées par cycles, du type « M1as1000430C20090712CYCLES.txt » contiennent, en plus des valeurs des variables macroscopiques et microscopiques correspondantes à chacune des 240 périodes de 6 minutes, une variable indicatrice d'accident « IndAcc » et une variable indicatrice du type d'accident « TypeAcc », initialisées aux valeurs « 0 » et « NA ».

Nous allons donc, à l'aide d'un programme FORTRAN conçu par Maurice Aron et d'un programme R, « insérer » les accidents dans les données de trafic, en repérant, pour chaque accident, le cycle correspondant à la situation d'avant-accident. « IndAcc » prendra donc la valeur 1 lorsque le cycle concerné correspond à une situation d'avant-accident. Et, la variable « TypeAcc » prendra les valeurs « Seul » ou « PasSeul » selon la nature de l'accident. Ce traitement modifie donc les valeurs de ces deux variables dans les 14000 fichiers issus du traitement de la section 3.3. L'annexe C.6. présente un extrait du fichier «M7is1264687C20090706CYCLES.txt » traité par la procédure.

Détaillons les deux grandes étapes du traitement.

1ère étape : Création du fichier « acc_baac_marius.csv »

Il s'agit, en premier lieu, de faire correspondre à chaque accident les fichiers de trafic correspondant au capteur immédiatement en aval du lieu de l'accident, grâce aux informations de localisation des accidents présentes dans la table de synthèse des accidents.

Cependant, un problème apparait : lorsqu'un accident se produit, nous avons toutes les informations de localisation le concernant excepté la voie de circulation sur laquelle il s'est produit. Or, les fichiers de données de trafic sont classés par voie. Comme nous ne pouvons pas déterminer la voie exacte sur laquelle s'est produit l'accident, nous décidons d'associer l'accident aux deux ou trois fichiers de données de trafic correspondant aux deux ou trois voies de circulation. Et, chacun de ces fichiers sera alors modifié pour indiquer la présence de l'accident.

Maurice Aron a réalisé un programme en FORTRAN qui réalise ce traitement, et crée le fichier « acc_baac_marius.csv ». Ce fichier, qui reprend aussi toutes les informations de la table de synthèse des accidents, permet donc d'associer les accidents aux bons fichiers de trafic, qu'il s'agira alors de modifier dans la 2nde étape. Un extrait du fichier « acc_baac_marius.csv » est disponible dans l'annexe C.7.

2nde étape : Insertion des données relatives aux accidents dans les données de trafic

Le fichier « acc_baac_marius.csv » associe chacun des accidents aux fichiers de données de trafic correspondants, qu'il s'agit donc maintenant de modifier. Pour chaque accident, il s'agit donc à présent de repérer dans ces fichiers le cycle correspondant à la situation d'avant-accident : ceci est réalisé à l'aide de l'information de l'heure et de la minute à laquelle l'accident a eu lieu. Pour le moment, dans notre modélisation, le temps de parcours entre le lieu de l'accident et le capteur en aval n'est pas estimé.

Le lecteur trouvera dans l'annexe C.8. le code du programme R qui réalise ce travail ainsi que quelques explications supplémentaires.

3.5. Agrégation finale des fichiers de données

La dernière étape de transformation des données consiste à agréger bout-à-bout tous les fichiers du type « M1as1000430C20090712CYCLES.txt », à présent enrichis des informations relatives aux accidents.

Lors de l'agrégation des fichiers, nous distinguons les trois différentes voies de circulation : la voie de gauche (G), la voie centrale (C) et la voie de droite (D). En effet, les valeurs que prennent les différentes variables sur les trois voies ne sont comparables entre elles ; il ne s'agit donc pas de mélanger tout les fichiers.

Nous réalisons ainsi avec R deux programmes. Le premier programme permet de lister les fichiers de données selon les voies de circulation, et crée ainsi les fichiers « ListeG.txt », « ListeC.txt » et « ListeD.txt ». Le second programme va ensuite agréger bout-à-bout les fichiers de données qui apparaissent dans chacune des listes. On obtient ainsi trois gros fichiers de données : « DonneesCyclesG.txt », « DonneesCyclesC.txt » et « DonneesCyclesD.txt ». C'est le fichier « DonneesCyclesD.txt », dont la taille est d'environ 260 méga-octets, qui servira à l'étude statistique des indicateurs de risque dans le chapitre 5.

Le lecteur trouvera dans l'annexe C.9. le code des programmes réalisant le listing des fichiers et l'agrégation bout-à-bout des fichiers.

4. Les outils de la régression logistique

4.1. Introduction

L'objet de cette partie est de présenter de manière simple les principaux outils statistiques que nous avons à disposition pour expliquer une variable binaire Y à partir d'une autre variable X également binaire. Le but ici n'est pas de présenter la théorie de la régression logistique dans tous ses détails et toutes ses subtilités ; il s'agit de l'introduire de manière générale et de présenter les principaux outils dont nous disposons pour interpréter, en termes de performances du modèle, les tables de contingences obtenues en croisant deux variables binaires.

La variable Y que nous voulons expliquer est la variable indicatrice d'accident « IndAcc », c'est-à-dire celle qui associe au cycle la valeur 1 si le cycle concerné correspond à une situation d'avant-accident, et la valeur 0 sinon. La variable explicative X, candidate au statut d' « indicatrice de risque », correspond à la variable dont nous souhaiterions tester le pouvoir de discriminer les situations d'accident par rapport aux situations de non-accident. La variable X sera alors transformée en une variable binaire $IndX$ à travers le choix d'un seuil de coupure optimal. La valeur $IndX = 1$ correspondra aux valeurs de la variable quantitative X indiquant a priori un risque d'accident. La valeur $IndX = 0$, quant à elle, représentera les valeurs de X non-indicatrice a priori d'un risque d'accident.

Nous présenterons d'abord le principe général de la régression logistique et de la transformation logistique en insistant sur les différences qui apparaissent par rapport au modèle linéaire. Nous verrons ensuite des outils concrets permettant d'interpréter les tables de contingence et de mesurer la performance de la discrimination des valeurs de Y par la variable X. Nous appliquerons ces méthodes dans le chapitre 5., en précisant les notions de seuils de coupure et de binarisation.

Notons bien que nous aurions pu choisir d'étudier la réponse Y en gardant la variable X sous sa forme quantitative. Cependant, la littérature sur le sujet nous apprend que ce type d'étude est alors globalement moins performant. Par ailleurs, les tables de contingence obtenues en croisant deux variables binaires sont d'une approche plus simple, et les résultats sont d'une interprétation plus aisée.

4.2. Principe générale de la régression logistique

4.2.1. Introduction

Pour comprendre les fondements de la régression logistique, plaçons-nous dans un cadre général en considérant la réponse Y à valeurs dans le doublet $\{0,1\}$ que l'on veut expliquer par une variable X de nature quantitative.

Il est impossible de modéliser directement le lien entre la réponse et la variable explicative à l'aide d'une régression linéaire. En effet, une régression linéaire est un processus continu qui renvoie des valeurs dans $\mathbb{R}$ tout entier. Elle semble donc très mal adaptée à la situation. Nous allons donc tenter de modéliser la probabilité que survienne l'évènement $\{Y = 1\}$ en fonction des valeurs de la variable explicative X. C'est l'objet de la régression logistique.

Nous allons donc tenter de modéliser la probabilité $p(x) = \mathbb{P}(Y = 1 \mid X = x)$, dont les valeurs sont comprises entre 0 et 1. Lorsque la variable explicative a une valeur fixée x, on aimerait pouvoir prédire la probabilité qu'a la réponse de prendre l'une ou l'autre de ses modalités. Notons que si nous avons accès à $p(x) = \mathbb{P}(Y = 1 \mid X = x)$, nous pouvons directement en déduire $\mathbb{P}(Y = 0 \mid X = x)$ par la formule : $\mathbb{P}(Y = 0 \mid X = x) = 1 - \mathbb{P}(Y = 1 \mid X = x)$. Pour une valeur x fixée, La variable aléatoire $Y|X = x$ prend deux valeurs ; elle peut donc être modélisée par une loi de Bernoulli :

$$Y|X = x \sim \mathfrak{B}(1, p(x)), \text{ où } p(x) = \mathbb{P}(Y = 1 \mid X = x)$$

D'où, en particulier :

$$\mathbb{E}(Y|X = x) = p(x) \text{ et } VAR(Y|X = x) = p(x) * (1 - p(x))$$

Rappelons que, pour une régression linéaire classique, le modèle s'écrit :

$$Y = X\beta + \varepsilon$$

où ε est le vecteur aléatoire des erreurs, qui vérifie les conditions suivantes :

- Les erreurs sont indépendantes
- Les erreurs sont normales et centrées
- Les erreurs ont toutes la même variance σ^2 (hypothèse d'homoscédasticité)

C'est-à-dire : $Y|X = x \sim \mathcal{N}(x\beta, \sigma^2)$. et donc : $\mathbb{E}(Y|X = x) = x\beta$ et $VAR(Y|X = x) = \sigma^2$.

Ainsi, plusieurs différences par rapport au modèle linéaire apparaissent dans une modélisation logistique :

- La variable $Y|X = x$ ne suit pas une loi normale, mais une loi de Bernoulli
- $VAR(Y|X = x) = p(x) * (1 - p(x))$: la variance n'est pas constante et dépend de x, la condition d'homoscédasticité n'est donc pas validée

Il s'agit donc d'étendre le modèle linéaire classique aux cas où la réponse à modéliser est binaire (ce qui revient à dire que nous nous intéressons à des probabilités). Les résidus, qui ne prennent que deux valeurs, ne peuvent alors plus être modélisés par une loi normale, et l'hypothèse d'homoscédasticité n'est plus vérifiée. La méthode la plus couramment utilisée pour résoudre ce problème est la régression logistique, qui est une méthode statistique d'étude de variables qualitatives qui appartient à la classe du modèle linéaire généralisé (GLM).

4.2.2. La transformation logistique

Comme nous ne pouvons pas modéliser directement la probabilité $\mathbb{E}(Y|X = x) = p(x)$ par la formule linéaire : $p(x) = \beta_0 + \beta_1 * x$, nous allons nous servir d'une « fonction de lien », notée h, pour transformer les $\beta_0 + \beta_1 * x$ en valeurs dans l'intervalle $]0,1[$ (puisque nous voulons modéliser une probabilité). La fonction de lien h sur laquelle est basée la régression logistique est la suivante :

$$h : \mathbb{R} \to]0,1[\text{ définie par :}$$

$$h(x) = \frac{\exp(\beta_0 + \beta_1 * x)}{1 + \exp(\beta_0 + \beta_1 * x)}$$

La fonction h est appelée la « fonction de transformation logistique ».

Le modèle de régression logistique s'écrit alors de la manière suivante :

$$\mathbb{E}(Y|X = x) = p(x) = \frac{\exp(\beta_0 + \beta_1 * x)}{1 + \exp(\beta_0 + \beta_1 * x)}$$

Pour linéariser cette relation, écrivons :

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 * x$$

Si l'on définit la fonction logit, bijective de $]0,1[$ dans $\mathbb{R}$, par la relation :

$$logit\ y = \log\left(\frac{y}{1 - y}\right)$$

On obtient alors :

$$logit\ p(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 * x$$

Le lecteur trouvera dans l'annexe D.1. le graphe de la fonction logit.

En se ramenant à la variable Y, le modèle de régression logistique peut alors s'écrire :

$$Y = p(x) + \varepsilon, \text{ où } p(x) = \frac{\exp(\beta_0 + \beta_1 * x)}{1 + \exp(\beta_0 + \beta_1 * x)}$$

La variable Y peut prendre la valeur 1 ou 0, et donc l'erreur ε peut prendre deux valeurs :

- Si $Y = 1$, $\varepsilon = 1 - p(x)$
- Si $Y = 0$, $\varepsilon = -p(x)$

Ainsi, puisque $p(x) = \mathbb{P}(Y = 1 \mid X = x)$, ε prend la valeur $1 - p(x)$ avec probabilité $p(x)$ et la valeur $-p(x)$ avec probabilité $1 - p(x)$.

Considérons à présent le cas qui nous intéresse, à savoir le cas où X et Y sont deux variables binaires à valeurs dans $\{0,1\}$. Si la modalité $X = 0$ est prise comme modalité de référence, le modèle de régression logistique s'écrit :

$$logit\ p(x) = \beta_0 + \beta_1 * \mathbb{1}_1(x)$$

où l'indicatrice $\mathbb{1}_1$ est définie par :

$$\mathbb{1}_1(x) = \begin{cases} 1\ si\ x = 1 \\ 0\ sinon \end{cases}$$

En se ramenant à la réponse Y, le modèle s'écrit alors :

$$Y = p(x) + \varepsilon, \text{ où } p(x) = \frac{\exp(\beta_0 + \beta_1 * \mathbb{1}_1(x))}{1 + \exp(\beta_0 + \beta_1 * \mathbb{1}_1(x))}$$

où l'erreur ε prend la valeur $1 - p(x)$ avec probabilité $p(x)$ et la valeur $-p(x)$ avec probabilité $1 - p(x)$.

4.2.3. Estimation des paramètres et influence sur la réponse

Le vecteur des paramètres $\beta = (\beta_0, \beta_1)$ est estimé à partir d'un échantillon de n observations indépendantes $(x_1, y_1), \ldots, (x_n, y_n)$ par la méthode du maximum de vraisemblance. Par l'écriture de la vraisemblance et l'annulation des dérivés de la log-vraisemblance par rapport aux deux paramètres, nous obtenons deux équations non-linéaires à résoudre pour obtenir l'estimateur $\hat{\beta}$. Ces équations sont résolues, dans le cas général, à l'aide d'un algorithme itératif appelé IRLS (Iterative Reweighted Least Squares). Nous avons placé dans l'annexe D.2. les développements de ces calculs.

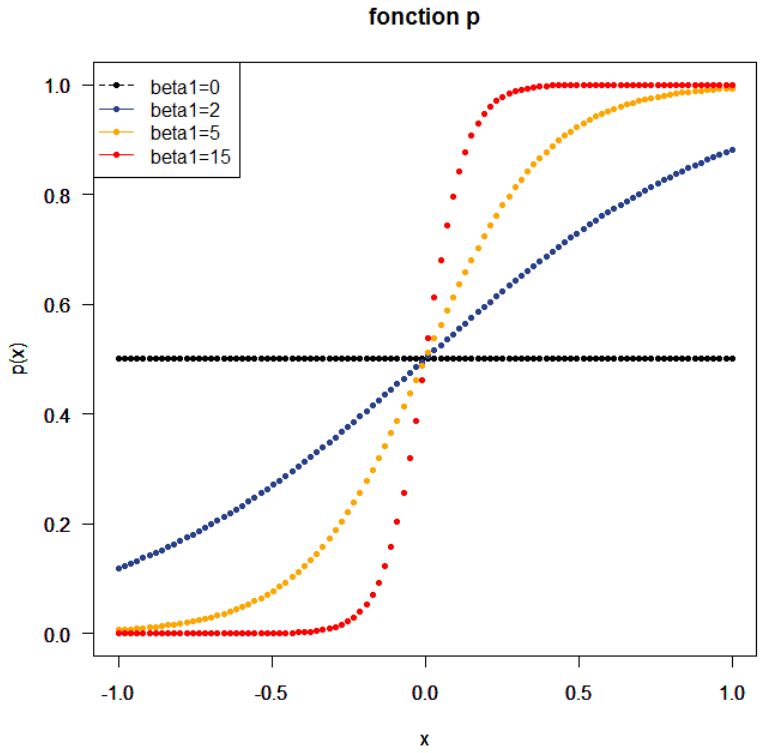
En ce qui concerne l'influence des paramètres sur la réponse $p(x)$, un coefficient positif indique une corrélation positive entre la modalité de la variable explicative et la probabilité $p(x)$, et inversement si le coefficient est négatif. Comme pour le modèle linéaire, l'intensité de la corrélation augmente avec la valeur absolue du coefficient.

Considérons à nouveau l'expression : $p(x) = \frac{\exp(\beta_0+\beta_1*x)}{1+\exp(\beta_0+\beta_1*x)}$

Pour avoir une idée de l'influence des paramètres sur l'aspect de la fonction p, fixons $\beta_0 = 0$ et faisons varier le paramètre β_1. La fonction p devient :

$$p(x) = \frac{\exp(\beta_1 * x)}{1 + \exp(\beta_1 * x)}$$

Voici le tracé de la fonction p pour 4 valeurs de β_1 : 0, 2, 5 et 15.



Nous pouvons faire les remarques suivantes :

- Lorsque β_1 prend de faibles valeurs ($\beta_1 = 2$ par exemple), il y a une large plage de valeurs de x pour lesquelles la probabilité se situe aux alentours de 0.5 (dans le cas extrême $\beta_1 = 0$, la probabilité est constante égale à 0.5). Ainsi, si le coefficient β_1 est faible, le modèle aura du mal à discriminer les 2 modalités de la réponse.
- Au contraire, lorsque β_1 prend des valeurs élevées ($\beta_1 = 15$ par exemple), les valeurs x de la variable explicative vont majoritairement être associées à une probabilité proche de 0 ou de 1, ce qui signifie que le modèle discrimine bien mieux et est donc plus performant.

En tant qu'estimateur obtenu par la méthode du maximum de vraisemblance, $\widehat{\beta_1}$ présente les propriétés de normalité et d'efficacité asymptotiques. $\widehat{\beta_1}$ étant de plus sans biais, si l'on note σ_1

son écart type, supposé connu, nous pouvons construire simplement, à l'aide d'un quantile de la loi normale, un intervalle de confiance asymptotique au seuil $\propto$ pour β_1 :

$$IC_{\propto}(\beta_1) = \left[\widehat{\beta_1} - u_{1-\propto/2} * \sigma_1 \; ; \; \widehat{\beta_1} + u_{1-\propto/2} * \sigma_1\right]$$

où : $u_{1-\propto/2}$ est le quantile au seuil $1 - \propto/2$ de la loi normale centrée réduite.

En pratique, σ_1 est estimé. Cependant, afin de simplifier le problème et d'être en cohérence avec les hypothèses que nous allons faire pour la construction du test de Wald dans la partie 4.3.1., nous supposons σ_1 connu. La théorie de la régression logistique fournit une estimation de ${\sigma_1}^2$: $\frac{1}{Vrai1} + \frac{1}{Faux1} + \frac{1}{Vrai0} + \frac{1}{Faux0}$, où les grandeurs Vrai1, Faux1, Vrai0 et Faux0 sont définis dans la section 4.3. Ils représentent le nombre d'accidents et de non-accidents correspondant à des valeurs « à risque » ou non « à risque » de l'indicateur testé.

4.3. Mesures de performances du modèle

Avant de présenter les principales mesures de performance d'un modèle qui permettent de juger de la qualité de la discrimination, il faut s'assurer que le modèle est valide, c'est-à-dire que sa qualité d'ajustement aux données est satisfaisante. Pour cela, plusieurs tests, basés sur la statistique de la déviance ou sur les résidus de Pearson, peuvent être mis en œuvre. Le test de Hosmer et Lemeshow peut aussi être intéressant. Il serait bon, par ailleurs, de faire une étude détaillée des données, en termes d'effets leviers et de distances de Cook. Cependant, l'étude de la validation d'un modèle logistique n'étant pas le cœur de ce rapport, nous ne la traiterons pas.

La performance du modèle, en termes de discrimination, peut être étudiée soit à l'aide d'un test, le test de Wald, qui juge de la pertinence de la discrimination réalisée par la variable explicative, soit à l'aide de mesures de performances intuitives. La construction du test de Wald et des mesures de performances repose sur l'étude du tableau de contingence obtenu en croisant les deux variables binaires suivantes :

- La variable binaire indicatrice d'accident « IndAcc », notée ici Y, qui associe la valeur 1 aux cycles correspondant à des situations d'avant-accident, et 0 sinon.
- La variable binaire testée $IndX$, qui est la variable binarisée associée à X, obtenue dès lors que l'on choisit un seuil de coupure. X est la variable quantitative dont nous souhaitons tester la capacité de bien discriminer les situations d'accident par rapport aux situations de non-accident.

	$IndX = 1$	$IndX = 0$	
$Y = 1$	*Vrai1*	*Faux1*	*Vrai1+Faux1=p*
$Y = 0$	*Faux0*	*Vrai0*	*Faux0+Vrai0=n-p*
	Vrai1+Faux0 = #{IndX=1}	*Faux1+Vrai0=#{IndX=0}*	*n*

p est le nombre de situations d'accident, n le nombre de cycles étudiés, #{IndX=1} le nombre de situations « à risque » pour la variable X et #{IndX=0}=n-#{IndX=1} le nombre de situations non « à risque ».

Le tableau de contingence est construit à partir des quatre nombres suivants :

- Vrai1 : le nombre d'observations de cycles correspondant à la fois à une situation d'accident $(Y = 1)$ et à une valeur « à risque » $(IndX = 1)$ de la variable testée X. C'est donc le nombre de situations d'accident correctement prédites par $IndX$.
- Vrai0 : le nombre d'observations de cycles correspondant à la fois à une situation de non-accident $(Y = 0)$ et à une valeur non-indicatrice de risque $(IndX = 0)$ de la variable testée X. C'est donc le nombre de situations de non-accident correctement prédites par $IndX$.
- Faux1 : le nombre d'observations de cycles correspondant à la fois à une situation d'accident $(Y = 1)$ et à une valeur non-indicatrice de risque $(IndX = 0)$ de la variable testée X. Il s'agit donc du nombre de « carences d'alarme ».
- Faux0 : le nombre d'observations de cycles correspondant à la fois à une situation de non-accident $(Y = 0)$ et à une valeur « à risque » $(IndX = 1)$ de la variable testée X. Il s'agit donc du nombre de « fausses alarme ».

4.3.1. Test de Wald

Étant donné que, dans la suite, nous n'ajusterons que des modèles de régression logistique univariées, nous ne présenterons ici que le test de Wald univarié, qui permet de juger de la significativité de l'influence d'une seule variable explicative.

Reprenons le modèle de régression suivant :

$$logit\ p(x) = \beta_0 + \beta_1 * \mathbb{1}_1(x)$$

Si nous voulons juger de l'influence de la variable $IndX$, c'est la nullité du coefficient β_1 qu'il faut tester. Nous voulons donc tester :

$$H_0 : \beta_1 = 0 \text{ contre } H_1 : \beta_1 \neq 0$$

La statistique à utiliser est : $T = {(\widehat{\beta_1} - \beta_1)}/{\sigma_1}$, où l'écart-type σ_1 de $\widehat{\beta_1}$ est supposé connu.

Sous l'hypothèse nulle : $T = {\widehat{\beta_1}}/{\sigma_1}$.

Dans le logiciel R, le test de Wald réalisé suppose que la statistique T suive sous H_0 une loi $\mathcal{N}(0,1)$. Pour cela, il faut supposer que $\widehat{\beta_1}$ suive une loi normale, ce qui est asymptotiquement vrai, et que l'écart-type σ_1 est connu. Or, en réalité, ${\sigma_1}^2$ est estimé par la grandeur $\frac{1}{Vrai1} + \frac{1}{Faux1} + \frac{1}{Vrai0} + \frac{1}{Faux0}$, que l'on peut approximer par $\frac{1}{Vrai1} + \frac{1}{Faux1}$. En effet, les nombres Vrai0 et Faux0, correspondant aux situations de non-accidents, sont beaucoup plus élevés que Vrai1 et Faux1 ; les termes $\frac{1}{Vrai0}$ et $\frac{1}{Faux0}$ peuvent donc être négligés.

Ainsi, afin de simplifier les calculs et d'être en cohérence avec R, nous supposerons que T suit sous H_0 une loi normale centrée réduite.

Le test sera réalisé avec une erreur de première espèce fixée à $\propto = 5\%$. Nous serons aussi amenés à nous intéresser à l'erreur de deuxième espèce β, étant donné que certaines applications de ce test vont décider de conserver l'hypothèse nulle.

Une étude de la puissance, réalisée dans l'annexe D.3., donne les résultats suivants :

- La puissance réelle du test est :

$$1 - \beta = \phi\left(\frac{\beta_{1,obs}}{\sqrt{\frac{1}{Vrai1} + \frac{1}{Faux1}}} - u_{1-{\alpha}/{2}}\right)$$

  où $\beta_{1,obs}$ la valeur estimée de β_1 (correspondant à l'hypothèse alternative) et $u_{1-{\alpha}/{2}}$ le quantile d'ordre $1 - {\alpha}/{2}$ de la loi normale centrée réduite.

- Le nombre d'accidents p à inclure dans l'étude pour obtenir une puissance $1 - \beta$ fixée (par exemple 80%) est :

$$p = k * (Vrai1 + Faux1)$$

$$o\grave{u}\ k = \left(u_{1-{\alpha}/{2}} + u_{1-\beta}\right)^2 * \left(\frac{1}{Vrai1} + \frac{1}{Faux1}\right) * \frac{1}{{\beta_1}^2}$$

  β_1 correspond, selon ce que l'on souhaite étudier, soit à la valeur estimée à partir des données $\beta_{1,obs}$, soit à une valeur de référence que l'on fixe de manière arbitraire et que l'on aimerait pouvoir détecter à la puissance $1 - \beta$. $u_{1-{\alpha}/{2}}$ et $u_{1-\beta}$ sont les quantiles d'ordre $1 - {\alpha}/{2}$ et$1 - \beta$ de la loi normale centrée réduite.

4.3.2. Mesures de performances intuitives

Par ailleurs, on peut s'intéresser à des mesures intuitives sur le pouvoir discriminant du modèle. Voici donc les différents outils dont nous allons nous servir, dans le chapitre 5., pour élaborer des méthodes permettant de juger du pouvoir discriminant de nos potentiels futurs indicateurs de risque :

- L'Odd Ratio (OR) : il s'agit d'un « rapport de côtes » ou « rapport de chances ». En effet, la chance relative des observations présentant le caractère $IndX = 1$ d'avoir une réponse $Y = 1$ plutôt que $Y = 0$ est OR fois la chance relative des observations présentant le caractère $IndX = 0$. OR est estimé par la grandeur suivante :

$$\widehat{OR} = \frac{Vrai1/_{Faux0}}{Faux1/_{Vrai0}}$$

Dans le contexte de notre étude, étant donné la très faible probabilité que survienne un accident, l'Odd Ratio peut être plus simplement interprété en termes de risques relatifs : il y a OR fois plus de chances de présenter la modalité $Y = 1$ lorsque $IndX = 1$ que lorsque $IndX = 0$. En d'autres termes, il y a OR fois plus de chances que la variable testée indique une situation d'accident lorsque les valeurs de la variable sont « à risque » que lorsqu'elles ne sont pas « à risque ». Ainsi, l'Odd Ratio est un bon indicateur de la qualité de la discrimination des situations d'accident et des situations de non-accident.

Par ailleurs, il existe une relation très simple qui fait le lien entre l'OR et le coefficient β_1 de la régression logistique :

$$OR = e^{\beta_1} \qquad (1)$$

Cette relation nous permet de construire un intervalle de confiance asymptotique au seuil $\propto$ pour OR, à partir de l'intervalle de confiance asymptotique au seuil $\propto$ de β_1. Cet intervalle de confiance s'écrit :

$$IC_{\propto}(OR) = \left[e^{\widehat{\beta_1} - u_{1-\propto/2} * \sigma_1} ; e^{\widehat{\beta_1} + u_{1-\propto/2} * \sigma_1}\right] = \left[\widehat{OR} * e^{-u_{1-\propto/2} * \sigma_1} ; \widehat{OR} * e^{u_{1-\propto/2} * \sigma_1}\right]$$

où $u_{1-\propto/2}$ est la quantile au seuil $1 - \propto/2$ de la loi normale centrée réduite et σ_1 la valeur supposée connue de l'écart-type de $\widehat{\beta_1}$.

En pratique, comme nous l'avons déjà remarqué, la variance ${\sigma_1}^2$ sera estimée par la grandeur : $\frac{1}{Vrai1} + \frac{1}{Faux1}$.

Le lecteur trouvera dans l'annexe D.4. des précisions sur la construction de l'Odd Ratio et des risques relatifs ainsi qu'une démonstration de la relation (1).

- La sensibilité : c'est la proportion de situations d'accident correctement prédites parmi les cas où l'indicateur est à risque.

$$Sensibilité = \frac{Vrai1}{Vrai1 + Faux0}$$

- La spécificité : c'est la proportion de situations de non-accident correctement prédites parmi les cas où l'indicateur n'est pas à risque.

$$Spécificité = \frac{Vrai0}{Vrai0 + Faux1}$$

- Taux de Faux1 : c'est la proportion de situations d'accident prédits à tort.

$$TauxFaux1 = \frac{Faux1}{Vrai1 + Faux1}$$

- Taux de Faux0 : c'est la proportion de situations de non-accident prédits à tort.

$$TauxFaux0 = \frac{Faux0}{Vrai0 + Faux0}$$

- Coefficient « c » : c'est un coefficient théorique compris entre 0 et 1, dont la construction, assez complexe, est détaillée dans l'annexe D.5. En voici une estimation :

$$c = \frac{Vrai1 * Vrai0 + 0.5 * Vrai1 * Faux1 + 0.5 * Vrai0 * Faux0}{(Vrai1 + Faux0) * (Faux1 + Vrai0)}$$

  Les règles d'interprétation de ce coefficient sont les suivantes :
  - Si $c = 0.5$, on considère qu'il n'y a pas de discrimination
  - Si $0.7 \leq c < 0.8$, la discrimination est jugée satisfaisante
  - Si $0.8 \leq c < 0.9$, la discrimination est jugée très bonne
  - Si $c \geq 0.9$, la discrimination est jugée excellente

Ainsi, en résumé, la performance de la discrimination sera d'autant meilleure que l'Odd Ratio (ou son inverse si celui-ci est inférieur à 1), la sensibilité, la spécificité et le coefficient c seront élevés, et TauxFaux1 et TauxFaux0 faibles.

L'idéal serait alors d'obtenir, à travers le choix d'un seuil de coupure, un tableau de contingence qui permet d'optimiser l'ensemble de ces mesures de performances. Cependant, nous verrons dans le chapitre 5. que cela est impossible, et qu'il nous faudra réaliser des compromis entre ces différentes mesures de performances pour réaliser le choix du seuil de coupure optimal.

5. Évaluation de la pertinence des indicateurs de risque

5.1. Introduction

Le cœur de l'étude consiste à juger de la qualité de nos potentiels indicateurs de risque à discriminer les situations d'accident et les situations de non-accident. Pour cela, nous mettrons au point deux méthodes différentes d'évaluation.

La première méthode consiste à comparer la valeur moyenne μ_{acc} que présente l'indicateur en cas d'accidents à la valeur moyenne $\mu_{non-acc}$ des situations de non-accident. Il s'agit alors de voir, à l'aide d'un test approprié, si les valeurs μ_{acc} et $\mu_{non-acc}$ sont significativement différentes, auquel cas le pouvoir de discrimination de l'indicateur sera considéré comme satisfaisant.

La deuxième méthode repose sur l'utilisation des mesures de performances, introduites dans le chapitre précédant ; il s'agit de choisir le seuil de coupure optimal permettant de binariser la variable X que nous voulons tester, de telle sorte que le tableau de contingence obtenu soit optimisé par rapport aux mesures de performances.

L'étude sera réalisée à partir des données agrégées, de type microscopique, de la voie de droite du mois de juillet 2009, c'est-à-dire à partir du fichier « DonneesCyclesD.txt ». Comme nous l'avons déjà mentionné, nous sommes contraints de nous limiter à l'étude d'un seul mois de données, étant donné le temps de calcul trop important nécessaire au traitement des données de toute l'année. Et, par ailleurs, les valeurs des variables n'étant pas comparables entre les différentes voies de circulation, l'étude se limite aux données de la voie de droite. Le nombre d'accident inclus dans l'étude, à savoir 16, est donc malheureusement très réduit. Étant donné que l'étude devra être réalisée en traitant de manière séparée les accidents isolés et les accidents non-isolés, le nombre d'accidents traités simultanément va encore diminuer. Ainsi, lors de l'application de nos méthodes statistiques et la réalisation de nos graphes de performances, nous serons très vite confrontés à des problèmes de significativité statistique. Nous ne pourrons donc pas mener à bout de manière satisfaisante l'étude de l'évaluation de la qualité des indicateurs. Nous ferons cependant une étude de dimensionnement, basée sur une analyse de la puissance des tests mis en œuvre.

Pour illustrer de manière satisfaisante nos méthodes d'optimisation, nous pouvons considérer l'idée suivante : au lieu de ne considérer que les accidents, en nombre trop réduit, nous pouvons considérer les situations « à risque d'accidents » pour un indicateur donné, dont la pertinence est supposée validée. Cela reviendra, comme nous le verrons dans la section 5.3.3., à comparer deux indicateurs de risque entre eux. Le nombre de « situations d'accident » sera alors largement suffisant pour illustrer correctement nos méthodes.

5.2. Méthodes d'évaluation

5.2.1. Comparaison de deux moyennes

La première méthode à laquelle nous pensons pour évaluer la qualité des indicateurs de risque consiste à comparer la valeur moyenne μ_{acc} que présente l'indicateur en cas d'accidents à la valeur moyenne $\mu_{non-acc}$ lorsqu'il n'y a pas d'accident. Il s'agit alors de voir si les valeurs μ_{acc} et $\mu_{non-acc}$ sont significativement différentes, auquel cas le pouvoir de discrimination de l'indicateur de risque considéré sera validé.

Pour comparer deux moyennes, le plus simple serait de mettre en œuvre un test de Student. Cependant, les hypothèses d'application de ce test ne sont pas satisfaites dans le cadre de notre étude : nous avons un nombre trop faible d'accidents (5 de type « Seul » et 11 de type « PasSeul ») et les données ne suivent pas une loi normale. Nous nous tournerons donc vers le test non-paramétrique de Mann-Whitney, qui ne nécessite que très peu de données pour fonctionner (au minimum de 4 à 8), et qui se passe de l'hypothèse de normalité.

Nous voulons donc tester :

$$H_0 : \mu_{acc} = \mu_{non-acc} \; contre \; H_1 : \mu_{acc} \neq \mu_{non-acc}$$

La construction du test de Mann-Whitney est expliquée de manière générale dans l'annexe E.1.

L'analyse de la puissance du test, réalisée dans l'annexe E.2., fournit les résultats suivants :

- La puissance réelle du test est :

$$1 - \beta = \phi\left(\frac{\delta}{\sigma\sqrt{\frac{1}{n1}}} - u_{1-\alpha/2}\right)$$

  où n1 est le nombre d'accident inclus dans l'étude, $\delta = \mu_{acc} - \mu_{non-acc}$ la différence des moyennes que nous voulons pouvoir détecter à la puissance $1 - \beta$, σ l'écart-type des observations, estimé en pratique à l'aide des données correspondant aux situations de non-accident, ϕ la fonction de répartition de la loi normale centrée réduite et $u_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

- Le nombre d'accidents $n1$ à inclure dans l'étude pour obtenir une puissance $1 - \beta$ fixé (par exemple 80%) est :

$$n1 = \left(u_{1-\alpha/2} + u_{1-\beta}\right)^2 * \frac{\sigma^2}{\delta^2}$$

  où σ est l'écart-type des observations, estimé en pratique à l'aide des données correspondant aux situations de non-accident, et $u_{1-\alpha/2}$ et $u_{1-\beta}$ les quantiles d'ordre respectifs $1 - \alpha/2$ et $1 - \beta$ de la loi normale centrée réduite. $\delta = \mu_{acc} - \mu_{non-acc}$ est la différence des moyennes que nous voulons pouvoir détecter à la puissance $1 - \beta$. δ correspond donc soit à une valeur estimée sur les données, soit à une valeur de référence fixée de manière non-statistique.

5.2.2. Optimisation du choix du seuil de coupure

Tout d'abord, fixons les notations suivantes : X est la variable quantitative dont nous voulons tester la capacité de bien discriminer les situations d'accident par rapport aux situations de non-accident, et $IndX$ est la variable binarisée associée à X, obtenue dès lors que l'on choisit un seuil de coupure. Nous allons tenter de choisir le seuil de coupure, de telle sorte qu'il optimise, en un certain sens, les mesures de performances, présentées dans la section 4.3.2. Reprenons donc le tableau de contingence de la section 4.3.

	$IndX = 1$	$IndX = 0$	
$IndAcc = 1$	*Vrai1*	*Faux1*	*Vrai1+Faux1=p*
$IndAcc = 0$	*Faux0*	*Vrai0*	*Faux0+Vrai0=n-p*
	Vrai1+Faux0 = #{IndX=1}	*Faux1+Vrai0=#{IndX=0}*	*n*

où p représente le nombre de situations d'accident, n le nombre de cycles étudiés, #{IndX=1} le nombre de situations « à risque » pour la variable X et #{IndX=0} le nombre de situations non « à risque ».

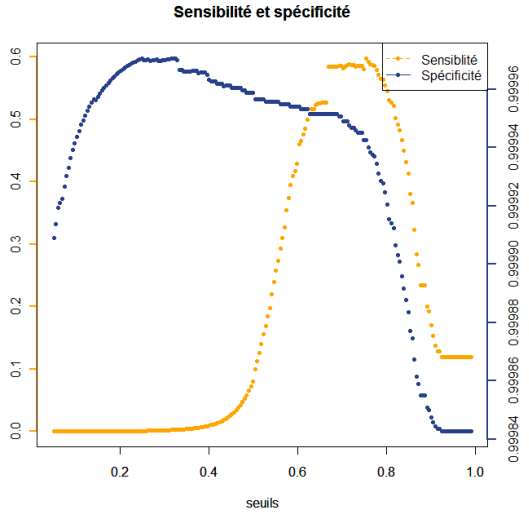
Comme nous l'avons déjà mentionné, l'idée serait de choisir un seuil de coupure qui maximise à la fois la sensibilité, la spécificité, l'Odd Ratio et le coefficient c, et qui minimise TauxFaux0 et TauxFaux1. En pratique, cela est impossible. En effet, sensibilité et spécificité évoluent globalement de manière opposée en fonction des seuils de coupure : lorsque le seuil est faible, la sensibilité est faible et la spécificité est forte, et inversement lorsque le seuil de coupure est élevé. De manière équivalente, TauxFaux1 et TauxFaux0 évoluent aussi en sens contraires. Ainsi, il est impossible de trouver un seuil qui maximise soit à la fois la sensibilité et la spécificité, soit à la fois TauxFaux1 et TauxFaux0, et encore moins ces quatre mesures de performances à la fois.

Ainsi, il s'agit donc, dans le choix du seuil de coupure optimal, de réaliser des compromis entre ces différentes variables. Nous développerons en ce sens plusieurs méthodes d'optimisation. L'application de chacune de ces méthodes fournit alors un seuil de coupure optimal, en un certain sens défini. L'idéal serait alors que toutes les méthodes d'optimisation convergent approximativement vers le choix d'un même seuil de coupure. Mais, en pratique, cela arrive rarement ; il n'y a alors pas de règles générales fixes pour choisir le seuil de binarisation de manière optimale.

Pour illustrer de manière satisfaisante les méthodes d'optimisation, on ne peut pas se servir des graphes de résultats obtenus pour l'étude des accidents ; comme nous le verrons dans les parties 5.3.1. et 5.3.2., du fait du nombre trop faible d'accidents inclus dans l'étude, ces graphes sont légèrement dégénérés. Ainsi, les graphes présentés sont ceux obtenus lors de l'étude croisée des variables $pPICUD0_j$ et $pTIV2_j$, dont l'objet est présentée dans la section 5.3.3. Ce qui importe dans cette partie est d'avoir un aperçu de l'allure des courbes considérées. Les résultats de seuils, ici, ne sont donnés qu'à titre d'exemple.

Méthode M.1. :

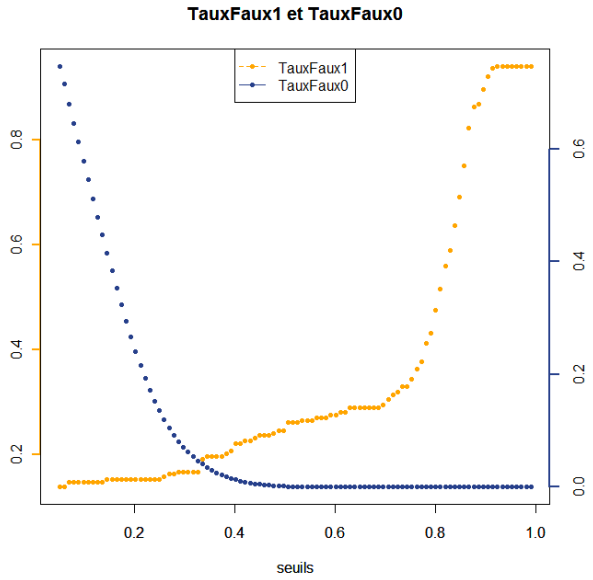
La première méthode d'optimisation du seuil de coupure consiste à choisir le seuil pour lequel les courbes associées respectivement à la sensibilité et à la spécificité s'intersectent. Si l'on applique cette méthode aux données de sensibilité et de spécificité représentées ci-dessous, le seuil choisi est 0.63.



Les méthodes M.2, M.3 et M.4 sont basées sur l'optimisation de TauxFaux1 et TauxFaux0, c'est-à-dire des fausses alarmes et des carences d'alarmes, qui évoluent en sens opposé en fonction des seuils de coupure.

Méthode M.2. :

La seconde méthode, analogue à la méthode précédente, consiste à choisir le seuil de coupure pour lequel les courbes associées à TauxFaux1 et TauxFaux0 s'intersectent. Pour le graphe ci-dessous, le seuil obtenu est alors 0.33.



Méthodes M.3. et M.4. :

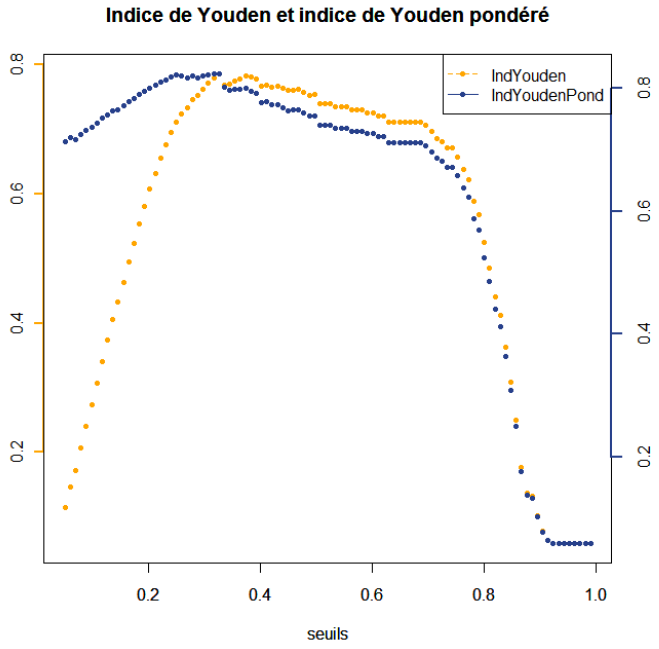
Dans la troisième méthode, au lieu de considérer le point d'intersection relatif à TauxFaux1 et à TauxFaux0, nous allons nous servir de l' « indice de Youden », qui est un indice construit à partir de TauxFaux0 et TauxFaux1. Afin de minimiser de manière optimale ces mesures de performances, à l'aide de cet indice, il s'agit de choisir le seuil qui maximise l'indice suivant :

$$IndYouden = 1 - (TauxFaux0 + TauxFaux1)$$

Cependant, l'intérêt de cette méthode est assez limité, dans la mesure où on accorde le même poids au taux de fausses alarmes (TauxFaux1) et au taux de carences d'alarme (TauxFaux0). En effet, le nombre de situations d'accident incluses dans l'étude étant très faible, nous accordons plus d'importance à la minimisation du taux de carences d'alarme qu'à celle du taux de fausses alarmes. La méthode M.4 prend en compte cela en introduisant un coefficient de pondération dans la formule de Youden :

$$IndYoudenPond = 1 - (TauxFaux0 + k * TauxFaux1)$$

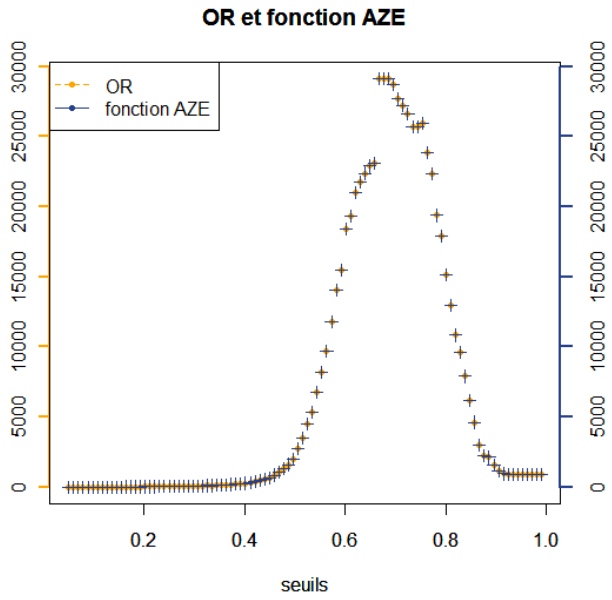
où k=0.2. Cela signifie que le préjudice causé par une carence d'alarme est 5 fois plus élevé que celui causé par une fausse alarme. Une étude spécifique pourrait être faite pour choisir de manière optimale la valeur de coefficient k.



Sur le graphe ci-dessus, la maximisation de l'indice de Youden et de l'indice de Youden pondéré conduit au choix d'un seuil de coupure identique, valant 0.33.

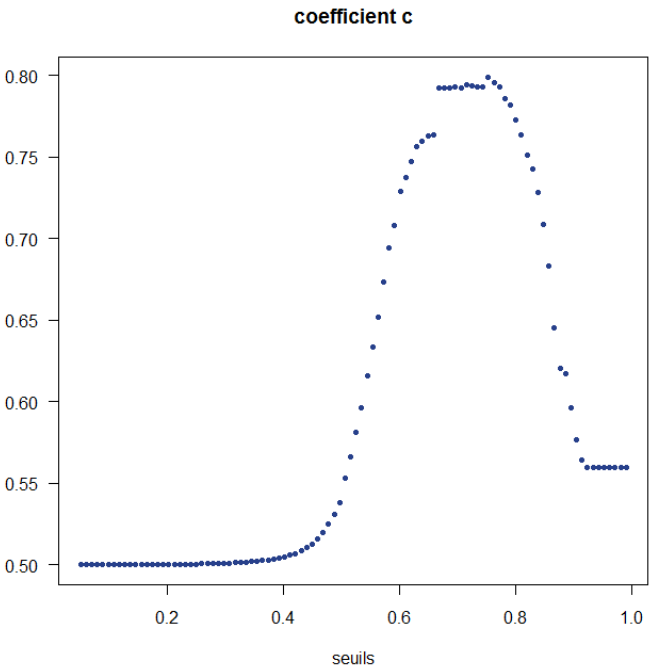
Méthode M.5. :

La cinquième méthode consiste simplement à choisir le seuil de coupure qui maximise l'Odd Ratio, l'une des principales mesures de performances. L'application de la méthode M.5 aux données du graphe ci-dessous conduit à choisir un seuil de coupure valant 0.67.



Méthode M.6. :

La sixième et dernière méthode d'optimisation consiste simplement à choisir le seuil qui maximise la valeur du coefficient c, qui est un très bon indicateur de la qualité de la discrimination. L'application de cette méthode aux données du graphe ci-dessous conduit à choisir un seuil de coupure égal à 0.67.



Au terme de l'application de ces six méthodes, nous disposons donc de plusieurs seuils de coupure différents, chaque seuil étant optimisé en un certain sens. Il s'agira alors, en pratique, de déterminer la valeur du seuil que l'on conserve de manière définitive pour binariser la variable X. Comme il n'y a pas de règles fixes pour réaliser ce choix final, le seuil définitif sera choisi au cas par cas, au vu des résultats obtenus.

5.3. Résultats

Nous avons, en tout, réalisé trois études distinctes :

- L'étude de la pertinence de la variable « vitesse moyenne » dans la discrimination des accidents isolés : « vitesse moyenne » est la seule variable a priori pertinente pour étudier ce type d'accident.
- L'étude de la pertinence de la variable « pPICUD0 » dans la discrimination des accidents non-isolés. Comme nous l'avons vu, les variables microscopiques sont corrélées entre elles ; leur comportement est similaire. Nous nous contenterons donc de l'étude de la variable « pPICUD0 ».
- L'étude croisée des variables « pTIV2 » et « pPICUD0 », dont nous expliquerons l'objet dans la section 5.3.3.

Pour chacune de ces études, nous mettrons en œuvre les différentes méthodes d'évaluation développées précédemment, à savoir :

- Le test de Mann-Whitney
- L'optimisation du choix du seuil de coupure, à travers l'application des six méthodes d'optimisation
- La mise en œuvre informatique de la régression logistique sous-jacente et du test de Wald associé, une fois le seuil de coupure définitif choisi

Une étude de la puissance des différents tests sera réalisée, dans le but d'avoir un ordre de grandeur du nombre d'accidents à inclure dans l'étude pour obtenir une puissance satisfaisante $1 - \beta$ de 80%. Les résultats numériques présentés ont été obtenus en appliquant les formules de puissance et de nombre d'accidents que nous avons démontrées précédemment.

Le lecteur trouvera dans l'annexe E.3. le code R qui a permis de réaliser l'étude de l'optimisation des seuils de coupure (création des tableaux de seuils et des graphes associées aux méthodes d'optimisation).

5.3.1. Étude des accidents isolés

Dans cette étude, seule la variable « vitesse moyenne » sera étudiée. Voici un tableau qui résume les valeurs des variables pour les cinq accidents de ce type dont nous disposons :

```
> donnees[donnees$IndAcc==1 & donnees$TypeAcc=="Seul",]
        Numcycles IndAcc TypeAcc debit     TauxOcc vitessemoy    TIVmoy
191570         50      1    Seul   250 0.017934629   95.04000 14.022000
466372         52      1    Seul   160 0.006992885   98.93750 23.046875
756787         67      1    Seul   890 0.054684711   84.03371  4.076180
1209119       239      1    Seul   450 0.023538864   90.33333  7.954667
1286327       167      1    Seul    NA          NA         NA        NA
           VRmoy    PICUDmoy PICUDBISmoy pTIV0.5      pTIV1      pTIV2
191570  2.177778 0.303111110           0       0 0.04000000 0.12000000
466372  2.586806 0.000000000           0       0 0.00000000 0.06250000
756787  1.282771 0.659765571           0       0 0.07865169 0.38202247
1209119 1.555556 0.008717421           0       0 0.00000000 0.06666667
1286327       NA          NA          NA      NA         NA         NA
            pPICUD0   pPICUD10 pPICUD20 pPICUDBIS0
191570  0.04000000 0.00000000        0          0
466372  0.00000000 0.00000000        0          0
756787  0.11235955 0.02247191        0          0
1209119 0.02222222 0.00000000        0          0
1286327         NA         NA       NA         NA
```

On constate qu'un accident parmi les cinq n'est pas exploitable : seuls quatre accidents seront donc inclus dans l'étude, ce qui est très peu.

Test de Mann-Whitney

Les deux moyennes à comparer sont : $\mu_{acc} = 92.09$ et $\mu_{non-acc} = 86.41$

L'écart-type des observations, estimé sur les données correspondant aux situations de non-accident, est : $\sigma = 14.11$

Le résultat du test, disponible dans l'annexe E.4., est le suivant : la p-valeur étant égal à 0.37, le test conserve l'hypothèse nulle au seuil $\alpha = 5\%$, et décide donc que $\mu_{acc} = \mu_{non-acc}$. L'erreur associée à cette décision est une erreur de seconde espèce, que l'on peut évaluer à l'aide d'une étude de la puissance $1 - \beta$.

Le calcul de la puissance du test fournit : $1 - \beta = 12.4\%$. L'erreur de seconde espèce (87.6%) est donc bien trop élevée pour que la décision prise par le test ait une quelconque valeur. On ne peut donc malheureusement conclure ni dans un sens, ni dans l'autre.

Cependant, nous pouvons calculer le nombre n1 d'accidents à inclure afin que la puissance obtenue soit satisfaisante, égale à 80%. Ce calcul conduit à : $n1 = 49$.

Et, si l'on avait voulu détecter de manière significative, à la puissance 80%, une différence de vitesses égale 3 km/h, il aurait fallu inclure : n1=174 accidents.

Optimisation du seuil de coupure

Les tableaux des seuils et les graphes associés aux méthodes d'optimisation sont disponibles dans les annexe E.5. et E.6. On constate rapidement que, à cause du trop faible nombre d'accidents inclus dans l'étude, les valeurs de certaines variables et l'allure de certains graphes obtenus ne sont pas satisfaisants. En effet, la sensibilité prend des valeurs extrêmement faible, la spécificité vaut quasiment toujours 1, et le coefficient c vaut environ 0.5 quelque soit la valeur du seuil. Malgré cela, nous allons tenter d'appliquer les six méthodes d'optimisation développées.

Les méthodes d'optimisation du choix du seuil de coupure fournissent les valeurs de seuils suivantes :

- Méthode M.1. : elle est inapplicable dans ce contexte.
- Méthode M.2. : 90.83371
- Méthodes M.3. et M.4. : 90.23371
- Méthode M.5. : 90.23371
- Méthode M.6. : 90.23371

On obtient donc deux seuils optimaux différents mais très proches l'un de l'autre. Le seuil de coupure gardé en définitive, qui permet d'optimiser de manière satisfaisante les mesures de performances, est le seuil 90.23371. Les mesures de performances associées sont les suivantes :

- Sensibilité : $5.75*10^{-6}$
- Spécificité : 0.9999986
- TauxFaux1 : 25%
- TauxFaux0 : 42.7%
- IndYouden et IndYoudenPond : 0.322662717 et 0.6645325
- OR : 4.020216
- c : 0.5000022

Le coefficient c, très proche de 0.5, semble indiquer une discrimination très mauvaise, voire inexistante, des situations d'accident et de non-accident. Cependant, la qualité des autres mesures de performances, excepté la sensibilité, semble satisfaisante.

A présent, nous pouvons donc, à l'aide d'un programme R très simple, binariser la variable étudiée à l'aide du seuil choisi et mettre en œuvre avec R le modèle de régression logistique sous-jacent, ainsi que le test de Wald associé. Celui-ci indiquera si la valeur de l'Odd Ratio obtenue, à savoir environ 4.02, est statistiquement significative, c'est-à-dire statistiquement différente de 1. Etant donné que la valeur du coefficient c indique l'inexistence d'une discrimination entre les deux situations, il est très probable que ce test ne soit pas significatif.

Régression logistique et test de Wald

Les sorties R associées sont disponibles dans l'annexe E.4.

Le calibrage du modèle de régression logistique, expliquant la variable binaire indicatrice d'accident « IndAcc » à l'aide de la variable binarisée associée à « vitesse moyenne », fournit l'estimation suivante du coefficient β_1 : 1.391. Nous vérifions aisément qu'en prenant l'exponentielle de cette valeur, nous retrouvons bien l'estimation de l'Odd Ratio obtenue : $e^{1.391} = 4.020216$.

L'intervalle de confiance asymptotique au seuil $\alpha = 5\%$ pour β_1 est alors :

$$IC_\alpha(\beta_1) = [\,-0.87\,;3.65]$$

En passant à l'exponentiel, on obtient donc un intervalle de confiance asymptotique au seuil $\alpha = 5\%$ pour l'Odd Ratio :

$$IC_\alpha(OR) = [\,0.42\,;38.7]$$

La valeur 1 appartient à $IC_\alpha(OR)$, ce qui signifie que l'Odd Ratio n'est pas significativement différent de 1. Cela est naturellement confirmé par l'application du test de Wald. Le résultat du test est le suivant : la p-valeur étant égal à 0.228, le test conserve l'hypothèse nulle au seuil $\alpha = 5\%$, et décide donc que $\beta_1 = 0$, ce qui correspond en effet à $OR = 1$. L'erreur associée à cette décision est une erreur de seconde espèce, que l'on peut évaluer à l'aide d'une étude de la puissance $1 - \beta$.

Le calcul de la puissance du test fournit : $1 - \beta = 22.5\%$, qui est bien trop faible pour que la décision prise par le test ait une quelconque valeur. On ne peut donc pas conclure, avec le nombre présent d'accident inclus dans l'étude, que la variable n'est en effet pas pertinente.

Cependant, nous pouvons calculer le nombre p d'accidents à inclure afin que la puissance obtenue soit satisfaisante, par exemple $1 - \beta = 80\%$. Ce calcul conduit à : $p = 22$.

Si on avait voulu détecter de manière significative un Odd Ratio de 1.2, qui correspond donc à un coefficient $\beta_1 = \ln(1.2) = 0.182$, le nombre p d'accidents à inclure afin que la puissance du test soit égale à 80% aurait été : $p = 1260$.

5.3.2. Étude des accidents non-isolés

Nous étudions à présent la pertinence de la variable « pPICUD0 » dans la discrimination des accidents non-isolés. Voici un tableau qui résume les valeurs des variables pour les accidents de type non-isolé inclus dans l'étude :

```
> donnees[donnees$IndAcc==1 & donnees$TypeAcc=="PasSeul",]
        Numcycles IndAcc TypeAcc debit    TauxOcc vitessemoy   TIVmoy     VRmoy  PICUDmoy
692105        185      1 PasSeul   770 0.04191925  95.532470 4.653896 2.0634920 1.8902116
694947        147      1 PasSeul   420 0.02094815 103.642860 8.534286 2.4669310 1.0577161
710556        156      1 PasSeul    NA         NA         NA       NA        NA        NA
762421        181      1 PasSeul   440 0.37646573   9.395349 7.921818 0.4298942 0.4362654
934260        180      1 PasSeul  1530 0.11341960  74.607840 2.369869 0.7389252 2.1849875
993425         65      1 PasSeul   780 0.06482996  63.858970 4.625641 1.1716524 0.9343938
1039070       110      1 PasSeul   960 0.08186021  84.778950 3.759896 1.4154846 1.8445823
1226373       213      1 PasSeul   740 0.04391862  86.380280 4.819324 1.3153595 1.2119599
1229059        19      1 PasSeul   590 0.03302628  83.135590 6.016271 1.3465160 0.2008213
1236187       187      1 PasSeul   960 0.18022931  43.236560 3.660729 1.0185185 0.9646468
1256088       168      1 PasSeul  1130 0.07777871  85.274340 3.185398 1.0988201 1.9241669
        PICUDBISmoy    pTIV0.5      pTIV1      pTIV2    pPICUD0   pPICUD10    pPICUD20  pPICUDBIS0
692105   0.00000000 0.00000000 0.07792208 0.24675325 0.16883117 0.07792208 0.012987013 0.000000000
694947   0.00000000 0.00000000 0.04761905 0.11904762 0.04761905 0.04761905 0.023809520 0.000000000
710556           NA         NA         NA         NA         NA         NA          NA          NA
762421   0.23487654 0.00000000 0.00000000 0.04545455 0.09523810 0.00000000 0.000000000 0.047619048
934260   0.00000000 0.01960784 0.17647059 0.50980392 0.28104575 0.08496732 0.006535948 0.000000000
993425   0.02718819 0.01282051 0.03846154 0.26923077 0.11538462 0.05128205 0.000000000 0.012820513
1039070  0.14479905 0.01041667 0.04166667 0.26041667 0.13829787 0.08510638 0.021276596 0.021276596
1226373  0.17549020 0.00000000 0.04054054 0.18918919 0.08823529 0.04411765 0.014705882 0.014705882
1229059  0.00000000 0.00000000 0.03389831 0.15254237 0.05084746 0.00000000 0.000000000 0.000000000
1236187  0.18710562 0.00000000 0.04166667 0.29166667 0.15555556 0.04444444 0.000000000 0.055555556
1256088  0.02910248 0.00000000 0.11504425 0.46902655 0.22123894 0.10619469 0.008849558 0.008849558
```

Ainsi, on constate que 10 accidents sont exploitables, et peuvent être inclus dans l'étude.

Test de Mann-Whitney

Les deux moyennes à comparer sont : $\mu_{acc} = 0.136$ et $\mu_{non-acc} = 0.132$.

L'écart-type des observations, estimé sur les données correspondant aux situations de non-accident, est : $\sigma = 0.103$

Le résultat du test, disponible dans l'annexe E.7., est le suivant : la p-valeur étant égal à 0.70, le test conserve l'hypothèse nulle au seuil $\alpha = 5\%$, et décide donc que $\mu_{acc} = \mu_{non-acc}$. L'erreur associée à cette décision est une erreur de seconde espèce, que l'on peut évaluer à l'aide d'une étude de la puissance $1 - \beta$.

Le calcul de la puissance du test fournit : $1 - \beta = 3.2\%$. La valeur de la puissance est donc bien trop faible pour que la décision prise par le test ait une quelconque valeur. On ne donc conclure ni dans un sens, ni dans l'autre.

Cependant, nous pouvons calculer le nombre n1 d'accidents à inclure afin que la puissance obtenue soit satisfaisante, égale à 80%. Ce calcul conduit à : $n1 = 6057$.

Si l'on avait voulu détecter de manière significative à la puissance 80% une différence de pPICUD0 égale 3%, il aurait fallu inclure : n1=93 accidents.

Optimisation du seuil de coupure

Les tableaux des seuils et les graphes associés aux méthodes d'optimisation sont disponibles dans les annexe E.8. et E.9. On constate, de manière similaire à l'étude précédente, que les valeurs de certaines variables ne sont pas satisfaisantes et que les graphes sont passablement dégénérés. Par ailleurs, les deux premiers points des graphes, qui paraissent isolés, ne sont pas pris en compte.

Les méthodes d'optimisation fournissent les valeurs suivantes pour les seuils optimaux :

- Méthode M.1. : elle est inapplicable dans ce contexte.
- Méthode M.2. : 0.13761905
- Méthodes M.3. et M.4. : 0.08661905
- Méthode M.5. : 0.08661905
- Méthode M.6. : 0.08661905

Le seuil de coupure définitif gardé, qui permet d'optimiser de manière satisfaisante les mesures de performances, est donc le seuil 0.08661905. Les mesures de performances associées sont les suivantes :

- Sensibilité : $1.06*10^{-5}$
- Spécificité : 0.9999957
- TauxFaux1 : 20%
- TauxFaux0 : 61.8%
- IndYouden et IndYoudenPond : 0.18 et 0.68
- OR : 2.473578
- c : 0.5000032

L'interprétation de ces valeurs est la même que pour l'étude des accidents isolés.

Régression logistique et test de Wald

Les sorties R associées sont disponibles dans l'annexe E.7.

Le calibrage du modèle de régression logistique, expliquant la variable binaire indicatrice d'accident « IndAcc » à l'aide de la variable binarisée associée à « pPICUD0 » fournit l'estimation suivante du coefficient β_1 : 0.9057. Nous vérifions aisément qu'en prenant l'exponentielle de cette valeur, nous retrouvons bien l'estimation de l'Odd Ratio obtenue : $e^{0.9057} = 2.473578$.

L'intervalle de confiance asymptotique au seuil $\alpha = 5\%$ pour β_1 est alors :

$$IC_\alpha(\beta_1) = [\,-0.64\,;\,2.46]$$

En passant à l'exponentiel, on obtient donc un intervalle de confiance asymptotique au seuil $\alpha = 5\%$ pour l'Odd Ratio, qui contient encore une fois la valeur 1 :

$$IC_\alpha(OR) = [\,0.53\,;\,11.65]$$

De même que dans le cas de l'étude des accidents isolés, le test de Wald conserve l'hypothèse nulle au seuil $\alpha = 5\%$.

Le calcul de la puissance du test fournit : $1 - \beta = 20.8\%$, qui est bien trop faible pour conclure à une effective absence d'influence de la variable.

Le nombre p d'accidents à inclure afin que la puissance obtenue soit satisfaisante, égale à 80% est : $p = 60$.

Si on avait voulu détecter de manière significative un Odd Ratio de 1.2, qui correspond donc à un coefficient $\beta_1 = \ln(1{,}2) = 0.182$, le nombre p d'accidents à inclure afin que la puissance du test soit égale à 80% aurait été : $p = 1476$.

En conclusion, comme nous venons de le voir, nous n'avons pas pu mener à bout l'étude de la pertinence des variables « vitesse moyenne » et « pPICUD0 », en raison d'un nombre trop faible d'accidents inclus dans l'étude. Par ailleurs, pour la même raison, nous n'avons pas pu illustrer de manière satisfaisante les méthodes d'optimisation du choix du seuil de coupure : certaines méthodes étaient inapplicables, et la plupart des graphes étaient dégénérés. Pour remédier à cela, nous allons réaliser dans la section 5.3.3. l'étude croisée des variables microscopiques « pTIV2 » et « pPICUD0 », dont le principal intérêt est de permettre d'illustrer correctement les méthodes d'optimisation. Les résultats obtenus, en termes de significativité de la pertinence de la discrimination, ne nous intéressent donc pas dans le cadre de cette étude.

5.3.3. Étude croisée de deux indicateurs

Considérons à présent l'idée suivante : au lieu de considérer uniquement les situations d'accident, dont le nombre est très faible, nous pouvons considérer les situations « à risque d'accident » pour un indicateur donné. Considérons ainsi l'indicateur de risque « pTIV2 », dont la pertinence est supposée validée. Les accidents étant a priori corrélés aux valeurs élevées de pTIV2, le seuil de coupure permettant de binariser la variable pTIV2 est choisi à 0.90.

Nous dirons alors qu'un cycle correspond à une « situation d'accident » si la valeur de pTIV2 associée au cycle est supérieure à 0.90. Dans ce contexte, la variable pTIV2, ainsi binarisée, jouera le rôle de l'indicatrice d'accident « IndAcc ».

Comme lors des deux études précédentes, nous allons alors tester la qualité d'un indicateur de risque, par exemple « pPICUD0 ».

Test de Mann-Whitney

Les deux moyennes à comparer sont : $\mu_{acc} = 0.66$ et $\mu_{non-acc} = 0.13$.

L'écart-type des observations, estimé sur les données correspondant aux situations de non-accident, est : $\sigma = 0.103$.

Le résultat du test, disponible dans l'annexe E.10., est le suivant : la p-valeur étant égal à 0, le test rejette l'hypothèse nulle au seuil $\alpha = 5\%$, et décide donc que $\mu_{acc} \neq \mu_{non-acc}$. L'erreur associée à cette décision est l'erreur de première espèce $\alpha = 5\%$. Ce résultat indique donc que la discrimination des situations d'accident et de non-accident est satisfaisante.

Optimisation du seuil de coupure

Les tableaux des seuils et les graphes associés aux méthodes d'optimisation sont disponibles dans les annexe E.11. et E.12.

Les méthodes d'optimisation du choix du seuil de coupure fournissent les valeurs des seuils optimaux :

- Méthode M.1. : 0.62919192
- Méthode M.2. : 0.33484848
- Méthodes M.3. et M.4. : 0.66717172
- Méthode M.5. : 0.32535354
- Méthode M.6. : 0.66717172

Ainsi, on obtient globalement deux seuils optimaux différents. Parmi ces deux seuils, le seuil définitif gardé est 0.66717172. En effet, c'est celui qui correspond à la fois à la maximisation de l'Odd Ratio (méthode M.5) et du coefficient c (méthode M.6), qui sont les deux mesures de performances les plus importantes. Par ailleurs, ce seuil correspond aussi à l'optimisation de la sensibilité et de la spécificité. Les mesures de performances associées au choix de ce seuil sont très bonnes, même en termes de taux de fausse alarme et de taux de carence d'alarme :

- Sensibilité : 58.4%
- Spécificité : 99.9%
- TauxFaux1 : 28.9%
- TauxFaux0 : $8.44*10^{-5}$
- IndYouden et IndYoudenPond : 0.71 et 0.71
- OR : 29114.9
- c : 0.792

Le coefficient c vaut environ 79%, ce qui signifie que la discrimination des situations d'accident et de non-accident est très bonne. La valeur très élevée de l'Odd Ratio (29114.9) confirme a priori cela.

Régression logistique et test de Wald

Les sorties R associées sont disponibles dans l'annexe E.10.

Le calibrage du modèle de régression logistique fournit l'estimation suivante du coefficient β_1 : 10.27901. Nous vérifions aisément qu'en prenant l'exponentielle de cette valeur, nous retrouvons bien environ l'estimation de l'Odd Ratio obtenue : $e^{10.27901} = 29115$.

L'intervalle de confiance asymptotique au seuil $\alpha = 5\%$ pour β_1 est alors :

$$IC_\alpha(\beta_1) = [\ 9.92\ ;\ 10.64]$$

L'intervalle de confiance asymptotique au seuil $\alpha = 5\%$ pour l'Odd Ratio est alors :

$$IC_\alpha(OR) = [\ 20332.4\ ;\ 41690.9]$$

L'intervalle de confiance ne contient pas la valeur 1, cela signifie que le test de Wald décidera de rejeter l'hypothèse nulle. En effet, le résultat du test est le suivant : la p-valeur étant égal à 0, le test rejette l'hypothèse nulle au seuil $\alpha = 5\%$, et décide donc que $\beta_1 \neq 0$, ce qui signifie que l'Odd Ratio est significativement différent de 1. L'erreur associée à cette décision est l'erreur de première espèce $\alpha = 5\%$.

6. Conclusion

L'objet de cette étude a été d'explorer, à travers l'analyse de la pertinence d'indicateurs de risque judicieusement élaborés, le lien entre la survenue des accidents et les caractéristiques cinématiques des véhicules. Ce type d'étude contribue à améliorer nos connaissances en matière de sécurité routière et de prévention des accidents, et s'insère dans le débat émergeant relatif à l'évolution de la législation existante.

L'étude a été réalisée à partir d'une base très volumineuse de données microscopiques de trafic, issues des données du réseau autoroutier « Marius » situé autour de Marseille. Les données d'accident, quant à elles, proviennent des BAAC (Bulletins d'Analyse des Accidents Corporels) établis par les forces de l'ordre lors des accidents.

Le traitement informatique préliminaire des données de trafic (tri, création des variables de trafic, etc.) a été long et fastidieux, surtout en termes de temps de calcul. Nous n'avons donc pu traiter, pour l'étude concrète réalisée dans le rapport, qu'un seul mois de données parmi les 150 giga-octets de données relatives aux douze mois de données de trafic disponibles.

Deux types de variables ont été étudiés : des variables de type macroscopique (débit, taux d'occupation, vitesse moyenne) qui témoignent de la situation ambiante du trafic sur une période donnée, et des variables de type microscopique (temps intervéhiculaires, vitesses relatives, etc.) modélisant le comportement, parfois accidentogène, des conducteurs. Nous sommes parvenus à exhiber des corrélations entre certaines variables microscopiques agrégées, comme la proportion de temps intervéhiculaires inférieurs à 2 secondes (pTIV2) et la proportion de PICUD négatifs (pPICUD0). La variable PICUD est une variable complexe, qui correspond à une prédiction de la distance intervéhiculaire finale de deux véhicules consécutifs, en cas d'arrêt brutal du premier véhicule. C'est une variable encore peu étudiée, contrairement aux temps intervéhiculaire ; ce résultat de corrélation est donc intéressant en soi. Nous avons, par ailleurs, mis à jour de nombreuses corrélations, en situation hors-congestion, entre variables macroscopiques et variables microscopiques. Cela indique que les variables microscopiques qui nous intéressent contiennent, en plus des informations permettant de modéliser le comportement individuel des conducteurs, des informations de type macroscopique.

Le cœur de l'étude a été de mettre au point des méthodes statistiques permettant d'évaluer la pertinence de la discrimination des situations d'accident et des situations de non-accident réalisée par les variables élaborées. Pour cela, nous avons dû distinguer deux grands types d'accident : les accidents isolés, impliquant un seul véhicule, et les accidents non-isolés, impliquant au moins deux véhicules. La vitesse moyenne est, a priori, la seule variable pertinente dans la discrimination des accidents de type isolés. C'est donc, pour ce type d'accident, la seule que nous avons étudiée. Toutes les variables microscopiques agrégées, quant à elles, tendent à quantifier le risque d'accident de type non-isolés. Pour ce type d'accident, nous avons uniquement étudié la variable pPICUD0, le but de ce rapport étant avant tout de présenter les méthodes statistiques utilisées.

Deux approches ont été testées pour évaluer la qualité des variables, en termes de discrimination des situations d'accident. La première consiste, à l'aide du test non-paramétrique de Mann-Whitney, à comparer les deux moyennes de la variable testée, correspondant aux situations d'accident et de non-accident. La deuxième approche repose sur l'étude de tableaux de contingence obtenus en croisant une variable « indicatrice d'accident » et la variable testée binarisée. Il s'agit alors d'élaborer des méthodes statistiques permettant d'optimiser des mesures de performances de la discrimination (telles que l'Odd Ratio ou le coefficient c), associées à un modèle de régression logistique sous-jacent. Un test de Wald a, par ailleurs, été mis en œuvre pour juger de la significativité statistique de l'Odd Ratio.

Les résultats obtenus sont les suivants : concernant l'application du test de Mann-Whitney pour l'étude des accidents isolés ou celle des accidents non-isolés, le test est non significatif au seuil $\alpha = 5\%$, ce qui infirmerait la qualité de discrimination des variables testées (vitesse moyenne et pPICUD0). Or, une analyse de la puissance des deux tests réalisés nous indique que celle-ci est bien trop faible (12,4% et 3.2%) pour pouvoir conclure quoi que ce soit de ces deux tests. Nous avons cependant calculé, en particulier, les nombres d'accidents nécessaires pour pouvoir détecter, avec une puissance de 80%, une différence significative de vitesse (pour les accidents isolés) égale à 3km/h, ou une différence significative de la variable pPICUD0 de 3% (pour les accidents non-isolés). Dans le premier cas, 174 accidents sont nécessaires, et 93 dans le deuxième cas.

Nous avons aussi calculé, pour chacune de ces deux variables, les valeurs optimisées des mesures de performances. Bien que les Odd Ratio semblent plutôt élevés (4.02 pour la variable « vitesse moyenne » dans le cas des accidents isolés, et 2.47 pour la variable « pPICUD0 » dans le cas des accidents non-isolés), le coefficient c, égal à 0.5, révèle une discrimination très médiocre, voire inexistante. Bien que les valeurs des Odd Ratios semblent importantes, ce qui compte le plus est leur niveau de significativité. Des tests de Wald, dont l'objet est précisément de juger de la significativité d'un Odd Ratio, ont donc été mis en œuvre. Dans les deux cas, le test n'est pas significatif au seuil $\alpha = 5\%$. Mais, comme avant, en analysant la puissance de ces deux tests, on s'aperçoit qu'elle est bien trop faible (22.5% dans le cas des accidents isolés, et 20.8% dans le cas des accidents non-isolés) pour que l'on puisse conclure à la non-pertinence des variables testées. Une étude de dimensionnement fournit les nombres d'accidents qu'il faudrait inclure pour pouvoir détecter, avec une puissance de 80%, un Odd Ratio valant 1.2 ; 1260 accidents sont nécessaires dans le cas des accidents isolés, et 1476 dans le cas des accidents non-isolés.

En résumé, et il s'agit là du résultat principal de l'étude (ou pré-étude) qui a été menée, pour pouvoir conclure, avec une puissance de 80%, de la pertinence ou non des indicateurs de risque élaborés, un ordre de grandeur de 100 à 150 accidents, correspondant environ à une année de données microscopiques de trafic, est nécessaire si on se sert du test de Mann-Whitney. Si on se base sur les mesures de performances et le test de Wald, encore environ dix fois plus d'accidents (de 1000 à 1500), correspondant donc à une dizaine d'années de données microscopiques de trafic, sont nécessaires.

Une des solutions qui permettrait d'augmenter le nombre d'accidents traités serait d'utiliser des données de type macroscopique. Ce type de données, moins riche en termes d'informations, est cependant beaucoup plus compact et donc plus facile à traiter d'un point de vue informatique. Et, les corrélations que nous avons établies entre variables microscopiques et macroscopiques semblent

indiquer, tout du moins en situation courante de non-accidents, que ces deux types de variables sont très liés, ce qui permettrait de substituer une variable macroscopique pertinente (comme le débit) à une variable microscopique corrélée, dont la pertinence en termes de mesure de risque d'accident a été validée.

Enfin, il faut noter que l'approche que nous avons testée, consistant à découper les journées en 240 cycles de six minutes et à associer, à chaque accident, un cycle correspondant à une situation moyenne d'avant accident, pourrait être optimisée en envisageant l'idée suivante : nous pourrions tenter de repérer sur les données de trafic, dans le cadre des accidents non-isolés, le peloton de véhicules dans lequel était inséré le véhicule avant l'accident, et de quantifier alors le risque d'accident à l'aide de la situation cinématique moyenne de ce peloton de véhicules. L'estimation du risque serait alors bien plus précise.

Tables des annexes

Annexe A.1. : Extrait des données de trafic bruts prélevées dans le fichier « 031454.dat » 61
Annexe A.2. : Extrait des quatre fichiers d'accidents 62
Annexe B.1. : Démonstration de la formule du temps de collision (TTC) 63
Annexe B.2. : Démonstration de la formule du PICUD 63
Annexe B.3. : Graphes des corrélations entre variables microscopiques 65
Annexe B.4. : Régression linéaire entre pPICUD0 et pTIV2 68
Annexe B.5. : Statistiques descriptives sur une journée de données 69
Annexe B.6. : Graphe des corrélations entre variables macroscopiques 71
Annexe B.7. : Graphes de la relation entre débit moyen et temps intervéhiculaire moyen 72
Annexe B.8. : Régression linéaire entre débit moyen et inverse du temps intervéhiculaire moyen 73
Annexe B.9. : Graphe des corrélations entre variables microscopiques et macroscopiques 74
Annexe B.10. : Exemples de graphes de corrélations entre variables microscopiques et macroscopiques en situation de congestion 76
Annexe C.1. : Extrait de la table SAS de synthèse des données d'accidents 77
Annexe C.2. : Code SAS de la réalisation de la table de synthèse des données d'accidents 78
Annexe C.3. : Extrait du fichier « M1as1000430C20090712.txt » 81
Annexe C.4. : Extrait du fichier « M1as1000430C20090712CYCLES.txt » 82
Annexe C.5. : Code R de création des fichiers de données agrégées par cycles 84
Annexe C.6. : Extrait du ficher «M7is1264687C20090706CYCLES.txt » enrichi des informations d'accidents 87
Annexe C.7. : Extrait du fichier « acc_baac_marius.csv » 87
Annexe C.8. : Code R du programme d'insertion des données relatives aux accidents dans les données de trafic 89
Annexe C.9. : Code R des programmes de listings et d'agrégation bout-à-bout des fichiers 90
Annexe D.1. : Graphe de la fonction logit 91
Annexe D.2. : Développements des calculs de l'estimation des paramètres 91
Annexe D.3. : Puissance du test de Wald 93
Annexe D.4. : Odd Ratio, risques relatifs et lien avec le coefficient $\beta 1$ de la régression 95
Annexe D.5. : Détails de la construction du coefficient « c » 96
Annexe E.1. : Construction du test de Mann-Whitney 97
Annexe E.2. : Approximation de la puissance du test de Mann-Whitney 98
Annexe E.3. : Code R de création du tableau des seuils et des graphes de performances 100
Annexe E.4. : Résultats des tests dans le cadre de l'étude des accidents isolés 102
Annexe E.5. : Tableau des seuils pour l'étude des accidents isolés 103
Annexe E.6. : Graphes des performances pour l'étude des accidents isolés 105
Annexe E.7. : Résultats des tests dans le cadre de l'étude des accidents non-isolés 106
Annexe E.8. : Tableau des seuils pour l'étude des accidents non-isolés 107
Annexe E.9. : Graphes des performances pour l'étude des accidents non-isolés 109
Annexe E.10. : Résultats des tests dans le cadre de l'étude croisée entre pPICUD0 et pTIV2 110
Annexe E.11. : Tableau des seuils pour l'étude croisée entre pPICUD0 et pTIV2 111
Annexe E.12. : Graphes des performances pour l'étude croisée entre pPICUD0 et pTIV2 113

Annexes

Annexe A.1. : Extrait des données de trafic bruts prélevées dans le fichier « 031454.dat »

```
317#M3q;8;285;Ve;03/07/09;14:54;5729;D;93;42;NEUTRE; ;
317#M3q;8;285;Ve;03/07/09;14:54;5743;C;99;41;NEUTRE; ;
504#M5d;1;439;Ve;03/07/09;14:54;1903;C;110;32;NEUTRE; ;
504#M5d;1;439;Ve;03/07/09;14:54;1973;D;88;37;NEUTRE; ;
504#M5d;1;439;Ve;03/07/09;14:54;2308;D;72;22;NEUTRE; ;
304#M3d;1;871;Ve;03/07/09;14:54;5411;C;73;24;NEUTRE; ;
304#M3d;1;871;Ve;03/07/09;14:54;5473;G;83;36;NEUTRE; ;
304#M3d;1;871;Ve;03/07/09;14:54;5609;D;61;39;NEUTRE; ;
304#M3d;1;871;Ve;03/07/09;14:54;5624;C;86;53;NEUTRE; ;
304#M3d;1;871;Ve;03/07/09;14:54;5787;C;78;37;NEUTRE; ;
151#M1A;0;430;Ve;03/07/09;14:54;5603;C;103;39;NEUTRE; ;
151#M1A;0;430;Ve;03/07/09;14:54;5721;C;110;40;NEUTRE; ;
151#M1A;0;430;Ve;03/07/09;14:54;5726;D;90;47;NEUTRE; ;
```

Chacune des lignes de ces données de trafic correspond à l'enregistrement des données relatives au passage d'un véhicule. Les données sont les suivantes :

- 1ère colonne : c'est l'identifiant du capteur qui a pris la mesure.
- 2ème et 3ème colonne : c'est le point routier du capteur qui a pris la mesure. Par exemple, « 8 ; 285 » correspond au point routier 8,285 km.
- 4ème et 5ème colonne : le jour et la date correspondants.
- 6ème colonne : l'heure et la minute correspondantes au passage du véhicule.
- 7ème colonne : la seconde, le dixième de seconde et le centième de seconde correspondant au passage du véhicule. Par exemple, « 5729 » correspond à 57 secondes et 29 centièmes.
- 8ème colonne : c'est la voie de circulation (D pour voie de droite, G pour voie de gauche et C pour voie centrale).
- 9ème colonne : c'est la mesure de la vitesse en km/h du véhicule.
- 10ème colonne : c'est la mesure de la longueur en décimètre du véhicule.

Annexe A.2. : Extrait des quatre fichiers d'accidents

	A	B	C	D	E	F	G	H
1	Identifiant N1	Date accident Niveau 1	NUMPV	Code postal	Code unité	Organisme	Date accident Niveau 1	Type de jour
2	543831	20090114	15	13170	973531	3	20090114	Mercredi
3								
4	Semaine/Week-end	HEURE	Tranches horaires (6)	Luminosité	Département	COMMUNE	Code INSEE	Code postal
5	Semaine	1811	15h à 18h59	Nuit sans éclairage public	BOUCHES-DU-RHONE	71	13071	13170
6								
7	Commune	En ou hors agglomération	INTERSECTION	ONDITION ATMOSPHERIQU	LIEU	LATITUDE	LONGITUDE	
8	LES PENN	Hors agglomération	hors intersection	Normale	A7-SENS	4340670	532882	
9								

Extrait du 1er fichier d'accident

	A	B	C	D	E	F	G	H
1	Identifiant N2	CODE ROUTE	Catégorie de route	VOIE	Bis ou Ter	Lettre indice	Régime de circulation	Nombre de voies de circulation
2	543831	1	Autoroute	7			à sens unique	3
3								
4	MARQUAGE CHAUSSEE	VOIES SPECIALES	Profil en long de la route	BORNE	OINT ROUTIE	TRACE EN PLAN	ETAT ROUTE	ETAT SURFACE
5			Plat	269	0	Partie rectiligne		normale
6								
7	AMENAGEMENT INFRASTUCTURE	SIGNALISATION	ENVIRONNEMENT	Situation de l'accident				
8				sur chaussée				

Extrait du 2ème fichier d'accident

Identifiant N3	IDENT3	Date accident Niveau 3	Sens de circulation	Catégorie de véhicules 4 classes
543831	1	20090114	PR Croissants	Véhicules légers
543831	2	20090114	PR Croissants	Véhicules légers
543831	3	20090114	PR Croissants	Véhicules légers
Année de mise en circulation	Ancienneté du véhicule (5)	Année du contrôle technique	Validité du contrôle technique	PUISSANCE DU VEHICULE
19990901	Plus de 8 ans	19	Contrôle tech valide	
19891201	Plus de 8 ans	19	Contrôle tech valide	
20040101	de 4 à 5 ans	19	Contrôle tech valide	
VEHICULE SPECIAL	ASSURANCE	OBSTACLE FIXE HEURTE		Obstacle mobile heurté
	Oui			Véhicule
	Oui			Véhicule
	Oui			Véhicule
MANOEUVRE PRINCIPALE AVANT L'ACCIDENT	POINT DE CHOC INITIAL	NOMBRE OCCUPANT TC	CNIT	Facteur lié au véhicule
meme sens,meme file	arrière	0	MJT5202P	
meme sens,meme file	avant	0	WFOAXXGC	
meme sens,meme file	arrière	0	MPE5204K	
CATEVEH40	IMMATRICULATION	APPARTENANT A		
VL seul	4	conducteur		
VL seul	4	conducteur		
VL seul	4	conducteur		

Extrait du 3ème fichier d'accident

Identifiant N4	Présumé responsable	Catégorie d'usager	GRAVITE	Détection et vérification de l'alcoolémie	Taux d'alcool g/l pour alcoolémies illégales
543831	A	Conducteur	Indemne	Dépistage négatif	
543831	B	Conducteur	Indemne	Dépistage négatif	
543831	C	Conducteur	Blessé léger	Dépistage négatif	
SEXE	Date naissance	AGE	Classe d'âge (7)	Catégorie socio-professionnelle	FACTEUR LIE A L'USAGER
Féminin	19491201	60	60 ans et plus	Retraité	
Masculin	19600401	49	de 40 à 59 ans	Artisan commerçant	
Féminin	19751201	34	de 30 à 39 ans	Ouvrier	
PERMIS DE CONDUIRE	Année de permis	TRAJET DU CONDUCTEUR	INFRAC1	INFRAC2	
Valide	1973	promenade-loisirs			
Valide	2007	promenade-loisirs			
Valide	2004	promenade-loisirs			
Equipement de sécurité	UTILISATION EQUIPEMENT DE SECURITE	Manoeuvre du piéton	Action du piéton	Nombre de piéton	
Ceinture	oui				
Ceinture	oui				
Ceinture	oui				

Extrait du 4ème fichier d'accident

Annexe B.1. : Démonstration de la formule du temps de collision (TTC)

$$TTC_i = \frac{v_{i-1} * TIV_i}{VR_i}$$

La différence des vitesses VR_i va entrer naturellement en jeu lors du calcul du TTC. De façon simple, nous avons : $TTC_i = \frac{DIV_i}{v_i - v_{i-1}}$, où DIV_i (Distance InterVéhiculaire) est la distance entre les deux véhicules lorsque le premier véhicule se trouve au niveau du capteur. La mesure de temps associée à cette distance intervéhiculaire est le temps intervéhiculaire. Comme nous cherchons à trouver le temps hypothétique au bout duquel le véhicule de derrière percute celui de devant, la formule pour DIV_i est : $DIV_i = v_{i-1} * TIV_i$. Ainsi, nous trouvons bien la relation annoncée.

Annexe B.2. : Démonstration de la formule du PICUD

Notons que DIV_i nous fournit la distance entre l'avant du second véhicule et l'avant du premier véhicule ; pour obtenir alors la distance entre l'avant du second véhicule et l'arrière du premier, il suffit de retrancher la longueur du premier véhicule L_{i-1}.

Le PICUD se définit par la différence de ces deux grandeurs :

$$PICUD_i = \left(D_{i-1}^{arrêt} + DIV_i - L_{i-1}\right) - D_i^{arrêt}$$

Il y a risque de collision si $D_{i-1}^{arrêt} + DIV_i - L_{i-1} < D_i^{arrêt}$, c'est-à-dire si $PICUD_i < 0$.

Exprimons les deux distances d'arrêt en fonction des données de trafic. Étant donné que c'est le véhicule $i-1$ qui freine le premier, nous obtenons :

$$D_{i-1}^{arrêt} = D_{i-1}^{freinage}$$
$$D_i^{arrêt} = D_i^{réaction} + D_i^{freinage}$$

Or, puisque nous faisons l'hypothèse que le freinage se fait à décélération constante γ, nous avons :

$$Décélération_i = \gamma$$

D'où, en intégrant une fois, puis deux fois (en prenant des constantes d'intégration nulles) :

$$v_i = \gamma * t_i$$
$$D_i^{freinage} = \frac{\gamma}{2} * {t_i}^2$$

Puis, en regroupant les deux formules, on obtient :

$$D_i^{freinage} = \frac{v_i^2}{2 * \gamma}$$

Et, de même :

$$D_{i-1}^{freinage} = \frac{v_{i-1}^2}{2 * \gamma}$$

Et, comme nous l'avons vu dans la section consacrée au TTC, nous avons :

$$DIV_i = TIV_i * v_{i-1}$$

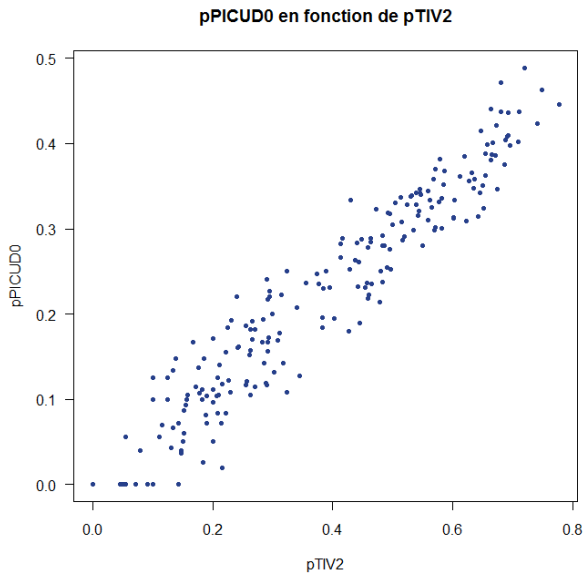
La distance parcourue par le véhicule de derrière pendant le temps de réaction se calcule aussi de manière très simple :

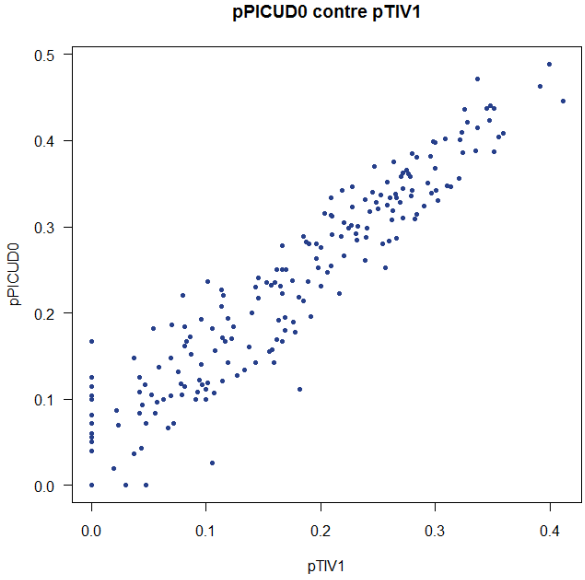
$$D_i^{réaction} = v_i * T_r$$

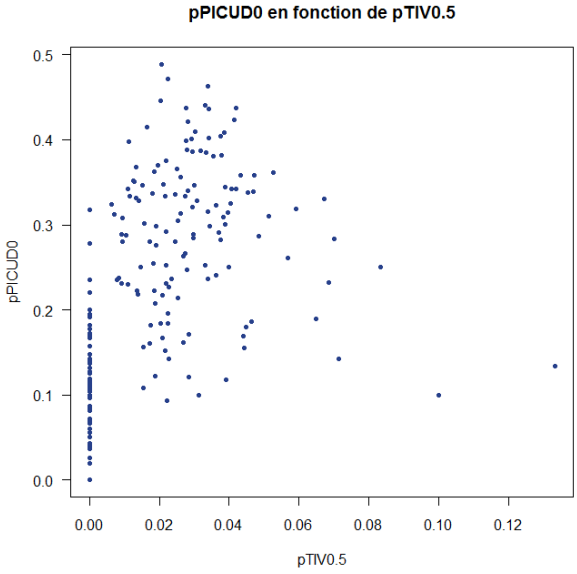
Puisque $PICUD_i = \left(D_{i-1}^{arrêt} + DIV_i - L_{i-1}\right) - D_i^{arrêt}$, nous obtenons alors la formule suivante :

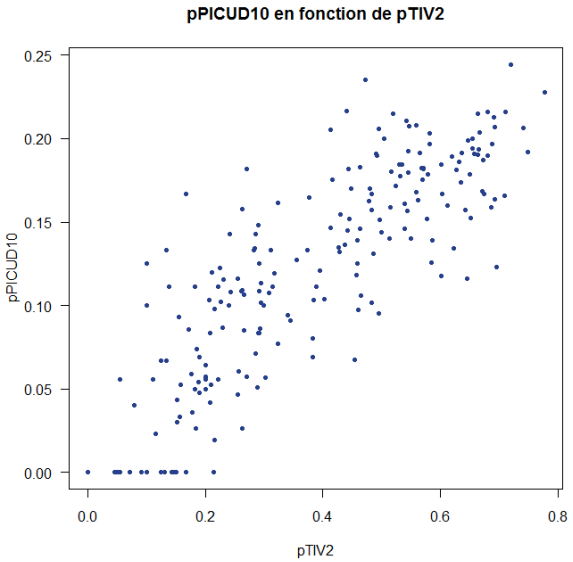
$$PICUD_i = \frac{v_{i-1}^2 - v_i^2}{2 * \gamma} + TIV_i * v_{i-1} - v_i * T_r - L_{i-1}$$

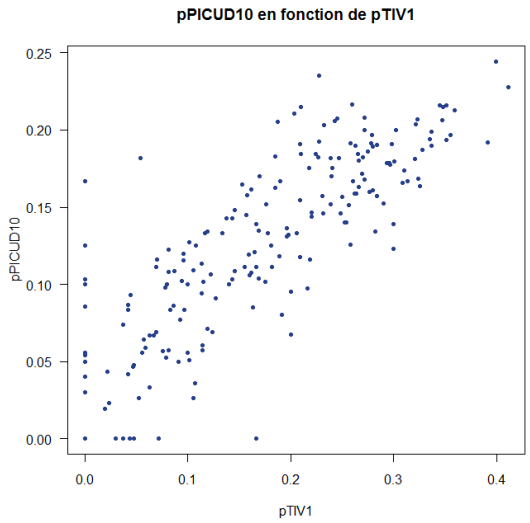
Annexe B.3. : Graphes des corrélations entre variables microscopiques

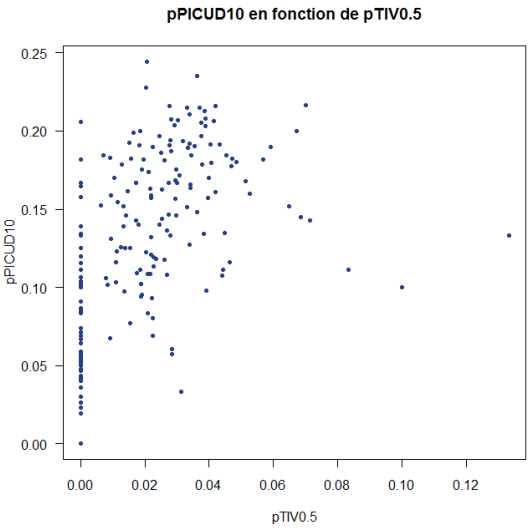


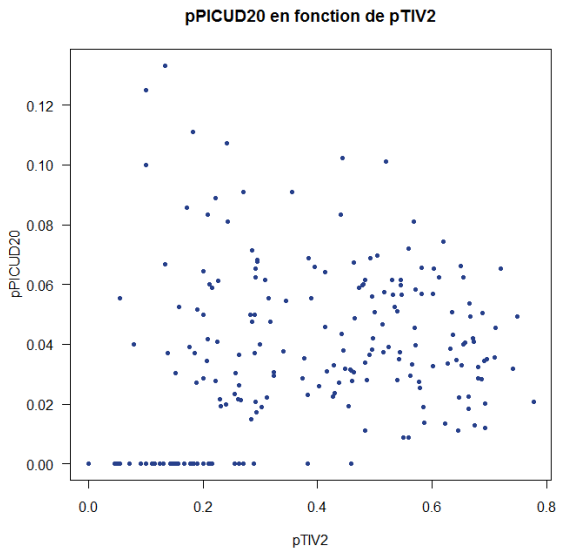


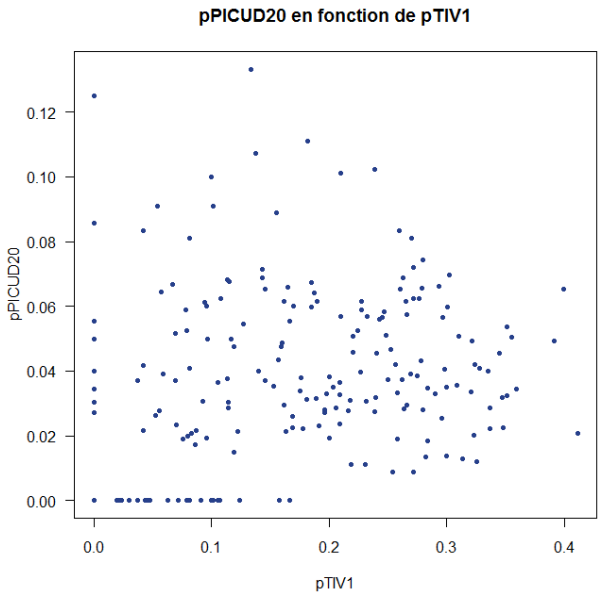


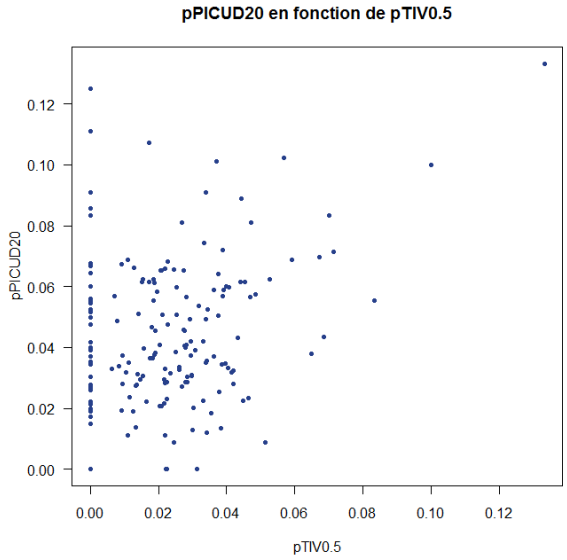


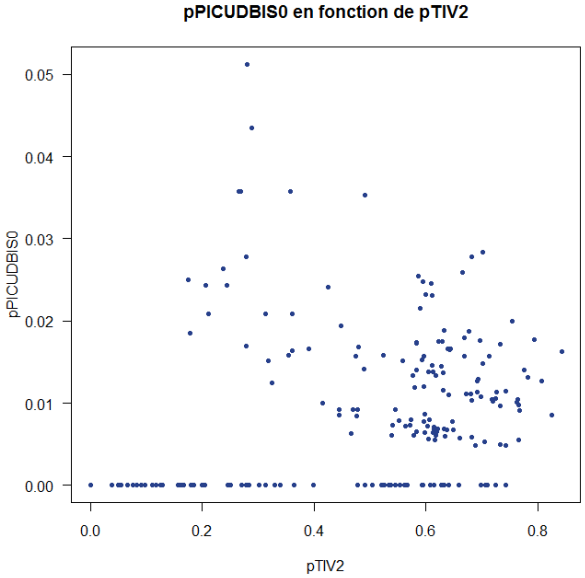


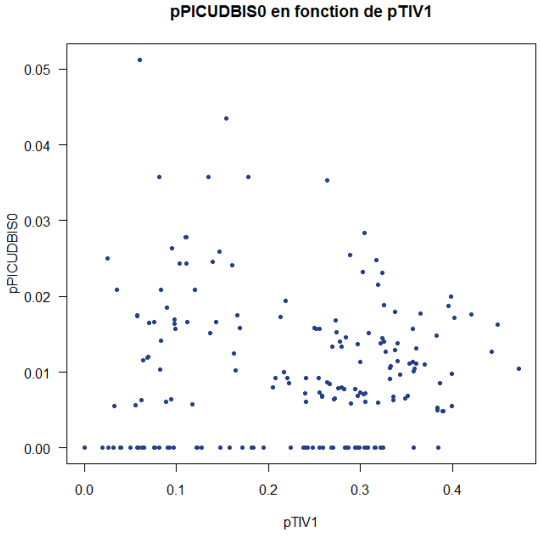


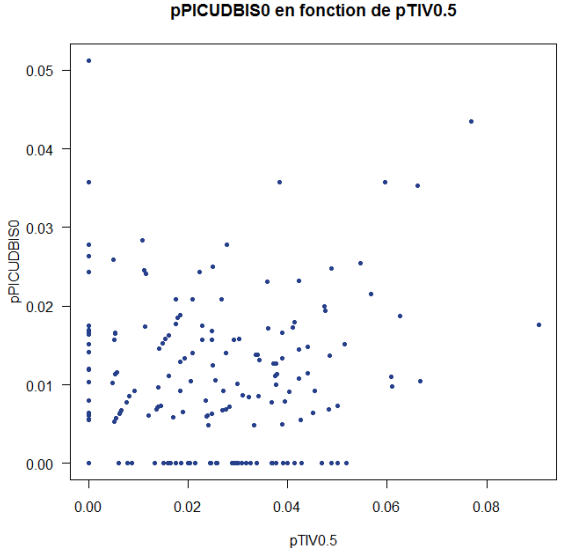

pPICUD20 en fonction de pTIV2
pPICUD20
pTIV2
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.0
0.2
0.4
0.6
0.8

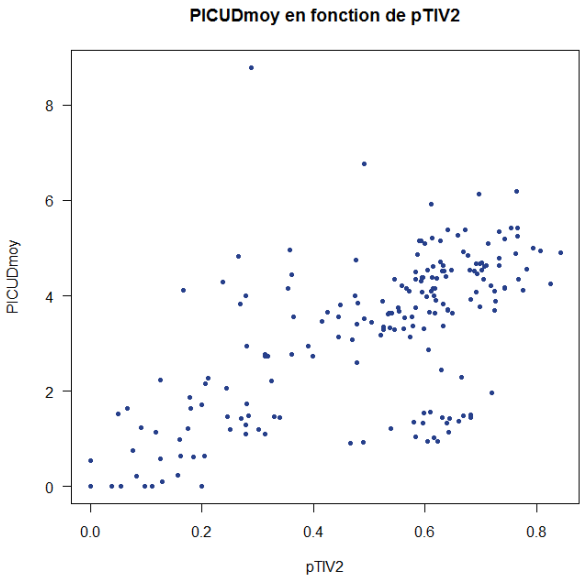

pPICUD20 en fonction de pTIV1
pPICUD20
pTIV1
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.0
0.1
0.2
0.3
0.4

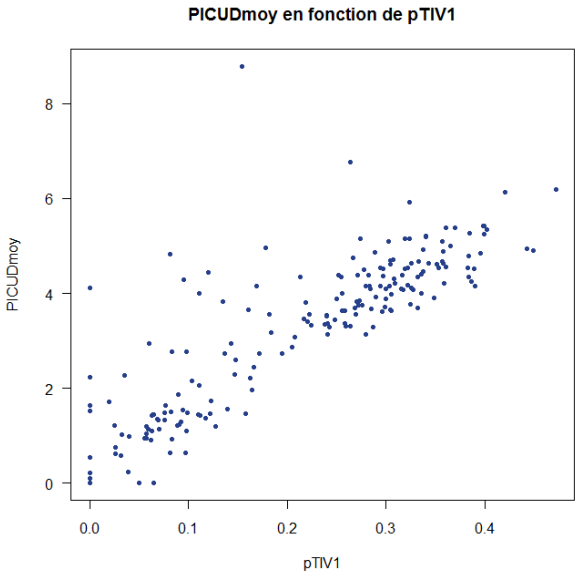

pPICUD20 en fonction de pTIV0.5
pPICUD20
pTIV0.5
0.00
0.02
0.04
0.06
0.08
0.10
0.12

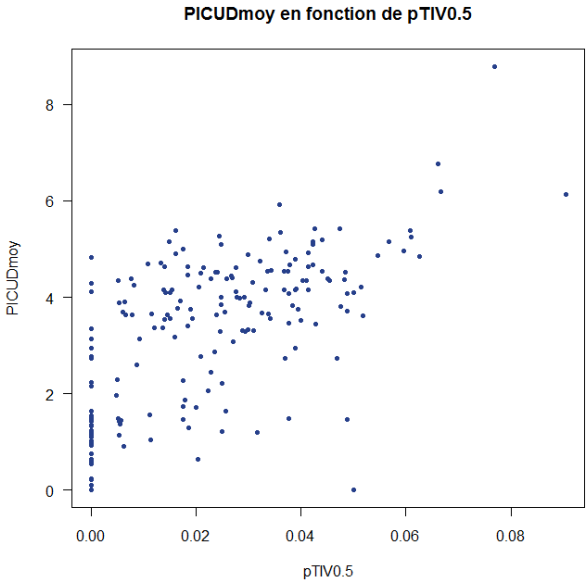

pPICUDBIS0 en fonction de pTIV2
pPICUDBIS0
pTIV2
0.00
0.01
0.02
0.03
0.04
0.05
0.0
0.2
0.4
0.6
0.8


pPICUDBIS0 en fonction de pTIV1
pPICUDBIS0
pTIV1
0.00
0.01
0.02
0.03
0.04
0.05
0.0
0.1
0.2
0.3
0.4


pPICUDBIS0 en fonction de pTIV0.5
pPICUDBIS0
pTIV0.5
0.00
0.01
0.02
0.03
0.04
0.05
0.00
0.02
0.04
0.06
0.08

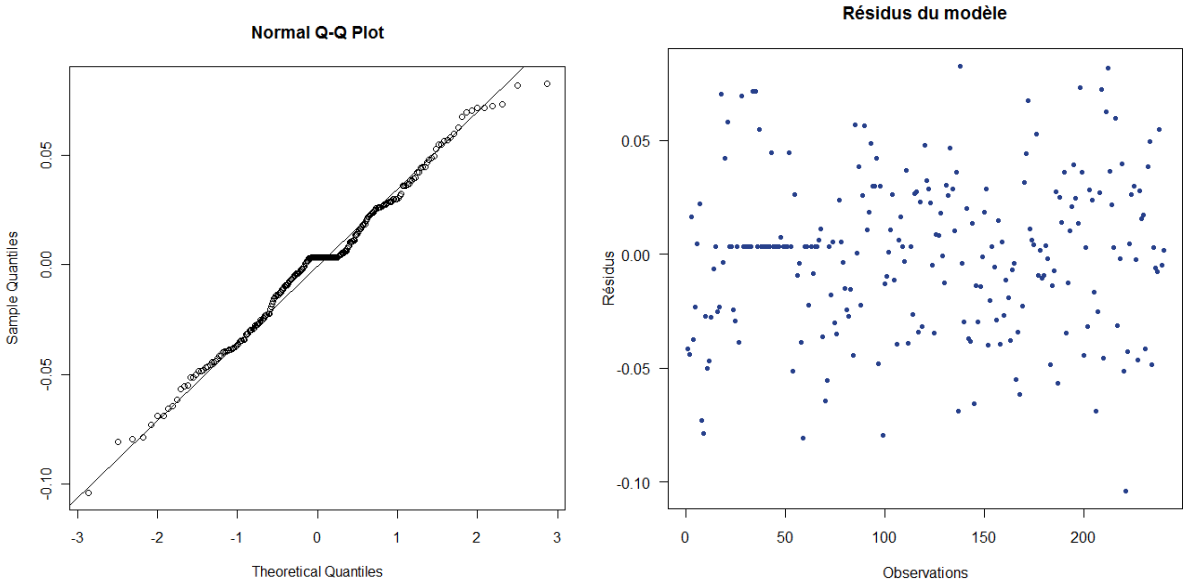

PICUDmoy en fonction de pTIV2
PICUDmoy
pTIV2
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8


PICUDmoy en fonction de pTIV1
PICUDmoy
pTIV1
0
2
4
6
8
0.0
0.1
0.2
0.3
0.4


PICUDmoy en fonction de pTIV0.5
PICUDmoy
pTIV0.5
0
2
4
6
8
0.00
0.02
0.04
0.06
0.08

Annexe B.4. : Régression linéaire entre pPICUD0 et pTIV2

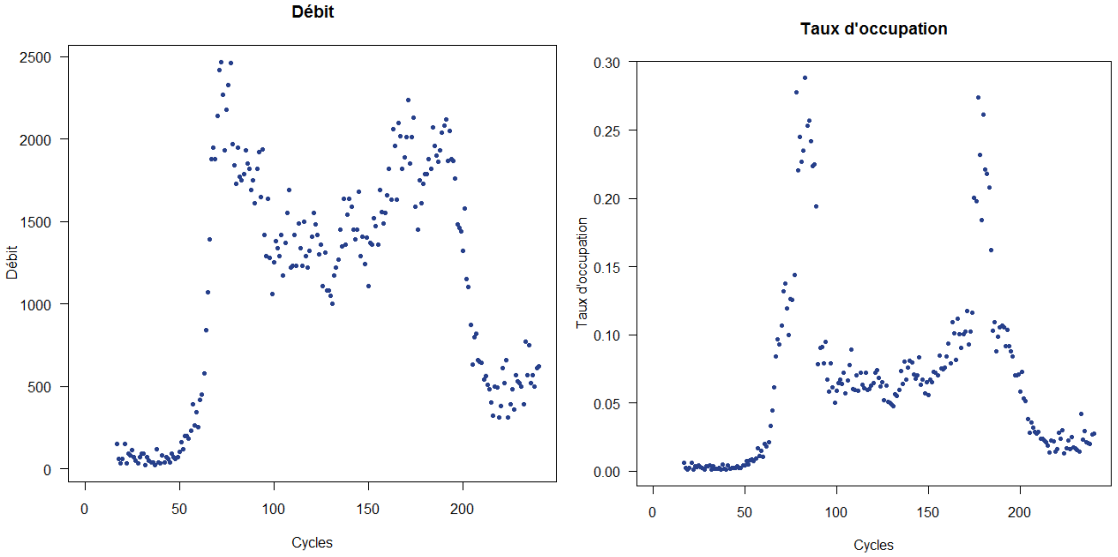
```
> shapiro.test(residuals(modele1))

        Shapiro-Wilk normality test

data:  residuals(modele1)
W = 0.9904, p-value = 0.1156
```

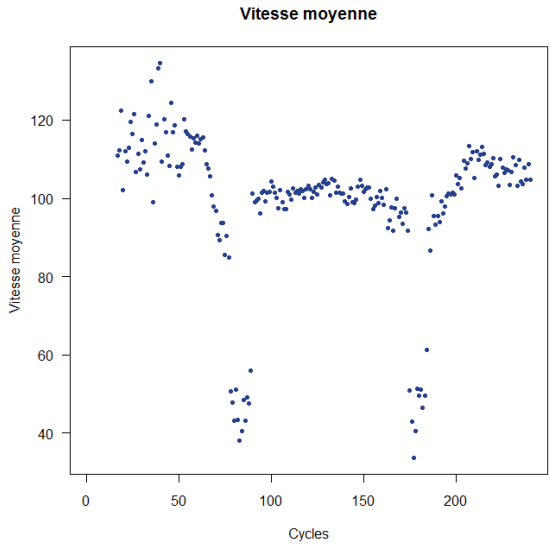



```
> summary(modele1)

Call:
lm(formula = pPICUD0 ~ pTIV2, data = donnees)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10418 -0.02454  0.00358  0.02306  0.08285

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.003580   0.004002  -0.894    0.372
pTIV2        0.590518   0.009830  60.070   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03394 on 238 degrees of freedom
Multiple R-squared: 0.9381,     Adjusted R-squared: 0.9379
F-statistic:  3608 on 1 and 238 DF,  p-value: < 2.2e-16
```

Annexe B.5. : Statistiques descriptives sur une journée de données

```
> summary(donnees)
   Numcycles          IndAcc  TypeAcc            debit          TauxOcc            vitessemoy         TIVmoy
Min.   :  1.00   Min.   :0   Mode:logical   Min.   :  20.0   Min.   :8.787e-04   Min.   : 33.62   Min.   :  1.442
1st Qu.: 60.75   1st Qu.:0   NA's:240       1st Qu.: 487.5   1st Qu.:1.948e-02   1st Qu.: 98.29   1st Qu.:  2.095
Median :120.50   Median :0                  Median :1305.0   Median :6.217e-02   Median :102.38   Median :  2.771
Mean   :120.50   Mean   :0                  Mean   :1126.4   Mean   :6.877e-02   Mean   : 99.04   Mean   : 13.583
3rd Qu.:180.25   3rd Qu.:0                  3rd Qu.:1700.0   3rd Qu.:8.932e-02   3rd Qu.:108.64   3rd Qu.:  7.310
Max.   :240.00   Max.   :0                  Max.   :2470.0   Max.   :2.887e-01   Max.   :134.67   Max.   :151.445
                                            NA's   :  16.0   NA's   :1.600e+01   NA's   : 16.00   NA's   : 16.000
     VRmoy           PICUDmoy       PICUDBISmoy          pTIV0.5            pTIV1              pTIV2
Min.   : 0.0000   Min.   : 0.000   Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.0000   Min.   : 0.0000
1st Qu.: 0.8096   1st Qu.: 1.217   1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.: 0.0624   1st Qu.: 0.2312
Median : 1.0023   Median : 3.425   Median : 0.01432   Median : 0.01754   Median : 0.2102   Median : 0.5483
Mean   : 1.3528   Mean   : 2.855   Mean   : 0.12739   Mean   : 0.01978   Mean   : 0.1882   Mean   : 0.4423
3rd Qu.: 1.5376   3rd Qu.: 4.356   3rd Qu.: 0.10390   3rd Qu.: 0.03406   3rd Qu.: 0.3049   3rd Qu.: 0.6342
Max.   : 8.2407   Max.   : 8.791   Max.   : 5.22524   Max.   : 0.09043   Max.   : 0.4718   Max.   : 0.8421
NA's   :16.0000   NA's   :16.000   NA's   :16.00000   NA's   :16.00000   NA's   :16.0000   NA's   :16.0000
    pPICUD0           pPICUD10          pPICUD20          pPICUDBIS0
Min.   : 0.0000   Min.   : 0.00000   Min.   : 0.000000   Min.   : 0.000000
1st Qu.: 0.1148   1st Qu.: 0.03101   1st Qu.: 0.005548   1st Qu.: 0.000000
Median : 0.3040   Median : 0.14384   Median : 0.029427   Median : 0.006757
Mean   : 0.2583   Mean   : 0.11910   Mean   : 0.029965   Mean   : 0.008526
3rd Qu.: 0.3894   3rd Qu.: 0.19186   3rd Qu.: 0.047426   3rd Qu.: 0.014519
Max.   : 0.5755   Max.   : 0.29843   Max.   : 0.166667   Max.   : 0.051282
NA's   :16.0000   NA's   :16.00000   NA's   :16.000000   NA's   :16.000000
```





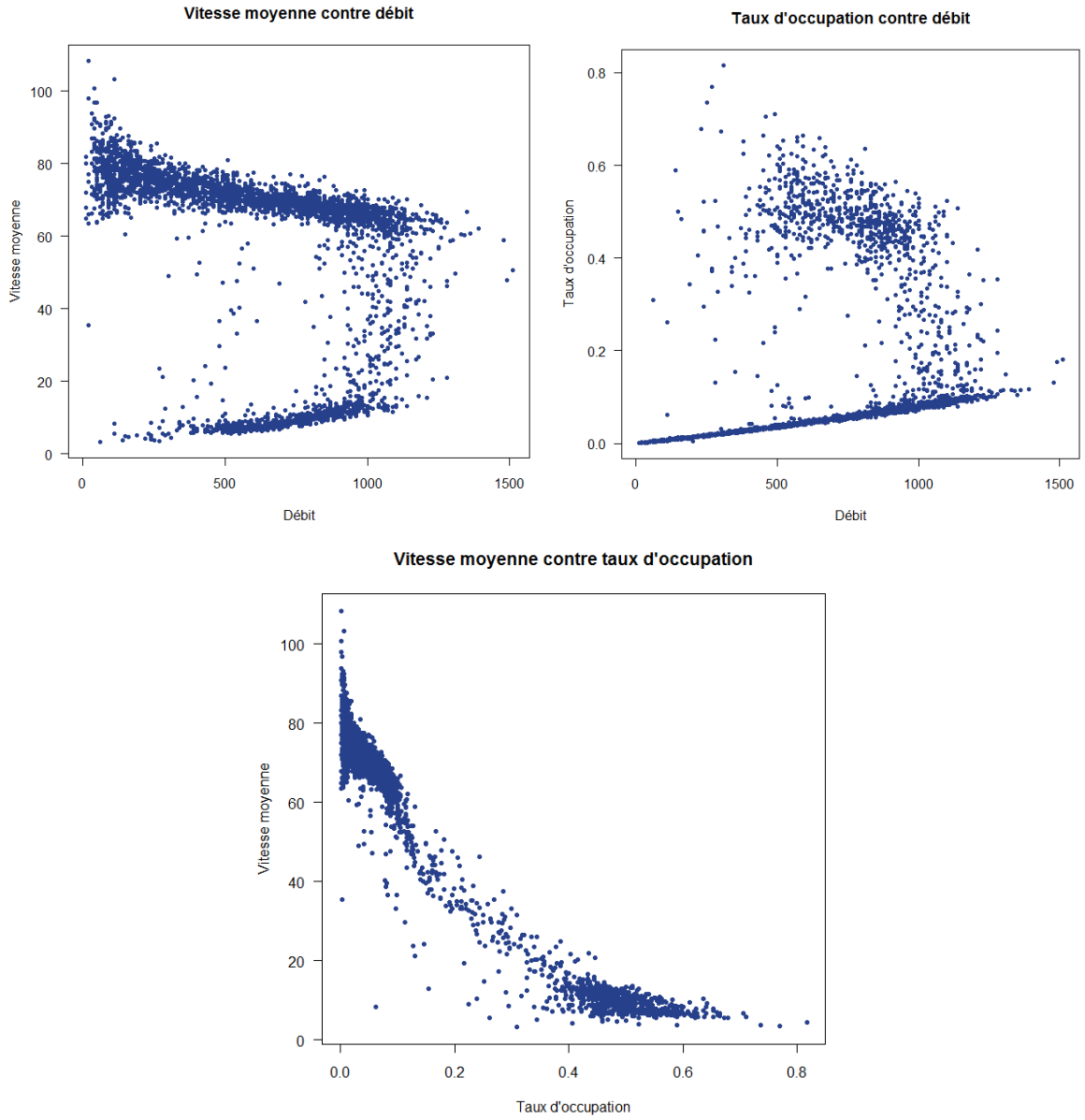
Il s'agit des graphes, sur une journée de données de trafic, des variables macroscopiques introduites. Ces données sont issues du fichier « M1as1000430C20090702CYCLES.txt ». L'analyse de ces graphes permet d'identifier clairement trois situations de trafic différentes :

- Trafic de nuit : le taux d'occupation est très faible (inférieur à 4%). Les cycles correspondant à cette situation sont les cycles 204 à 240 (fin de soirée) et 1 à 64 (pleine nuit et aube). Cette période s'étant donc de 20h24 au soir à 06h24 le matin. En pleine nuit et à l'aube, le débit moyen est très faible : entre 2 et 58 véhicules par cycle de 6 minutes, avec une moyenne de 8. En fin de soirée, le débit est un peu plus élevé : entre 31 et 82 véhicules par cycle de 6 minutes, avec une moyenne à 52. En revanche, les vitesses moyennes par cycle sont importantes : entre 99 km/h et 134.7 km/h, avec une moyenne à 114.6 km/h.

- Trafic congestionné : le taux d'occupation est supérieur à 15%. Cette situation correspond à deux séries de cycles, correspondant respectivement à la pointe du matin et à celle du soir : 78-89 (7h48-8h54) et 175-184 (17h30-18h24). Les débits sont alors très élevés : entre 145 et 247 véhicules par cycle de 6 minutes, avec une moyenne à 178. A cause du nombre trop important de voitures, le trafic est ralenti : les vitesses moyennes par cycle chutent, entre 33 km/h et 62 km/h, avec une vitesse moyenne de congestion à 48 km/h.

- Trafic ambiant : en milieu de journée, entre ces deux périodes de pointe, le taux d'occupation est compris entre 4% et 15%. Il s'agit de la situation ambiante de jour. Les débits sont modérés : entre 100 et 224 véhicules par cycle de 6 minutes, avec une moyenne à 149, ainsi que les vitesses moyennes par cycles : entre 91.7 km/h et 104.9 km/h, avec une moyenne à 100.4 km/h.

Annexe B.6. : Graphe des corrélations entre variables macroscopiques



Grâce à quelques calculs se servant des formules de la théorie du trafic, nous pouvons vérifier la validité de ces données.

La modélisation du trafic établit que le taux d'occupation s'exprime de la manière suivante :

$$TauxOcc_j = \frac{L_j + l}{v_j} * Q_j$$

où : l est la longueur de la boucle magnétique (environ 1m), L_j et v_j sont, respectivement, la longueur moyenne et la vitesse moyenne des véhicules appartenant au cycle, et Q_j est le débit moyen réel du cycle. Notons que le débit réel est égal au débit horaire, que l'on divise par 3600.

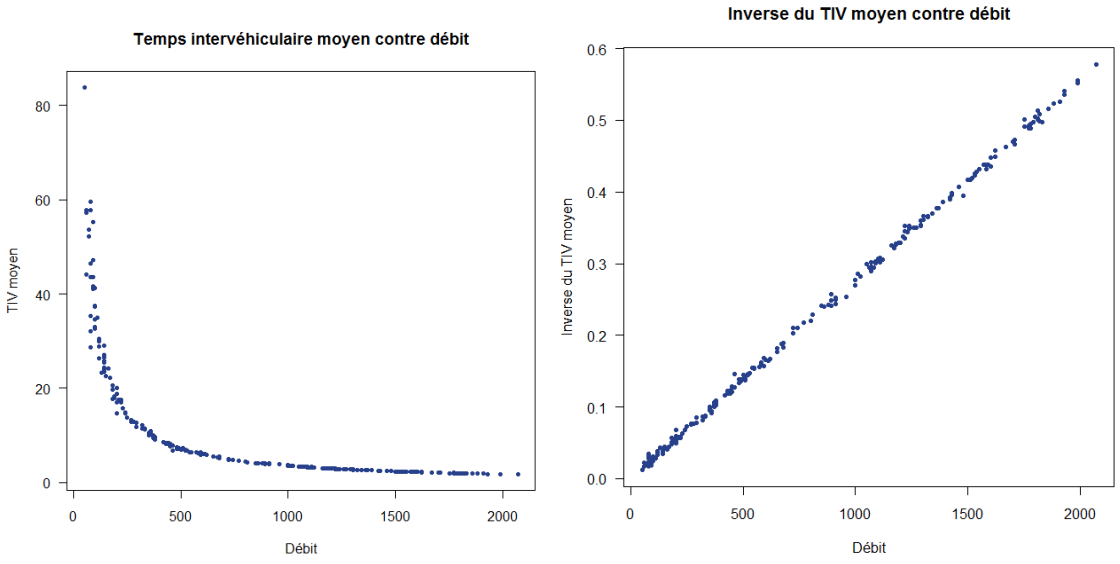
En considérant que L_j est de l'ordre de 5.5m, et si l'on suppose que la vitesse moyenne par cycles v_j est constante, ce qui est globalement le cas en situation de trafic libre, le taux d'occupation et le débit sont proportionnels. Le coefficient k de proportionnalité vaut alors : $k = \frac{L_j + l}{v_j} = \frac{6.5}{v_j}$.

Considérons donc, à présent, le graphe taux d'occupation – débit. Lorsqu'on ne se trouve pas en situation de congestion, c'est-à-dire lorsque le taux d'occupation est inférieur à 15%, on s'aperçoit que la relation semble, en effet, linéaire. La pente de la courbe, calculée à partir du graphe, vaut :

$$k' = \frac{0.114}{\frac{1300}{3600}} = 0.316$$

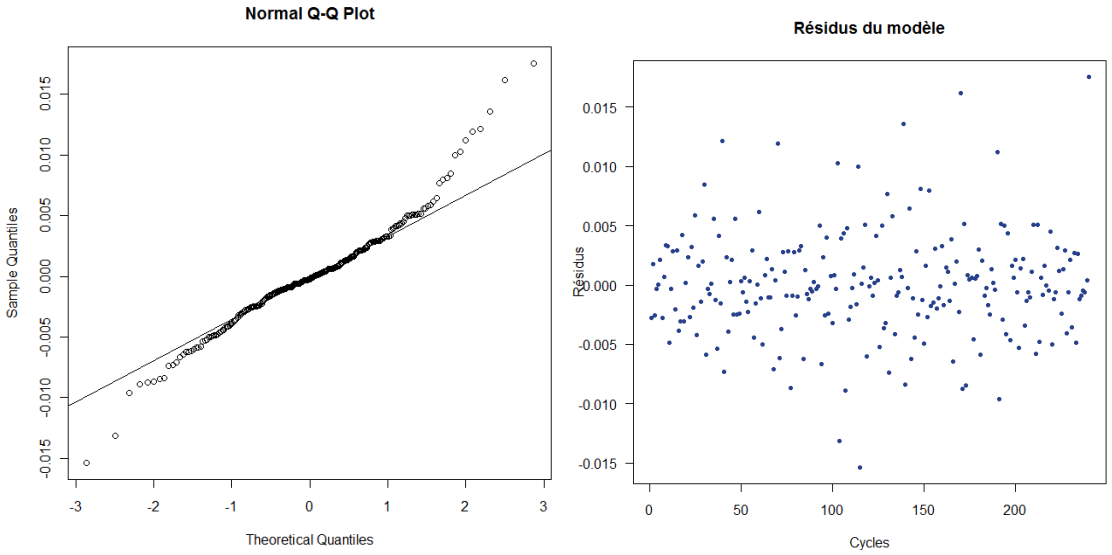
Ainsi, puisque $k = k'$, on trouve une vitesse moyenne par cycles égale à $v_j = \frac{6.5}{k'} = 20.57\ m/s$, soit environ 74 km/h. Cette valeur de vitesse correspond bien à la vitesse libre, que l'on peut lire sur le graphe vitesse – débit.

Annexe B.7. : Graphes de la relation entre débit moyen et temps intervéhiculaire moyen



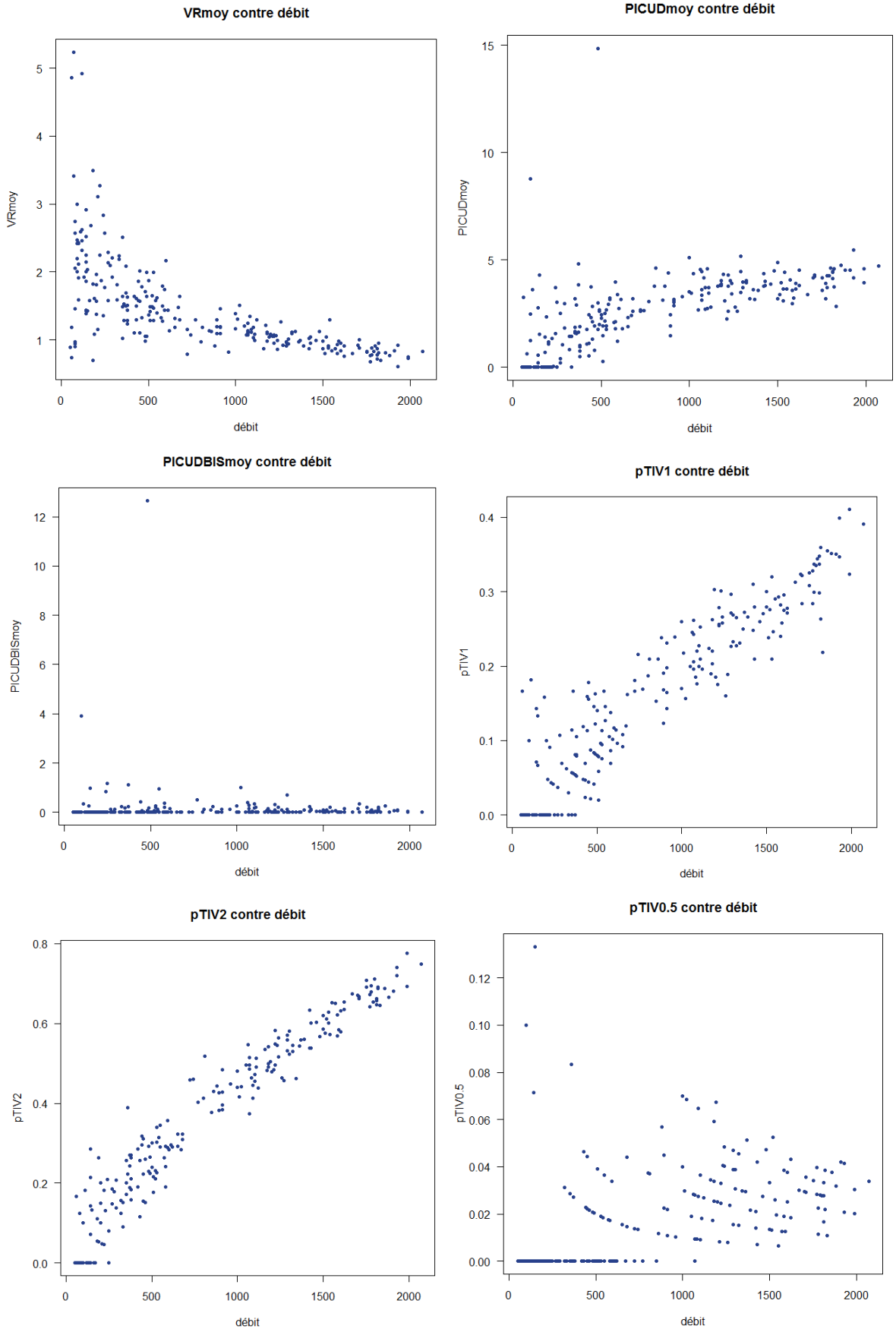
Annexe B.8. : Régression linéaire entre débit moyen et inverse du temps intervéhiculaire moyen

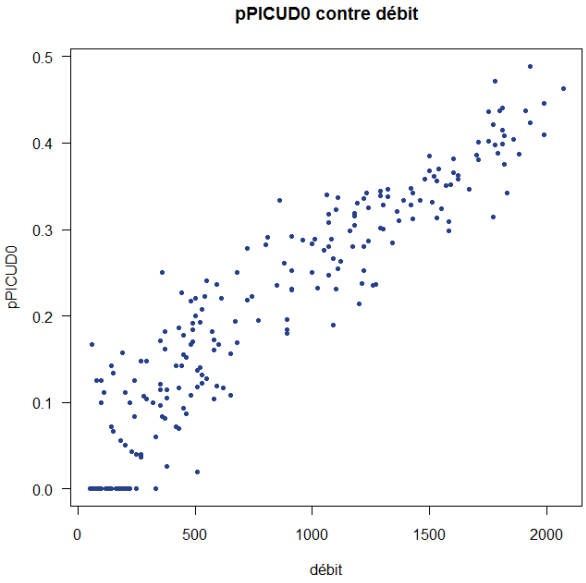
```
> shapiro.test(residuals(modele1))

        Shapiro-Wilk normality test

data:  residuals(modele1)
W = 0.9611, p-value = 4.264e-06
```

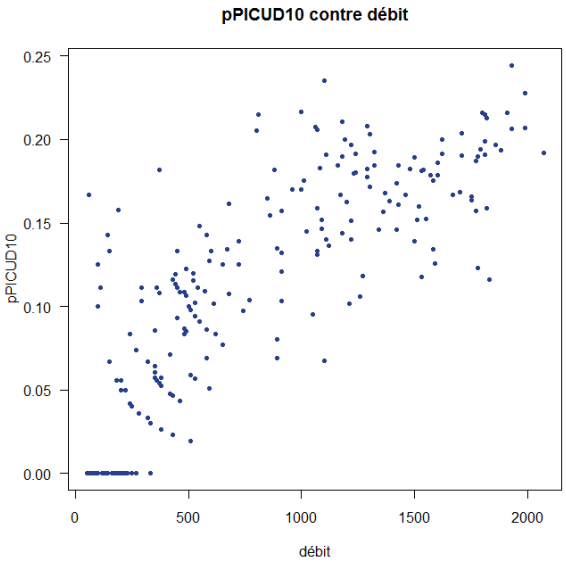


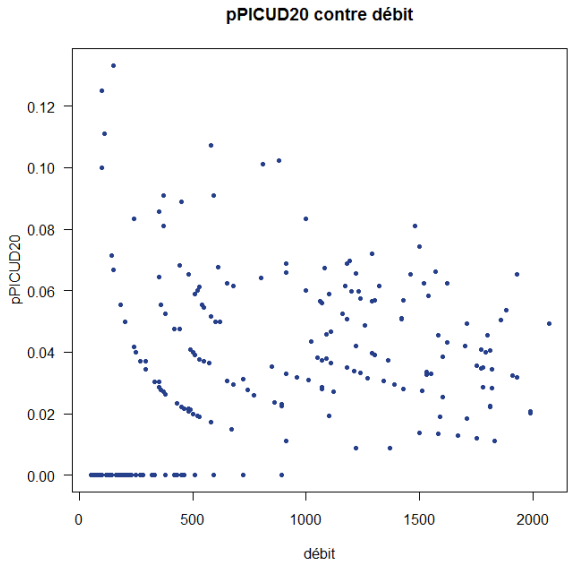

```
Call:
lm(formula = InvTIVmoy ~ VraiDebit)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0154035 -0.0024511 -0.0002484  0.0021639  0.0175872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0005117  0.0004735   1.081    0.281
VraiDebit   0.9976332  0.0017391 573.641   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.004379 on 238 degrees of freedom
Multiple R-squared: 0.9993,     Adjusted R-squared: 0.9993
F-statistic: 3.291e+05 on 1 and 238 DF,  p-value: < 2.2e-16
```

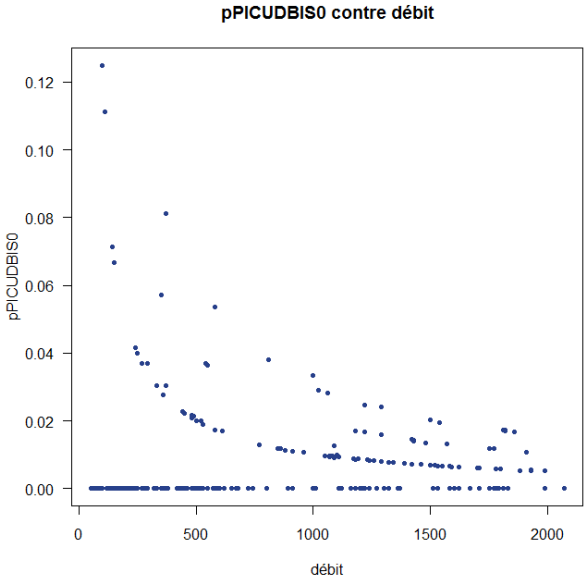
Annexe B.9. : Graphe des corrélations entre variables microscopiques et macroscopiques



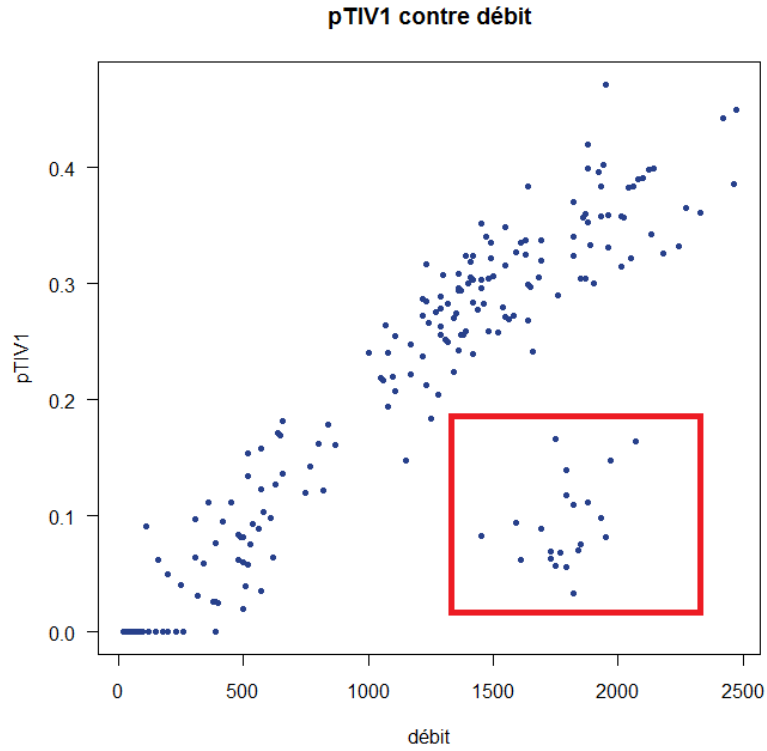

pPICUD0 contre débit
pPICUD0
débit
0.0
0.1
0.2
0.3
0.4
0.5
0
500
1000
1500
2000

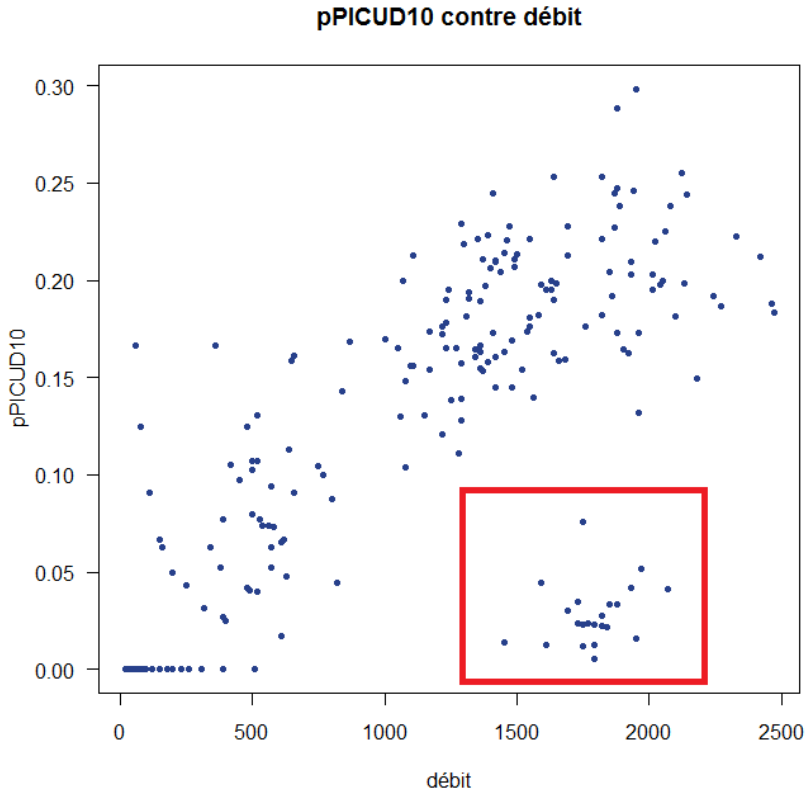

pPICUD10 contre débit
pPICUD10
débit
0.00
0.05
0.10
0.15
0.20
0.25
0
500
1000
1500
2000


pPICUD20 contre débit
pPICUD20
débit
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
500
1000
1500
2000


pPICUDBIS0 contre débit
pPICUDBIS0
débit
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
500
1000
1500
2000

Annexe B.10. : Exemples de graphes de corrélations entre variables microscopiques et macroscopiques en situation de congestion





Les zones encadrées par les rectangles rouges correspondent aux situations de congestion.

Annexe C.1. : Extrait de la table SAS de synthèse des données d'accidents

	Luminosite	ConditionAtmo	Identifiant	DateAccident	Heure	TraceEnPlan	EtatSurface	Amenageme	Aut	PR	SensCirculation	TypeAccident	TypeVehic
21	Plein jour	Normale	688984	20090607	1225	en courbe à droite	normale	bretelle	7	279.2	PR Décroissants	PasSeul	2roues
22	Plein jour	Normale	689001	20090607	1112	Partie rectiligne	normale		7	272.9	PR Décroissants	PasSeul	Pas2roues
23	Plein jour	Normale	688923	20090609	1702		normale		7	275	PR Croissants	PasSeul	Pas2roues
24	Plein jour	Normale	688935	20090609	1740	Partie rectiligne	normale		50	7.75	PR Décroissants	PasSeul	Pas2roues
25	Plein jour	Normale	688931	20090610	840	Partie rectiligne	normale	souterrain	55	1.706	PR Croissants	PasSeul	Pas2roues
26	Plein jour	Normale	688920	20090612	1315	Partie rectiligne	normale	bretelle	55	15	PR Décroissants	Seul	Pas2roues
27	Plein jour	Normale	688922	20090615	815	Partie rectiligne	normale		50	12.7	PR Décroissants	PasSeul	2roues
28	Plein jour	Normale	688944	20090615	800	Partie rectiligne	normale		55	6	PR Décroissants	Seul	2roues
29	Plein jour	Normale	688930	20090617	1255	Partie rectiligne	normale		50	10.35	PR Décroissants	PasSeul	Pas2roues
30	Nuit sans éclairage public	Normale	688925	20090623	2316	Partie rectiligne	normale		7	280.027	PR Décroissants	PasSeul	2roues
31	Plein jour	Normale	688929	20090624	1015	Partie rectiligne	normale		51	9.1	PR Croissants	PasSeul	2roues
32	Plein jour	Normale	688919	20090625	1800	Partie rectiligne	normale		50	10.7	PR Croissants	PasSeul	2roues
33	Plein jour	Normale	688997	20090626	1210	en courbe à droite	normale	souterrain	55	0.54	PR Croissants	PasSeul	Pas2roues
34	Plein jour	Normale	688942	20090630	1710	Partie rectiligne			7	279.33	PR Croissants	PasSeul	Pas2roues
35	Plein jour	Normale	689442	20090702	1830	Partie rectiligne	normale		50	10.97	PR Décroissants	PasSeul	Pas2roues
36	Plein jour	Normale	689682	20090705	635	Partie rectiligne	normale		7	269	PR Décroissants	PasSeul	Pas2roues
37	Plein jour	Normale	689432	20090706	1800	Partie rectiligne	normale		7	265.1	PR Décroissants	PasSeul	Pas2roues
38	Nuit sans éclairage public	Normale	689425	20090707	500	Partie rectiligne	normale		51	11.6	PR Décroissants	Seul	Pas2roues
39	Plein jour	Normale	689429	20090708	1645	Partie rectiligne	normale		7	279.5	PR Croissants	Seul	2roues
40	Plein jour	Normale	689437	20090709	720	Partie rectiligne	normale		51	2.6	PR Décroissants	Seul	Pas2roues
41	Plein jour	Normale	689439	20090710	1845	Partie rectiligne	normale		7	278	PR Décroissants	PasSeul	Pas2roues
42	Nuit sans éclairage public	Normale	689686	20090712	157	en courbe à gauche	normale		7	277	PR Croissants	PasSeul	Pas2roues
43	Plein jour	Normale	689426	20090713	1534	en courbe à gauche	normale		7	269.631	PR Décroissants	PasSeul	2roues
44	Plein jour	Normale	689684	20090714	1445	Partie rectiligne	normale		50	6.6	PR Décroissants	PasSeul	Pas2roues
45	Nuit sans éclairage public	Normale	689424	20090719	515	Partie rectiligne	normale		50	10.7	PR Croissants	Seul	Pas2roues
46	Nuit sans éclairage public	Normale	689440	20090721	2355	Partie rectiligne	normale		7	277	PR Décroissants	Seul	Pas2roues
47	Plein jour	Normale	689441	20090722	1100	Partie rectiligne	normale		7	271.5	PR Décroissants	PasSeul	2roues
48	Nuit sans éclairage public	Normale	689427	20090723	2303	Partie rectiligne	normale		51	3	PR Décroissants	PasSeul	Pas2roues
49	Plein jour	Normale	689428	20090723	1540	Partie rectiligne	normale		55	1	PR Décroissants	PasSeul	Pas2roues
50	Plein jour	Normale	689438	20090723	1810	Partie rectiligne	normale		55	3.5	PR Décroissants	PasSeul	2roues

Cette table SAS de synthèse des accidents est constituée de 289 lignes, chaque ligne correspondant à un accident. Les variables qui apparaissent et leurs modalités sont les suivantes :

- « Luminosite » : Plein jour, Nuit sans éclairage public, Nuit avec EP allumée, Crépuscule ou aube.
- « ConditionAtmospherique » : Normale, Pluie légère, Pluie forte, Temps couvert, Neige-grêle, Autre.
- « Identifiant » : c'est l'identifiant de l'accident dans la base de données.
- « DateAccident » et « Heure » : c'est la date et l'heure à laquelle s'est produit l'accident.
- « TraceEnPlan » : Partie rectiligne, en courbe à gauche, en courbe à droite.
- « EtatSurface » : normale, mouillée, autre.
- « Amenagement » : souterrain, bretelle, zone de péage, pont, ou valeur manquante. Une valeur manquante correspond à l'absence d'aménagements particuliers.
- « Autoroute » : 7, 50, 51, 55. C'est le symbole de l'autoroute sur laquelle l'accident s'est produit.
- « PR » :c'est le point routier qui repère l'accident sur l'autoroute.
- « SensCirculation » : PR Croissants, PR Décroissants. C'est le sens de circulation du (ou des) véhicule(s) impliqué(s) dans l'accident.
- « TypeAccident » : Seul, PasSeul. Cette variable distingue les accidents isolés des accidents non-isolés.
- « TypeVehicule » : 2roues, Pas2roues. Cette variable distingue les accidents impliquant un véhicule type « 2roues » des autres accidents.

Annexe C.2. : Code SAS de la réalisation de la table de synthèse des données d'accidents

```
*importation des données d'accidents;

data accidents1;
     infile 'C:\Documents and Settings\Yannick\Mes documents\Echantillon\accidents\Baac13-Autoroutes_2010-2009_N1.csv'
     firstobs=2 dlm=';' missover;
     length TypeJour$10 SemaineOuWE$10 TrancheHoraire$15 Luminosite$40 Departement$25 Commune2$20 EnOuHorsAgglomeration$25
            Intersection$25 ConditionAtmospherique$12 Lieu$12;
     input Identifiant DateAccidentNiv1 NumPV CodePostal CodeUnite Organisme DateAccidentNiv1 TypeJour$ SemaineOuWE$ Heure
           TrancheHoraire$ Luminosite$ Departement$ Commune CodeINSEE CodePostal Commune2$ EnOuHorsAgglomeration$
           Intersection$ ConditionAtmospherique$ Lieu$ Latitude Longitude;
run;

data accidents2;
     infile 'C:\Documents and Settings\Yannick\Mes documents\Echantillon\accidents\Baac13-Autoroutes_2010-2009_N2.csv'
     firstobs=2 dlm=';' missover;
     length CategorieRoute$12 BisOuTer$10 LettreIndice$10 RegimeCirculation$25 MarquageChaussee$10 VoiesSpeciales$10
            ProfilEnLongDeLaRoute$12 TraceEnPlan$20 EtatRoute$15 EtatSurface$15 AmenagementInfrastructure$20
            Signalisation$15 Environnement$15 SituationAccident$25;
     input Identifiant CodeRoute CategorieRoute$ Autoroute BisOuTer$ LettreIndice$ RegimeCirculation$ NbreVoiesCirculation
           MarquageChaussee$ VoiesSpeciales$ ProfilEnLongDeLaRoute$ Borne PointRoutier TraceEnPlan$ EtatRoute$
           EtatSurface$ AmenagementInfrastructure$ Signalisation$ Environnement$ SituationAccident$;
run;

data accidents3;
     infile 'C:\Documents and Settings\Yannick\Mes documents\Echantillon\accidents\Baac13-Autoroutes_2010-2009_N3.csv'
     firstobs=2 dlm=';' missover;
     length SensCirculation$20 CategorieVehicule$20 CatVeh40$30 AncienneteVehicule$20 ValiditeControleTechnique$25
            PuissanceVehicule$20 AppartenantA$20 VehiculeSpecial$25 Assurance$15 ObstacleFixeHeurte$30
            ObstacleMobileHeurte$30 FacteurLieAuVehicule$20 ManoeuvrePrincipaleAvantAccident$30 PointChocInitial$20 CNIT$15;
     input Identifiant Ident3 DateAccidentNiv3 SensCirculation$ CategorieVehicule$ CatVeh40$ Immatriculation
           AnneeMiseEnCirculation AncienneteVehicule$ AnneeControleTechnique ValiditeControleTechnique$ PuissanceVehicule$
           AppartenantA$ VehiculeSpecial$ Assurance$ ObstacleFixeHeurte$ ObstacleMobileHeurte$ FacteurLieAuVehicule$
           ManoeuvrePrincipaleAvantAccident$ PointChocInitial$ NombreOccupantTC CNIT$;
run;

data accidents4;
     infile 'C:\Documents and Settings\Yannick\Mes documents\Echantillon\accidents\Baac13-Autoroutes_2010-2009_N4.csv'
     firstobs=2 dlm=';' missover;
     length PresumeResponsable$5 CategorieUsager$15 Gravite$15 DetectionEtVerifAlcoolemie$20 Sexe$15 ClasseAge$20
            CategorieSociopro$30 FacteurLieAlUsager$30 PermisDeConduire$20 TrajetConducteur$30 Infrac1$20 Infrac2$20
            EquipementSecurite$20 UtilisationEquipementSecurite$10 ManoeuvreDuPieton$30 ActionDuPieton$20 NombreDePieton$20;
     input Identifiant PresumeResponsable$ CategorieUsager$ Gravite$ DetectionEtVerifAlcoolemie$
           TauxAlcoolPourAlcoolemieIllegale Sexe$ DateNaissance Age ClasseAge$ CategorieSociopro$ FacteurLieAlUsager$
           PermisDeConduire$ AnneeDePermis TrajetConducteur$ Infrac1$ Infrac2$ EquipementSecurite$
           UtilisationEquipementSecurite$ ManoeuvreDuPieton$ ActionDuPieton$ NombreDePieton$;
run;
```

1ère partie du code SAS

```
*calcul de la vrai valeur du point routier (PR);

data accidents2;
     set accidents2;
     PR=Borne+PointRoutier/1000;
run;



*fusion des deux premieres tables, à travers la variable "Identifiant"
(ces deux tables ayant le même nombre de lignes);

proc sort data=accidents1;
     by Identifiant;
run;

proc sort data=accidents2;
     by Identifiant;
run;

data accidents1ETaccidents2;
     merge accidents1 accidents2;
run;



*création de la variable "SensCirculation";

data SensCirculation;
     set accidents3;
     by Identifiant;
     if first.Identifiant;
     keep Identifiant SensCirculation;
run;

*création de la variable "TypeAccident";

proc sort data=accidents3;
     by Identifiant;
run;

 data TypeAccident;
     length TypeAccident$8;
     set accidents3;
     by Identifiant;
     if (first.Identifiant= 1 and last.Identifiant=1) then TypeAccident='Seul';
     else TypeAccident='PasSeul';
     if first.Identifiant;
     keep Identifiant TypeAccident;
run;



*création de la variable "TypeVehicule";

proc sort data=accidents3;
     by Identifiant CategorieVehicule;
run;

data TypeVehicule;
     length TypeVehicule$10;
     set accidents3;
     by Identifiant;
     if ((first.Identifiant=1 and CategorieVehicule=NA) or (last.Identifiant=1 and CategorieVehicule=NA))
     then TypeVehicule='2roues';
     else TypeVehicule='Pas2roues';
     if first.Identifiant;
     keep Identifiant TypeVehicule;
run;
```

2ème partie du code SAS

```
*création de la table "accidents3ETaccidents4", qui regroupe les 3 variables que nous avons créees à partir
des informations extraites des tables 3 et 4.
la fusion des tables se fait à travers la variable commune "Identifiant";

data accidents3ETaccidents4;
    set SensCirculation TypeAccident TypeVehicule;
    merge SensCirculation TypeAccident TypeVehicule;
run;


*fusion des tables accidents1ETaccidents2 et accidents3ETaccidents4,
en ne conservant que les accidents qui se sont produits au cours de notre période d'étude,
à savoir du 17 mai 2009 au 31 mai 2010;

proc sort data=accidents1ETaccidents2;
    by Identifiant;
run;

proc sort data=accidents3ETaccidents4;
    by Identifiant;
run;

data DonneesAccidents;
    set accidents1ETaccidents2 accidents3ETaccidents4;
    merge accidents1ETaccidents2 accidents3ETaccidents4;
    if (DateAccidentNiv1<=20090516 or DateAccidentNiv1>=20100601) then delete;
    keep  Identifiant Autoroute PR DateAccidentNiv1 Heure Luminosite ConditionAtmospherique TraceEnPlan
          EtatSurface AmenagementInfrastructure SensCirculation TypeAccident TypeVehicule;
run;


*exportation de la la table sous format csv, pour un traitement ultérieur;

proc sort data=DonneesAccidents;
    by DateAccidentNiv1;
run;

proc export data=DonneesAccidents
            outfile='C:\Documents and Settings\Yannick\Mes documents\Echantillon\DonneesAccidents.csv'
            replace
            dbms=dlm;
            delimiter=';';
run;
```

3ème partie du code SAS

Il s'agit tout d'abord, dans la 1ère partie du code, d'importer les données présentes dans les quatre fichiers d'accidents, et de créer ainsi quatre tables SAS « accidents1 », « accidents2 », « accidents3 » et « accidents4 ».

Ensuite, dans la 2ème partie du code, nous créons dans la table « accidents2 » la variable « PR » correspondant au point routier de l'accident. Puis, nous fusionnons les deux tables « accidents1 » et « accidents2 », qui ont le même nombre de lignes, à travers la variable « Identifiant » commune aux deux tables. Nous obtenons la table « accidents1ETaccidents2 ». Ensuite, nous construisons les variables « SensCirculation », « TypeAccident » et « TypeVehicule », en utilisant des astuces de codage des modalités que nous ne détaillerons pas. Ces variables sont stockées, avec la variable « Identifiant », dans des tables portant le même nom que les variables.

Dans la 3ème partie, nous créons la table « accidents3ETaccidents4 » en fusionnant les trois tables précédentes. Puis, nous fusionnons les tables « accidents1ETaccidents2 » et « accidents3ETaccidents4 », en nous limitant aux accidents correspondant à nos données de trafic, c'est-à-dire à ceux qui se sont produits entre le 17 mai 2009 et le 31 mai 2010. Nous obtenons ainsi la table SAS « DonneesAccidents » qui synthétise l'ensemble des informations nécessaires. Pour finir, nous exportons la table en format csv dans le fichier « DonneesAccidents.csv », car nous en aurons besoin lorsque nous insérerons les donnés d'accidents dans les données de trafic. Un extrait de la table est disponible dans l'annexe C.1.

Annexe C.3. : Extrait du fichier « M1as1000430C20090712.txt »

```
12 15 59 3972  36  92
12 15 59 4119  39  92
12 15 59 4218  40  90
12 15 59 4903  26 107
12 15 59 4975  36  92
12 15 59 5258  35  94
12 15 59 5605  32  87
12 15 59 5812  27 111
12 16  0  442  39 100
12 16  0  801  41  97
12 16  0 1409  39  97
12 16  0 1493  34 100
12 16  0 1811  34  87
12 16  0 2350  42  90
12 16  0 2548  37  74
12 16  0 2723  38  75
12 16  0 3021  34  75
12 16  0 3271  37  89
12 16  0 3412  44  78
12 16  0 3912  36  98
12 16  0 4415  43  85
```

Il s'agit d'un extrait de données individuelles des véhicules. Les variables sont les suivantes :

- 1ère colonne : la date du jour. Cette variable est donc constante, et prend, pour ce fichier, la valeur 12 sur toutes les lignes
- 2ème, 3ème et 4ème colonnes : l'heure, la minute et la seconde de passage du véhicule. Par exemple, les données 15 59 3972 correspondent au temps de passage 15h59min et 39.72 secondes.
- 5ème colonne : la longueur en décimètre du véhicule.
- 6ème colonne : la vitesse en km/h du véhicule.

Annexe C.4. : Extrait du fichier « M1as1000430C20090712CYCLES.txt »

165	0	NA	1010	0.048955803	103.44000	3.569901
166	0	NA	1200	0.056665261	101.49153	3.007000
167	0	NA	1090	0.052500528	101.25688	3.284128
168	0	NA	1060	0.050014103	101.19048	3.404528
169	0	NA	1130	0.053588415	101.81982	3.191327
170	0	NA	1230	0.059344395	100.37705	2.928618
171	0	NA	1080	0.049457454	103.16981	3.330093
172	0	NA	1610	0.079238958	96.76730	2.224286
173	0	NA	1700	0.084996176	95.90419	2.127294
174	0	NA	1600	0.079444341	98.48428	2.232375
175	0	NA	1200	0.056999319	98.47009	3.030084
176	0	NA	1490	0.078978421	92.25342	2.429329
177	0	NA	1330	0.066241802	97.54962	2.693308
178	0	NA	1550	0.075938821	98.32680	2.342922
179	0	NA	1830	0.089512231	97.79444	1.961257
180	0	NA	1460	0.070635321	98.80420	2.475205
181	0	NA	1950	0.101755646	93.44330	1.851897
182	0	NA	1590	0.079883177	96.10256	2.259937
183	0	NA	1800	0.090088186	96.46067	2.008389
184	0	NA	1920	0.099183569	93.52941	1.892684
185	0	NA	2010	0.104863153	90.87179	1.789602

1.76165357	0.0000000000	0.000000000	0.19801980	0.40594059	0.18181818
2.98159323	0.0000000000	0.033333333	0.24166667	0.55000000	0.29310345
2.45940084	0.0193679918	0.009174312	0.14678899	0.49541284	0.23853211
3.58305734	0.0122536800	0.028301887	0.22641509	0.54716981	0.33653846
2.48342961	0.1394721939	0.000000000	0.16814159	0.40707965	0.24770642
3.19146261	0.0000000000	0.016260163	0.23577236	0.51219512	0.29752066
3.23420881	0.0008636040	0.018518519	0.15740741	0.37962963	0.24038462
3.50874420	0.0000000000	0.012422360	0.27950311	0.62732919	0.40127389
3.09025175	0.0040160643	0.029411765	0.27647059	0.69411765	0.36144578
3.26561296	0.0106589604	0.025000000	0.23750000	0.55625000	0.28662420
2.73959474	0.0000000000	0.016806723	0.22689076	0.55462185	0.28695652
3.36850989	0.0000000000	0.013422819	0.24161074	0.63087248	0.33566434
2.82563403	0.0000000000	0.015037594	0.21804511	0.56390977	0.31007752
3.57515229	0.0000000000	0.032467532	0.26623377	0.62337662	0.34210526
3.16011195	0.0139656134	0.021857923	0.27868852	0.67759563	0.38418079
2.05512433	0.0000000000	0.027397260	0.18493151	0.55479452	0.26241135
3.39525523	0.0200473358	0.025641026	0.28717949	0.70769231	0.39378238
3.48475830	0.2322510823	0.037974684	0.26582278	0.61392405	0.31818182
4.69626999	0.0000000000	0.044444444	0.36111111	0.70000000	0.49431818
3.52201087	0.0218464842	0.021052632	0.30526316	0.70000000	0.39673913
4.13966699	0.0000000000	0.024875622	0.31840796	0.72636816	0.42631579

165	0.11111111	0.000000000	0.000000000
166	0.14655172	0.008620690	0.000000000
167	0.08256881	0.036697248	0.009174312
168	0.19230769	0.038461538	0.009615385
169	0.08256881	0.027522936	0.009174312
170	0.14876033	0.024793388	0.000000000
171	0.15384615	0.048076923	0.009615385
172	0.16560510	0.006369427	0.000000000
173	0.14457831	0.012048193	0.006024096
174	0.14012739	0.031847134	0.012738854
175	0.13043478	0.026086957	0.000000000
176	0.15384615	0.013986014	0.000000000
177	0.13953488	0.015503876	0.000000000
178	0.16447368	0.019736842	0.000000000
179	0.12429379	0.016949153	0.005649718
180	0.09219858	0.007092199	0.000000000
181	0.14507772	0.020725389	0.010362694
182	0.14285714	0.032467532	0.012987013
183	0.17613636	0.028409091	0.000000000
184	0.13586957	0.021739130	0.005434783
185	0.19473684	0.031578947	0.000000000

Il s'agit d'un extrait des données agrégées par cycles, du 165ème au 185ème, issues du fichier « M1as1000430C20090712CYCLES.txt ». Les variables sont les suivantes (dans l'ordre de gauche à droite) :

- NumCycles : le numéro du cycle
- IndAcc : la variable indicatrice d'accident, initialisée à 0
- TypeAcc : la variable indicatrice du type d'accident, initialisée à NA
- debit : le débit horaire
- TauxOcc : le taux d'occupation
- vitessemoy : la vitesse moyenne
- TIVmoy : le temps intervéhiculaire moyen
- VRmoy : la vitesse relative moyenne
- PICUDmoy : le PICUD moyen
- PICUDBISmoy : le PICUDBIS moyen
- pTIV0.5 : la proportion de TIV inférieure à 0.5 secondes
- pTIV1 : la proportion de TIV inférieure à 1 seconde
- pTIV2 : la proportion de TIV inférieure à 2 secondes
- pPICUD0 : la proportion de PICUD inférieure à 0
- pPICUD10 : la proportion de PICUD inférieure à -10
- pPICUD20 : la proportion de PICUD inférieure à-20
- pPICUDBIS0 : la proportion de PICUDBIS inférieure à 0

Annexe C.5. : Code R de création des fichiers de données agrégées par cycles

```
CreerFichiersCycles=function(FichierListe)
{
      liste=read.table(file=FichierListe,dec=".",sep="",quote="\"",header=F,stringsAsFactors=F) #liste des fichiers à traiter
      DimListe=dim(liste)[1]

      for (k in 1:DimListe)
      {
            donnees=read.table(file=liste[k,1],dec=".",sep="",quote="\"",header=F) #on traite un à un les fichiers de la liste
            names(donnees)=c("jour","heure","minute","seconde","longueur","vitesse")
            n=dim(donnees)[1]

            #la procédure ne fonctionne pas avec un fichier pathologique qui ne contient qu'une seule ligne.

            if (n>1)
            {
                  #on filtre les données individuelles aberrantes, dues à des pannes ou des erreurs de mesure du capteur,
                  #mais on conserve les temps de passage.

                  ListeIndices=which(donnees$longueur<5 | donnees$longueur>250 | donnees$vitesse==0 | donnees$vitesse>300)
                  p=length(ListeIndices)
                  donnees=as.matrix(donnees)
                  donnees[ListeIndices,"longueur"]=rep(NA,p)
                  donnees[ListeIndices,"vitesse"]=rep(NA,p)
                  donnees=as.data.frame(donnees)

                  gamma=6.25 #constante de décélerration
                  Tr=1 #temps de réaction
                  l=1 #longueur de la boucle magnétique

                  #on associe un temps de passage à chacun des véhicules.
                  temps=1:n
                  temps = donnees$heure*3600 + donnees$minute*60 + donnees$seconde/100

                  #création des variables microscopiques individuelles :

                  TIV=1:n
                  TIV[1]=NA
                  VR=1:n
                  VR[1]=NA
                  PICUD=1:n
                  PICUD[1]=NA
                  PICUDBIS=1:n
                  PICUDBIS[1]=NA

      for (i in 2:n)
      {
            TIV[i]= temps[i]-temps[i-1]
            VR[i]=1/3.6*(donnees$vitesse[i]-donnees$vitesse[i-1])

            PICUD[i]=((1/3.6)^2*donnees$vitesse[i-1]^2-(1/3.6)^2*donnees$vitesse[i]^2)/(2*gamma)+(temps[i]-temps[i-1])*1/3.6*donnees$vitesse[i-1]
            -Tr*1/3.6*donnees$vitesse[i]-donnees$longueur[i-1]/10

            PICUDBIS[i]=((1/3.6)^2*donnees$vitesse[i-1]^2-(1/3.6)^2*donnees$vitesse[i]^2)/(2*gamma)+
            (temps[i]-temps[i-1])*1/3.6*donnees$vitesse[i-1]-donnees$longueur[i-1]/10
      }

      donnees1=as.data.frame(matrix(c(donnees$longueur,donnees$vitesse,TIV,VR,PICUD,PICUDBIS),ncol=6))
      names(donnees1)=c("longueur","vitesse","TIV","VR","PICUD","PICUDBIS")
      n=dim(donnees1)[1]

      #on filtre les mesures individuelles correspondantes à des remorques de camion:

      ListeIndices=which(donnees1$TIV<0.8 & donnees1$longueur<10)
      p=length(ListeIndices)
      donnees1=as.matrix(donnees1)
      donnees1[ListeIndices,"longueur"]=rep(NA,p)
      donnees1[ListeIndices,"vitesse"]=rep(NA,p)
      donnees1[ListeIndices,"TIV"]=rep(NA,p)
      donnees1[ListeIndices,"VR"]=rep(NA,p)
      donnees1[ListeIndices,"PICUD"]=rep(NA,p)
      donnees1[ListeIndices,"PICUDBIS"]=rep(NA,p)
      donnees1=as.data.frame(donnees1)


      NbreCycles=240
      NumCycles=1:NbreCycles

      IndAcc=rep(0,NbreCycles) #0 ou 1 que l'on complètera par la suite. Initalisé à "0".
      TypeAcc=rep(NA,NbreCycles) #"Seul" ou "PasSeul", que l'on complètera par la suite. Initialité à "NA".

      IndCycles=1:n
      IndCycles=temps%/%360+1 #on fait correspondre chaque véhicule à un cycle grace à la cette variable.
```

```
#on calcule les différentes variables "NbreVeh", qui serviront à calculer le débit
#ainsi que les moyennes pour créer les indicateurs microscopiques agrégées de 1er type.

NbreVeh=1:NbreCycles
NbreVehMesTIV=1:NbreCycles
NbreVehMesPICUD=1:NbreCycles
NbreVehMesPICUDBIS=1:NbreCycles
NbreVehMesvitesse=1:NbreCycles
NbreVehMesVR=1:NbreCycles

for (j in 1:NbreCycles)
{
      NbreVeh[j]=sum(IndCycles==j)
      NbreVehMesTIV[j]=NbreVeh[j]-sum(IndCycles==j & is.na(donnees1$TIV))
      NbreVehMesPICUD[j]=NbreVeh[j]-sum(IndCycles==j & is.na(donnees1$PICUD))
      NbreVehMesPICUDBIS[j]=NbreVeh[j]-sum(IndCycles==j & is.na(donnees1$PICUDBIS))
      NbreVehMesvitesse[j]=NbreVeh[j]-sum(IndCycles==j & is.na(donnees1$vitesse))
      NbreVehMesVR[j]=NbreVeh[j]-sum(IndCycles==j & is.na(donnees1$VR))

}


#création des variables microscopiques agrégées par cycles et des variables macroscopiques :

TIVmoy=1:NbreCycles
VRmoy=1:NbreCycles
PICUDmoy=1:NbreCycles
PICUDBISmoy=1:NbreCycles
pTIV0.5=1:NbreCycles
pTIV1=1:NbreCycles
pTIV2=1:NbreCycles
pPICUD0=1:NbreCycles
pPICUD10=1:NbreCycles
pPICUD20=1:NbreCycles
pPICUDBIS0=1:NbreCycles
vitessemoy=1:NbreCycles

TauxOcc=1:NbreCycles
debit=NbreVeh*10 #pour le calcul du débit, on prend en compte tout les véhicules, y compris ceux qui ont été filtrés.
#mais, pour les moyennes, on ne prend pas en compte les véhicules filtrés. On fait la moyenne uniquement avec les véhicules
#qui ont des mesures valides.


for (j in 1:NbreCycles)
{

if (NbreVehMesTIV[j]>=1 & NbreVehMesPICUD[j]>=1 & NbreVehMesPICUDBIS[j]>=1 & NbreVehMesvitesse[j]>=1 & NbreVehMesVR[j]>=1 & debit[j]<3800)
{

      TauxOcc[j]=1/360*sum((1/10*donnees1$longueur[IndCycles==j]+1)/(1/3.6*donnees1$vitesse[IndCycles==j]),na.rm=T)
      vitessemoy[j]=sum(donnees1$vitesse[IndCycles==j],na.rm=T)/NbreVehMesvitesse[j]
      TIVmoy[j]=sum(donnees1$TIV[IndCycles==j],na.rm=T)/NbreVehMesTIV[j]
      VRmoy[j]=sum(donnees1$VR[IndCycles==j & donnees1$VR>0],na.rm=T)/NbreVehMesVR[j]
      PICUDmoy[j]=abs(sum(donnees1$PICUD[IndCycles==j & donnees1$PICUD<0],na.rm=T)/NbreVehMesPICUD[j])
      PICUDBISmoy[j]=abs(sum(donnees1$PICUDBIS[IndCycles==j & donnees1$PICUDBIS<0],na.rm=T)/NbreVehMesPICUDBIS[j])
      pTIV0.5[j]=sum(donnees1$TIV<0.5 & IndCycles==j,na.rm=T)/NbreVehMesTIV[j]
      pTIV1[j]=sum(donnees1$TIV<1 & IndCycles==j,na.rm=T)/NbreVehMesTIV[j]
      pTIV2[j]=sum(donnees1$TIV<2 & IndCycles==j,na.rm=T)/NbreVehMesTIV[j]
      pPICUD0[j]=sum(donnees1$PICUD<0 & IndCycles==j,na.rm=T)/NbreVehMesPICUD[j]
      pPICUD10[j]=sum(donnees1$PICUD<(-10) & IndCycles==j,na.rm=T)/NbreVehMesPICUD[j]
      pPICUD20[j]=sum(donnees1$PICUD<(-20) & IndCycles==j,na.rm=T)/NbreVehMesPICUD[j]
      pPICUDBIS0[j]=sum(donnees1$PICUDBIS<0 & IndCycles==j,na.rm=T)/NbreVehMesPICUDBIS[j]
}

else
{
      debit[j]=NA
      TauxOcc[j]=NA
      vitessemoy[j]=NA
      TIVmoy[j]=NA
      VRmoy[j]=NA
      PICUDmoy[j]=NA
      PICUDBISmoy[j]=NA
      pTIV0.5[j]=NA
      pTIV1[j]=NA
      pTIV2[j]=NA
      pPICUD0[j]=NA
      pPICUD10[j]=NA
      pPICUD20[j]=NA
      pPICUDBIS0[j]=NA
}

}
```

La fonction « CreerFichiersCycles » prend en entrée la liste des 14000 fichiers environ, du type « M1as1000430C20090712.txt », issus du tri préliminaire, et va créer les 14000 fichiers de données agrégées par cycles, du type « M1as1000430C20090712CYCLES.txt ».

Il a fallu, tout d'abord, procéder au traitement des données aberrantes, issues d'erreurs ou de pannes de capteurs. Nous avons ainsi réalisé trois niveaux de filtrage :

- Au niveau des données microscopiques individuelles brutes : certaines données, issues des fichiers de données de trafic sources, sont aberrantes. On y trouve des vitesses nulles ou trop élevées ou des longueurs nulles, trop faibles ou trop élevées. Nous avons donc filtré ces données de manière à conserver les données des temps de passage qui, elles, sont exactes, afin de ne pas faire baisser le débit de manière artificielle. Les valeurs des variables correspondant à des données aberrantes sont, quant à elles, affectés de la valeur « NA ». Elles ne seront ainsi pas prises en compte dans la suite du traitement.

- Au niveau des variables microscopiques individuelles créées : nous avons remarqué que les capteurs traitent souvent les cabines de camions et les remorques de camions comme deux véhicules distincts, à cause de l'espace qui existe entre les deux. Pour remédier à ce problème, nous avons filtré les données individuelles correspondant à la fois à un TIV inférieur à 0.8 secondes et à une longueur inférieure à 1m, pour repérer ce type de situation. Les véhicules ainsi filtrés ne seront pas pris en compte dans la suite du traitement.

- Au niveau des variables macroscopiques : nous avons constaté que certains cycles possédaient des valeurs de débit beaucoup trop élevées (supérieures à 3800). Or, il est impossible en pratique que plus de 380 véhicules passent devant un capteur en six minutes. Il s'agit en fait d'erreurs de données contenues dans les fichiers sources. Nous avons donc filtré ces cycles, en assignant la valeur « NA » à l'ensemble des variables.

Le principe de base de la création des variables macroscopiques et microscopiques individuelles agrégées par cycles est d'associer chaque véhicule individuel à 1 des 240 cycles de 6 minutes de la journée, à travers la variable « IndCycles ». Elle se sert des instants de passage devant le capteur pour associer les bons cycles aux bons véhicules.

Enfin, nous devons apporter quelques précisions concernant les variables « NbreVeh », « NbreVehMesTIV », « NbreVehMesPICUD » etc. Pour chaque cycle, le débit réel est calculé à partir du nombre total de véhicules appartenant au cycle, y compris ceux dont les données ont été filtrées. En effet, les données des temps de passage étant correctes, nous devons les prendre en compte pour ne pas faire baisser le débit de manière artificielle. Cependant, lors du calcul des moyennes pour les variables microscopiques agrégées, nous devons uniquement comptabiliser les véhicules dont les données des variables impliquées dans le calcul sont correctes. C'est pour cela que nous avons défini différentes notions de nombre de véhicules comme « NbreVehMesTIV » ou « NbreVehMesPICUD ».

Annexe C.6. : Extrait du ficher «M7is1264687C20090706CYCLES.txt » enrichi des informations d'accidents

175	0	<NA>	940	0.071506965	74.16129	3.791277	1.4613527	1.59270666	0.109541063	0.021276596	0.09574468	0.32978723	0.19565217
176	0	<NA>	1030	0.094325243	74.59804	3.582255	1.2690632	1.71409162	0.088525781	0.000000000	0.07843137	0.27450980	0.21568627
177	0	<NA>	980	0.081321159	82.32990	3.697320	1.5635739	3.21918035	0.181809215	0.000000000	0.13402062	0.39175258	0.26804124
178	0	<NA>	950	0.075354440	72.05319	3.695684	1.4844683	1.22403093	0.039894464	0.010526316	0.08421053	0.37894737	0.11827957
179	0	<NA>	700	0.047866459	93.92754	5.252429	1.7483660	1.62252179	0.000000000	0.000000000	0.02857143	0.17142857	0.10294118
180	1	PasSeul	650	0.043338007	95.07812	5.494688	1.8055556	1.41014178	0.148644869	0.015625000	0.03125000	0.17187500	0.09375000
181	0	<NA>	630	0.045363376	91.03175	5.851905	1.2389771	0.70797570	0.000000000	0.000000000	0.01587302	0.17460317	0.09523810
182	0	<NA>	860	0.049309578	92.43373	4.129070	1.4479167	1.47425154	0.004591049	0.000000000	0.03488372	0.24418605	0.12500000
183	0	<NA>	650	0.037908446	96.35484	5.406250	1.7129630	1.06160494	0.232901235	0.000000000	0.01562500	0.12500000	0.03333333
184	0	<NA>	770	0.043600549	96.34211	4.851688	1.7074074	1.10011934	0.056666667	0.000000000	0.03896104	0.15584416	0.09333333
185	0	<NA>	620	0.036598652	93.90164	5.872258	1.3796296	1.69397634	0.231255144	0.000000000	0.03225806	0.17741935	0.06666667

175	0.06521739	0.010869565	0.021739130
176	0.06862745	0.019607843	0.019607843
177	0.12371134	0.061855670	0.041237113
178	0.05376344	0.021505376	0.021505376
179	0.07352941	0.029411765	0.000000000
180	0.04687500	0.031250000	0.031250000
181	0.03174603	0.000000000	0.000000000
182	0.06250000	0.012500000	0.012500000
183	0.03333333	0.033333333	0.033333333
184	0.02666667	0.013333333	0.013333333
185	0.05000000	0.050000000	0.033333333

Le 180ème cycle de cette journée correspond à une situation d'avant-accident. Le programme R d'insertion des accidents dans les données de trafic, présenté dans l'annexe C.8., a donc modifié les deux premières colonnes de ce fichier, au niveau de la 180ème ligne : la variable indicatrice d'accident « IndAcc » a pris la valeur 1, et la variable indicatrice du type d'accident « TypeAcc » la valeur « PasSeul ».

Annexe C.7. : Extrait du fichier « acc_baac_marius.csv »

	A	B	C	D	E	F	G	H	I	J
1	numero	heure	minute_multiple_6	fichier	Luminosite	ConditionAtmospherique	Identifiant	DateAcciden	Heure	TraceEnPlan
2	13	12	12	2009\05\M4os1002482C20090529	Plein_jour	Normale	680789	20090529	1215	en_courbe_à_gauche
3	13	12	12	2009\05\M4os1002482D20090529	Plein_jour	Normale	680789	20090529	1215	en_courbe_à_gauche
4										
5										
6	EtatSurface	Amé	Autoroute	PR_metre	SensCirculation	TypeAccident	TypeVehicul	normf	b	Amé
7	normale	0	50	5600	PR_Décroissants	Seul	Pas2roues	norm	5600	0
8	normale	0	50	5600	PR_Décroissants	Seul	Pas2roues	norm	5600	0
9										
10	PR									
11	5,6									
12	5,6									

19	numero	heure	minute_multiple_6	fichier	Luminosite	ConditionAtmospherique	Identifiant	DateAcciden	Heure	TraceEnPlan
20	14	1	54	2009\05\M8Hs2275695C20090530	Nuit_sans_éclaira	Normale	680777	20090530	158	en_courbe_à_gauche
21	14	1	54	2009\05\M8Hs2275695D20090530	Nuit_sans_éclaira	Normale	680777	20090530	158	en_courbe_à_gauche
22	14	1	54	2009\05\M8Hs2275695G20090530	Nuit_sans_éclaira	Normale	680777	20090530	158	en_courbe_à_gauche
23										
24	EtatSurface	Amé	Autoroute	PR_metre	SensCirculation	TypeAccident	TypeVehicul	normf	b	Amé
25	xx	0	7	275235	PR_Croissants	PasSeul	Pas2roues	norm	275235	0
26	xx	0	7	275235	PR_Croissants	PasSeul	Pas2roues	norm	275235	0
27	xx	0	7	275235	PR_Croissants	PasSeul	Pas2roues	norm	275235	0
28										
29	PR									
30	275,235									
31	275,235									
32	275,235									

Ceci est un extrait qui présente le traitement réalisé par le programme FORTRAN pour deux accidents : les numéros 13 et 14. En plus des informations générales associées aux accidents, déjà disponibles dans la table de synthèse des accidents, la colonne « fichier » fournit pour chaque accident les fichiers de données de trafic correspondants, qui seront modifiés par le programme R présenté en annexe C.8.

Par exemple, l'accident numéro 13 est associé à deux fichiers de trafic, correspondant à la voie centrale (C) et à la voie de droite (D). La portion d'autoroute sur laquelle a eu lieu cet accident n'est, en effet, constitué que de deux voies de circulation. Quant à l'accident numéro 14, ce sont trois fichiers de trafic qui lui sont associés ; la portion d'autoroute correspondante comprenait en effet trois voies de circulation.

Annexe C.8. : Code R du programme d'insertion des données relatives aux accidents dans les données de trafic

```
DonneesAccidents=read.table(file="acc_baac_marius.csv",dec=".",sep=";",header=T)

#on filtre les accidents dont on ne connait pas les informations de PR ou de sens de circulation.
DonneesAccidents=DonneesAccidents[DonneesAccidents$PR!=0 & DonneesAccidents$SensCirculation!=0,]

K=dim(DonneesAccidents)[1]
NouveauNom=1:K

#on insère un accident à la fois dans les fichiers de données de trafic :
for (k in 1:K)
{
      ChiffresSeparesDate=unlist(strsplit(as.character(DonneesAccidents$DateAccidentNiv1[k]), ''))
      LettresSeparees=unlist(strsplit(as.character(DonneesAccidents$fichier[k]), ''))

      #on ne traite que les accidents correspondants au mois de juillet 20009, ainsi que ceux
      #dont on  a pas trouvé de fichiers de données de trafic correspondants :

      if (ChiffresSeparesDate[5]=="0" & ChiffresSeparesDate[6]=="7" & LettresSeparees[9]!="P")
      {
            IndiceDuDebut=min(match("m",LettresSeparees),match("M",LettresSeparees),na.rm=T)
            NomAncienFichier=substr(DonneesAccidents$fichier[k], IndiceDuDebut,length(LettresSeparees))
            NomFichierTrafficCorresp=paste(NomAncienFichier,"CYCLES.txt",sep="")

            #le fichier de lien "acc_baac_marius.csv" associe à certains accidents des fichiers de trafic inexistants,
            #on ne traitera donc pas ces accidents.

            if (NomFichierTrafficCorresp!="M7rs1269326D20090713CYCLES.txt" & NomFichierTrafficCorresp!="M5es1001892C20090723CYCLES.txt"
                & NomFichierTrafficCorresp!="M5es1001892D20090723CYCLES.txt" & NomFichierTrafficCorresp!="M5ds1001439G20090723CYCLES.txt")

            {
                  #pour l'accident traité, on repère le cycle correspondant à la situation d'avant-accident :
                  CycleCorresp=DonneesAccidents$heure[k]*10+DonneesAccidents$minute_multiple_6[k]/6

                  #puis, on ouvre le bon fichier de trafic :
                  FichierTrafficCorresp=read.table(file=NomFichierTrafficCorresp,dec=".",sep=";",header=T)

                  #puis, on implémente à la variable "IndAcc" la valeur 1 au niveau du bon cycle du bon fichier
                  #et l'information correspondante au type d'accident (Seul ou PasSeul) dans la variable TypeAcc :
                  FichierTrafficCorresp$IndAcc[CycleCorresp]=1
                  FichierTrafficCorresp$TypeAcc[CycleCorresp]=as.character(DonneesAccidents$TypeAccident[k])

                  NouveauNom[k]=NomFichierTrafficCorresp

                  #on écrase l'ancien fichier de trafic par le nouveau fichier de trafic enrichi des informations d'accidents
                  write.table(FichierTrafficCorresp,file=NomFichierTrafficCorresp, append=F,dec=".",sep=";",row.names=F,col.names=T,quote=T,qmethod="d")

            }
      }
}


#pour garder en mémoire les fichiers de trafic modifiés :
write.table(NouveauNom,file="ListeFichiersAccidents.txt",append=F)
```

Plusieurs problèmes apparaissent lors de l'insertion des données d'accidents dans les données de trafic :

- Pour certains accidents, la donnée du point routier (PR) ou du sens de circulation, qui permet de localiser l'endroit où a eu lieu l'accident, est manquante (le fichier « acc_baac_marius.csv » indique alors la valeur 0 au niveau des variables « PR » et « SensCirculation »). L'accident, non localisable dans l'espace, ne peut donc pas être inséré dans les données de trafic.
- Le fichier « acc_baac_marius.csv » associe certains accidents à des fichiers de données de trafic inexistants. Ces accidents ne seront donc pas non plus inclus dans l'étude.

Le fichier « ListeFichierAccidents.txt », créé au cours de la procédure, permet de garder en mémoire les fichiers de trafic modifiés.

Annexe C.9. : Code R des programmes de listings et d'agrégation bout-à-bout des fichiers

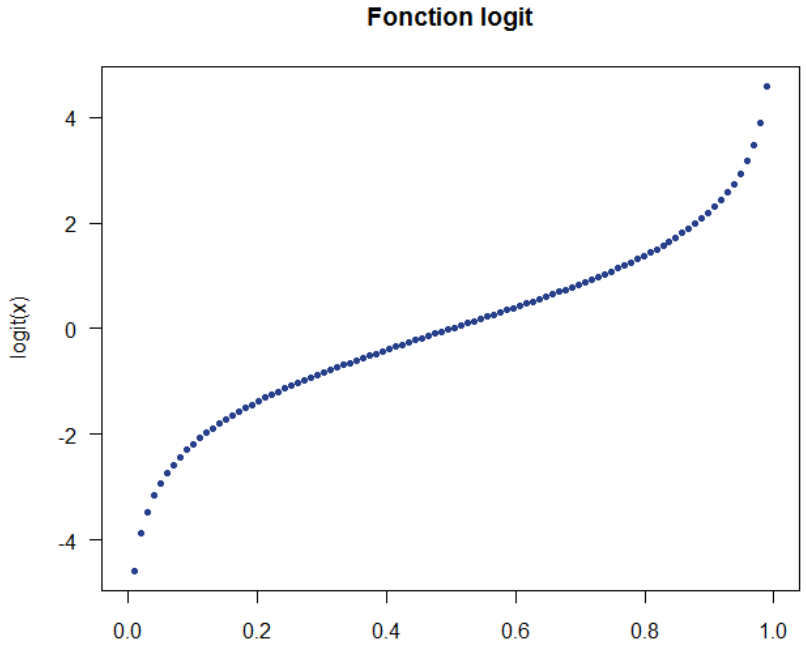
```
#Cette fonction va créer les fichiers "listeG.txt", "listeC.txt" et "listeG.txt"
#qui vont lister les fichiers de données, présents dans "liste.txt",
#en les distinguant selon les trois différentes voies de circulation.

CreerFichiersListes=function(FichierListe)
{
      liste=read.table(file=FichierListe,dec=".",sep="",quote="\"",header=F,stringsAsFactors=F)
      names(liste)=c("fichier")
      DimListe=dim(liste)[1]
      IndiceVoie=1:DimListe

      for (k in 1:DimListe)
      {
            LettresSeparees=unlist(strsplit(liste[k,1], ''))
            IndiceVoie[k]=LettresSeparees[12]
      }

      listeG=liste[IndiceVoie=="G",]
      listeC=liste[IndiceVoie=="C",]
      listeD=liste[IndiceVoie=="D",]

      write.table(listeG,file="listeG.txt", append=F,dec=".",sep=";",row.names=F,col.names=F,quote=T,qmethod="d")
      write.table(listeC,file="listeC.txt", append=F,dec=".",sep=";",row.names=F,col.names=F,quote=T,qmethod="d")
      write.table(listeD,file="listeD.txt", append=F,dec=".",sep=";",row.names=F,col.names=F,quote=T,qmethod="d")
}

CreerFichiersListes("liste.txt")
```

Fonction R de listings des fichiers « CreerFichiersListes »

```
#Cette fonction va agréger bout-à-bout les fichiers de données listées dans "FichierListe"
#et créer le fichier agrégée "NomFichierResultat".

AgregerFichiers=function(FichierListe,NomFichierResultat)
{
      liste=read.table(file=FichierListe,dec=".",sep="",quote="\"",header=F,stringsAsFactors=F)
      DimListe=dim(liste)[1]

      DonneesInitiales=as.matrix(read.table(file=liste[1,1],dec=".",sep=";",header=T))

      for (k in 2:DimListe)
      {
            DonneesARajouter=as.matrix(read.table(file=liste[k,1],dec=".",sep=";",header=T))
            DonneesInitiales=rbind(DonneesInitiales,DonneesARajouter)

      }


      write.table(DonneesInitiales,file=NomFichierResultat,col.names=T,row.names=F,sep=";",dec=".")

}

AgregerFichiers("listeD.txt","DonneesCyclesD.txt")
AgregerFichiers("listeG.txt","DonneesCyclesG.txt")
AgregerFichiers("listeC.txt","DonneesCyclesC.txt")
```

Fonction R d'agrégation bout-à-bout des fichiers « AgregerFichiers »

Annexe D.1. : Graphe de la fonction logit



Annexe D.2. : Développements des calculs de l'estimation des paramètres

Comme les observations sont indépendantes, la vraisemblance du modèle s'écrit de la manière suivante :

$$L(Y \mid X, \beta) = \prod_{i=1}^{n} \mathbb{P}(Y = y_i \mid X = x_i).$$

Or, nous avons vu que : $Y|X = x \sim \mathfrak{B}(1, p(x))$.

D'où, en utilisant la densité d'une loi de Bernoulli, nous pouvons écrire pour tout i :

$$\mathbb{P}(Y = y_i \mid X = x_i) = p_i^{y_i} * (1 - p_i)^{1-y_i} \text{ où } p_i = \mathbb{P}(Y = 1 \mid X = x_i)$$

Et donc :

$$L(Y \mid X, \beta) = \prod_{i=1}^{n} \{p_i^{y_i} * (1 - p_i)^{1-y_i}\}$$

En passant au logarithme, nous obtenons :

$$\ell(Y \mid X, \beta) = \log L(Y \mid X, \beta) = \sum_{i=1}^{n} \{y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)\}$$
$$= \sum_{i=1}^{n} \left\{y_i * \log\left(\frac{p_i}{1 - p_i}\right) + \log(1 - p_i)\right\}$$

Or, d'après le modèle de régression logistique :

$$p_i = \mathbb{P}(Y = 1 \mid X = x_i) = \frac{\exp(\beta_0 + \beta_1 * x_i)}{1 + \exp(\beta_0 + \beta_1 * x_i)}$$

D'où : $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * x_i$ et $\log(1 - p_i) = -\log(1 + \exp(\beta_0 + \beta_1 * x_i))$

Nous obtenons donc :

$$\begin{aligned}\ell(Y \mid X, \beta) &= \sum_{i=1}^{n} \{y_i * (\beta_0 + \beta_1 * x_i) - \log(1 + \exp(\beta_0 + \beta_1 * x_i))\} \\ &= \beta_0 * \sum_{i=1}^{n} y_i + \beta_1 * \sum_{i=1}^{n} (x_i * y_i) - \sum_{i=1}^{n} \log(1 + \exp(\beta_0 + \beta_1 * x_i))\end{aligned}$$

En dérivant cette expression par rapport à β_0 et àβ_1, nous obtenons :

- $\frac{\partial \ell(Y \mid X,\beta)}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0+\beta_1*x_i)}{1+\exp(\beta_0+\beta_1*x_i)} = \sum_{i=1}^{n} (y_i - p_i)$
- $\frac{\partial \ell(Y \mid X,\beta)}{\partial \beta_1} = \sum_{i=1}^{n} (x_i * y_i) - \sum_{i=1}^{n} \frac{x_i*\exp(\beta_0+\beta_1*x_i)}{1+\exp(\beta_0+\beta_1*x_i)} = \sum_{i=1}^{n} (xi * (y_i - p_i))$

Il s'agit donc de trouver les estimateurs $\widehat{\beta_0}$ et $\widehat{\beta_1}$ qui vérifient :

$$\begin{cases} \sum_{i=1}^{n} (y_i - \widehat{p_i}) = 0 \\ \sum_{i=1}^{n} (xi * (y_i - \widehat{p_i})) = 0 \end{cases}$$

où $\widehat{p_i} = \frac{\exp(\widehat{\beta_0}+\widehat{\beta_1}*x_i)}{1+\exp(\widehat{\beta_0}+\widehat{\beta_1}*x_i)}$

C'est un système de deux équations non-linéaires qui se résout, en pratique, à l'aide d'une méthode numérique basée sur un algorithme itératif que nous ne détaillerons pas.

Annexe D.3. : Puissance du test de Wald

Le test de Wald s'appuie sur la statistique : $T = \frac{\widehat{\beta_1} - \beta_1}{\sigma_1}$.

Sous l'hypothèse nulle, La statistique du test de Wald $T = \frac{\widehat{\beta_1}}{\sigma_1}$ suit une loi normale centrée réduite.

Rappelons que σ_1 est supposé connu, mais, qu'en pratique, il est estimée par la grandeur :

$$\sigma_1 = \sqrt{\frac{1}{Vrai1} + \frac{1}{Faux1}}$$

où Vrai1 représente le nombre d'accidents correspondant à une valeur « à risque » de X, et Faux1 représente le nombre d'accidents correspondant à une valeur non « à risque » de X.

La valeur critique au-delà de laquelle le test va rejeter l'hypothèse nulle est donc le quantile $u_{1-\alpha/2}$ de la loi normale centrée réduite. Nous pouvons à présent calculer la puissance $1 - \beta$:

$$1 - \beta = \mathbb{P}\left(T > u_{1-\alpha/2} \middle| H_1 \; est \; vraie\right) = \mathbb{P}\left(\frac{\widehat{\beta_1}}{\sigma_1} > u_{1-\alpha/2}\right)$$

Or, sous H_1, on a simplement :

$$\frac{\widehat{\beta_1} - \beta_1}{\sigma_1} \sim \mathcal{N}(0,1)$$

D'où :

$$1 - \beta = \mathbb{P}\left(\frac{\widehat{\beta_1}}{\sigma_1} > u_{1-\alpha/2}\right) = \mathbb{P}\left(\frac{\widehat{\beta_1} - \beta_1}{\sigma_1} > u_{1-\alpha/2} - \frac{\beta_1}{\sigma_1}\right)$$

Et donc, par symétrie de la $\mathcal{N}(0,1)$:

$$1 - \beta = \mathbb{P}\left(\frac{\widehat{\beta_1} - \beta_1}{\sigma_1} \leq \frac{\beta_1}{\sigma_1} - u_{1-\alpha/2}\right)$$

Or, puisque $\frac{\widehat{\beta_1} - \beta_1}{\sigma_1} \sim \mathcal{N}(0,1)$, si l'on note ϕ la fonction de répartition de la loi normale centrée réduite, on obtient finalement la relation :

$$1 - \beta = \phi\left(\frac{\beta_1}{\sqrt{\frac{1}{Vrai1} + \frac{1}{Faux1}}} - u_{1-\alpha/2}\right)$$

Cette formule permet de trouver le nombre p d'accidents qu'il faudrait inclure pour obtenir une puissance $1 - \beta$ fixée.

Le nombre d'accidents inclus dans l'étude initiale de la puissance est : Vrai1+Faux1. Considérons alors le coefficient k tel que :

$$p = k * (Vrai1 + Faux1)$$

L'interprétation du coefficient k est simple ; si, par exemple, la valeur du coefficient k trouvée est égale à 50, cela signifiera que, pour atteindre la puissance fixée $1 - \beta$, nous devrions inclure 50 fois plus d'accidents que le nombre Vrai1+Faux1 d'accidents initialement inclus.

Calculons à présent la valeur de k. Posons : $Vrai'1 = k * Vrai1$ et $Faux'1 = k * Faux1$. Alors, $p = Vrai'1 + Faux'1$.

Vrai'1 et Faux'1 représentent ainsi les nombres d'accidents à inclure, correspondant à une situation respectivement « à risque » et non « à risque » de la variable X testée, pour obtenir la puissance souhaitée $1 - \beta$.

Nous pouvons écrire :

$$1 - \beta = \phi\left(\frac{\beta_1}{\sqrt{\frac{1}{Vrai'1} + \frac{1}{Faux'1}}} - u_{1-\alpha/2}\right) = \phi\left(\frac{\beta_1}{\sqrt{\frac{1}{k*Vrai1} + \frac{1}{k*Faux1}}} - u_{1-\alpha/2}\right)$$

D'où :

$$1 - \beta = \phi\left(\frac{\beta_1 * \sqrt{k}}{\sqrt{\frac{1}{Vrai1} + \frac{1}{Faux1}}} - u_{1-\alpha/2}\right)$$

On obtient donc :

$$\sqrt{k} = \left(u_{1-\alpha/2} + u_{1-\beta}\right) * \frac{\sqrt{\frac{1}{Vrai1} + \frac{1}{Faux1}}}{\beta_1}$$

D'où :

$$k = \left(u_{1-\alpha/2} + u_{1-\beta}\right)^2 * \left(\frac{1}{Vrai1} + \frac{1}{Faux1}\right) * \frac{1}{{\beta_1}^2}$$

où β_1 correspond à la valeur que l'on souhaite détecter à la puissance $1 - \beta$.

On obtient finalement le nombre d'accidents p cherché, en calculant simplement :

$$p = k * (Vrai1 + Faux1)$$

Annexe D.4. : Odd Ratio, risques relatifs et lien avec le coefficient β_1 de la régression

Considérons le tableau de contingence présenté dans la section 4.3.2.

Notons $p_{i|j} = \mathbb{P}(Y = i \mid X = j)$ $i\epsilon\{0,1\}$, $j\epsilon\{0,1\}$ les probabilités estimées par le modèle d'être dans chacun des 4 cas de figures qui se présentent.

À partir de ces quatre probabilités, nous pouvons construire deux « Odds » différents, en fixant la modalité de X ; il s'agit alors de considérer le rapport des probabilités correspondant aux deux modalités de Y :

$$Odds_1 = \frac{p_{1|1}}{p_{0|1}} \text{ et } Odds_0 = \frac{p_{1|0}}{p_{0|0}}$$

L'Odd Ratio ou rapport de côtes, noté OR, est alors défini par le rapport des deux Odds :

$$OR = \frac{Odds_1}{Odds_0} = \frac{p_{1|1}/p_{0|1}}{p_{1|0}/p_{0|0}}$$

On peut aussi définir le « risque relatif » de présenter la réponse $Y = 1$: $RR = \frac{p_{1|1}}{p_{1|0}}$.

OR peut se réécrire de la manière suivante : $OR = \frac{p_{1|1}/1-p_{1|1}}{p_{1|0}/1-p_{1|0}}$. Ainsi, si $p_{1|1}$ et $p_{1|0}$ prennent des valeurs faibles, ce qui sera le cas dans notre étude puisque la probabilité qu'un accident se produise est très faible, alors l'Odd Ratio peut être approximé par le risque relatif, dont l'interprétation est plus aisée : $OR \approx RR$.

Démontrons à présent la relation :

$$OR = e^{\beta_1}$$

Pour cela, considérons le modèle de régression qui nous intéresse :

$$logit\ p(x) = \beta_0 + \beta_1 * \mathbb{1}_1(x)$$

D'après les formules de la régression logistique, nous avons que :

$$p_{1|0} = \mathbb{P}(Y = 1 \mid X = 0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \text{ et } p_{1|1} = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$$

Et donc, en passant aux complémentaires :

$$p_{0|0} = 1 - p_{1|0} = \frac{1}{1+\exp(\beta_0)} \text{ et } p_{0|1} = 1 - p_{1|1} = \frac{1}{1+\exp(\beta_0+\beta_1)}$$

D'où :

$$OR = \frac{p_{1|1}/1-p_{1|1}}{p_{1|0}/1-p_{1|0}} = \frac{\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \Big/ \frac{1}{1+\exp(\beta_0+\beta_1)}}{\frac{\exp(\beta_0)}{1+\exp(\beta_0)} \Big/ \frac{1}{1+\exp(\beta_0)}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Annexe D.5. : Détails de la construction du coefficient « c »

C'est un coefficient qui correspond à la probabilité qu'une paire quelconque de deux observations, correspondant respectivement a priori à un cas d'accident $(Y = 1)$et à un cas de non-accident $(Y = 0)$, soit correctement prédite par le modèle. Il est défini, de manière rigoureuse, à l'aide de la notion de score, que nous ne développerons pas.

On considère les quatre types de paires que l'on peut former.

La paire est dite :

- « concordante » si $\mathbb{P}(Y = 1|\, IndX = 1) > \mathbb{P}(Y = 1|\, IndX = 0)$
- « discordante » si $\mathbb{P}(Y = 1|\, IndX = 1) < \mathbb{P}(Y = 1|\, IndX = 0)$
- « liée » si $\mathbb{P}(Y = 1|\, IndX = 1) = \mathbb{P}(Y = 1|\, IndX = 0)$

Type de paire	$\boldsymbol{Y = 1}$	$\boldsymbol{Y = 0}$	**Nombre de paires**	**Classification**
1	$IndX = 1$	$IndX = 0$	Vrai1*Vrai0	Concordante
2	$IndX = 0$	$IndX = 1$	Faux1*Faux0	Discordante
3	$IndX = 1$	$IndX = 1$	Vrai1*Faux1	Liée
4	$IndX = 0$	$IndX = 0$	Vrai0*Faux0	Liée
Total			(Vrai1+Faux0)*(Faux1+Vrai0)	

Ainsi, si nous prenons un poids de 1 pour les paires concordantes, un poids de 0.5 pour les paires liées, et un poids de 0 pour les paires discordantes, nous obtenons une estimation du coefficient c, c'est-à-dire de la probabilité que le score d'une observation avec $Y = 1$ soit supérieure au score d'une observation avec $Y = 0$:

$$c = \frac{Vrai1 * Vrai0 + 0.5 * Vrai1 * Faux1 + 0.5 * Vrai0 * Faux0}{(Vrai1 + Faux0) * (Faux1 + Vrai0)}$$

Annexe E.1. : Construction du test de Mann-Whitney

Il s'agit d'un test non-paramétrique qui ne nécessite, contrairement au test de Student, aucune hypothèse sur la distribution des observations. Seule l'indépendance des observations et des échantillons doit être supposée.

Ce test compare les valeurs numériques de deux échantillons indépendants, notées $X_{i1}, \dots, X_{in1}$ et $Y_{j1}, \dots, Y_{jn2}$. Les échantillons X et Y sont d'effectifs respectifs n1 et n2. Le test consiste à regarder si ces deux échantillons sont issus d'une même loi et donc, en particulier, si les moyennes correspondantes respectives μ_1 et μ_2 sont égales.

Il s'agit donc de tester les hypothèses suivantes :

$$H_0 : L(X) = L(Y) \text{ contre } H_1 : L(X) \neq L(Y)$$

Comme un certain nombre de tests non-paramétriques, le test de Mann-Whitney s'intéresse aux rangs des observations. L'idée du test est simple : sous H_0, les deux séries de valeurs numériques ont la même moyenne, donc, si on regroupe ces deux séries de valeurs et qu'on les classe dans l'ordre croissant, les valeurs des deux échantillons devraient être réparties de manière homogène. Cela signifie que la somme des rangs W_X des éléments de X et la somme des rangs W_Y des éléments de Y sont significativement proches.

On s'intéresse alors à la loi que suit, par exemple, la variable W_X. Sans entrer dans les détails techniques, on peut montrer que, sous l'hypothèse nulle, W_X suit une loi dont on peut calculer l'espérance et la variance :

$$\mathbb{E}[W_X] = \frac{n1}{2} * (n1 + n2 + 1)$$
$$Var[W_X] = \frac{n1 * n2}{12} * (n1 + n2 + 1)$$

Ce qui nous permet alors de construire le test est le résultat asymptotique suivant, qui est une application du théorème centrale limite :

$$T = \frac{W_X - \frac{n1}{2} * (n1 + n2 + 1)}{\sqrt{\frac{n1*n2}{12} * (n1 + n2 + 1)}} \xrightarrow{+\infty} \mathcal{N}(0, 1)$$

Ainsi, sous l'hypothèse nulle, la statistique T suit une loi normale centrée réduite.
L'hypothèse nulle sera donc rejetée, au seuil $\alpha = 5\%$, si T est supérieure au quantile d'ordre $1-\propto/2$ de la loi normale centrée réduite.

Annexe E.2. : Approximation de la puissance du test de Mann-Whitney

La statistique du test de Mann-Whitney est difficile à manier lorsque l'on se place sous l'hypothèse H_1. L'étude de la puissance est donc compromise. Pour l'approximer, on va donc tenter d'évaluer la puissance d'un test plus simple, proche du test de Mann-Whitney, à savoir le test de Student.

Le test de Student que l'on considère est celui permettant de tester l'égalité des moyennes de deux échantillons issus d'une loi normale et dont les variances sont égales. Par ailleurs, la variance doit être supposée connue. Dans notre cas, la variance sera estimée sur les données correspondant aux situations de non-accident, dont l'effectif n2 est très grand par rapport à celui des données d'accidents n1.

Dans ce cadre, la puissance asymptotique du test de Student est :

$$1 - \beta = \phi\left(\frac{\delta}{\sigma\sqrt{\frac{1}{n1}}} - u_{1-\alpha/2}\right)$$

où n1 est le nombre d'accidents inclus, δ la différence observée des moyennes μ_{acc} et $\mu_{non-acc}$, σ l'écart-type supposé connu de la distribution, $u_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite et ϕ la fonction de répartition de la loi normale centrée réduite.

Démontrons cette relation.

On veut tester :

$$H_0 : \mu_{acc} = \mu_{non-acc} \; contre \; H_1 : \mu_{acc} \neq \mu_{non-acc}$$

où μ_{acc} est la moyenne des n1 valeurs de l'échantillon correspondant aux situations d'accident et $\mu_{non-acc}$ la moyenne des n2 valeurs de l'échantillon correspondant aux situations de non-accident

D'après les hypothèses qu'on a faites, les estimateurs $\widehat{\mu_{acc}}$ et $\widehat{\mu_{non-acc}}$ de ces moyennes suivent les lois suivantes :

$$\widehat{\mu_{acc}} \sim \mathcal{N}\left(\mu_{acc}, \frac{\sigma^2}{n1}\right) \text{ et } \widehat{\mu_{non-acc}} \sim \mathcal{N}\left(\mu_{non-acc}, \frac{\sigma^2}{n2}\right)$$

La statistique dont on va se servir pour le test est :

$$T = \frac{\widehat{\mu_{acc}} - \widehat{\mu_{non-acc}}}{\sigma * \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

D'après les propriétés bien connues sur les vecteurs gaussiens, T va suivre sous H_0 une loi normale centrée réduite. La valeur critique au-delà de laquelle le test va rejeter l'hypothèse nulle est donc le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite. Nous pouvons à présent calculer la puissance du test $1 - \beta$.

$$1 - \beta = \mathbb{P}\left(T > u_{1-\alpha/2} \middle| H_1 \text{ est vraie}\right) = \mathbb{P}\left(\frac{\widehat{\mu_{acc}} - \widehat{\mu_{non-acc}}}{\sigma * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} > u_{1-\alpha/2}\right)$$

Or, sous H_1, si on note $\delta = \mu_{acc} - \mu_{non-acc}$, alors :

$$\frac{\widehat{\mu_{acc}} - \widehat{\mu_{non-acc}} - \delta}{\sigma * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \sim \mathcal{N}(0,1)$$

D'où : $1 - \beta = \mathbb{P}\left(\frac{\widehat{\mu_{acc}} - \widehat{\mu_{non-acc}}}{\sigma * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} > u_{1-\alpha/2}\right) = \mathbb{P}\left(\frac{\widehat{\mu_{acc}} - \widehat{\mu_{non-acc}} - \delta}{\sigma * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} > u_{1-\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n1} + \frac{1}{n2}}}\right)$

Et donc, par symétrie de la $\mathcal{N}(0,1)$:

$$1 - \beta = \mathbb{P}\left(\frac{\widehat{\mu_{acc}} - \widehat{\mu_{non-acc}} - \delta}{\sigma * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \leq \frac{\delta}{\sigma\sqrt{\frac{1}{n1} + \frac{1}{n2}}} - u_{1-\alpha/2}\right)$$

Or, puisque $\frac{\widehat{\mu_{acc}} - \widehat{\mu_{non-acc}} - \delta}{\sigma * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \sim \mathcal{N}(0,1)$, si l'on note ϕ la fonction de répartition de la loi normale centrée réduite, on obtient finalement la relation :

$$1 - \beta = \phi\left(\frac{\delta}{\sigma\sqrt{\frac{1}{n1} + \frac{1}{n2}}} - u_{1-\alpha/2}\right)$$

Dans notre cas, comme n2 est très grand devant n1, la puissance devient :

$$1 - \beta = \phi\left(\frac{\delta}{\sigma\sqrt{\frac{1}{n1}}} - u_{1-\alpha/2}\right)$$

Cette formule permet de trouver facilement le nombre n1 d'accidents qu'il faudrait inclure pour obtenir une puissance $1 - \beta$ fixée (par exemple à 80%) :

$$n1 = \left(u_{1-\alpha/2} + u_{1-\beta}\right)^2 * \frac{\sigma^2}{\delta^2}$$

δ correspond à la différence que l'on souhaite détecter à la puissance $1 - \beta$: c'est donc soit une valeur estimée sur les données, soit une valeur de référence fixée de manière arbitraire.

Annexe E.3. : Code R de création du tableau des seuils et des graphes de performances

Il s'agit du code R présentant la construction du tableau des seuils, ainsi que des graphes de performances dans l'étude des accidents de type isolés. Les codes correspondant à l'étude des accidents de type non-isolés ou à l'étude croisée sont très similaires à celui-ci.

```
#méthodes d'optimisation des seuils de coupure pour les accidents de type "Seul"
#la seule variable à tester est "vitessemoy"

donneesACC=donnees$vitessemoy[donnees$IndAcc==1 & donnees$TypeAcc=="Seul"]
donneesNONACC=donnees$vitessemoy[donnees$IndAcc==0]

indicateur="vitessemoy" #c'est la variable à tester
DonneesInd=donnees[,indicateur]

seuils=seq(min(donneesACC,na.rm=T),max(donneesACC,na.rm=T),0.2) #on crée un vecteur de seuils
Taille=length(seuils)

#pour chaque seuil donné, on obtiendra un tableau de contingence,
#défini par les nombres Vrai1, Vrai0, Faux1, Faux0

Vrai1=1:Taille
Vrai0=1:Taille
Faux1=1:Taille
Faux0=1:Taille

for (k in 1:Taille)
{
      Vrai1[k]=sum(donnees$IndAcc==1 & donnees$TypeAcc=="Seul" & DonneesInd>seuils[k],na.rm=T)
      Vrai0[k]=sum(donnees$IndAcc==0 & DonneesInd<=seuils[k],na.rm=T)
      Faux1[k]=sum(donnees$IndAcc==1 & donnees$TypeAcc=="Seul" & DonneesInd<=seuils[k],na.rm=T)
      Faux0[k]=sum(donnees$IndAcc==0 & DonneesInd>seuils[k],na.rm=T)
}

#à partir de Vrai1, Vrai0, Faux1 et Faux0, on calcule les mesures de performances
#du tableau de contingence :

Sensibilite=Vrai1/(Vrai1+Faux0)
Specificite=Vrai0/(Vrai0+Faux1)
TauxFaux1=Faux1/(Vrai1+Faux1)
TauxFaux0=Faux0/(Vrai0+Faux0)

OR=1:Taille
OR=(Vrai1/Faux1)/(Faux0/Vrai0)
OR[OR<=1]=1/OR[OR<=1] #pour faire en sorte que l'OR soit toujours supérieur à 1

IndYouden=1-(TauxFaux0+TauxFaux1)
IndYoudenPond=1-(0.2*TauxFaux0+TauxFaux1)
```

```
Vrai1=as.numeric(Vrai1)
Vrai0=as.numeric(Vrai0)
Faux1=as.numeric(Faux1)
Faux0=as.numeric(Faux0)

c=(Vrai1*Vrai0+0.5*Vrai1*Faux1+0.5*Vrai0*Faux0)/((Vrai1+Faux0)*(Faux1+Vrai0))

#création du tableau de synthèse :
mat=matrix(c(seuils,Vrai1,Vrai0,Faux1,Faux0,Sensibilite,Specificite,TauxFaux1,TauxFaux0,OR,IndYouden,
IndYoudenPond,c),ncol=13)
colnames(mat)=c("seuils","Vrai1","Vrai0","Faux1","Faux0","Sensibilite","Specificite","TauxFaux1",
"TauxFaux0","OR","IndYouden","IndYoudenPond","c")

#affichage des différents graphes :

x11()
plot(seuils,Sensibilite,pch=20,col="orange",las=1,main="Sensibilité et spécificité",xlab="seuils",ylab=NA,yaxt="n")
axis(side=2,col="orange",lwd=2)
par(new=T)
plot(seuils,Specificite,pch=20,col="royalblue4",main="Sensibilité et spécificité",xlab="seuils",ylab=NA,yaxt="n")
axis(side=4,col="royalblue4",lwd=2)
legend("topright",c("Sensiblité","Spécificité"),col=c("orange","royalblue4"),lty=c(2,1,1,1,1),pch=20)

x11()
plot(seuils,TauxFaux1,pch=20,col="orange",las=1,main="TauxFaux1 et TauxFaux0",xlab="seuils",ylab=NA,yaxt="n")
axis(side=2,col="orange",lwd=2)
par(new=T)
plot(seuils,TauxFaux0,pch=20,col="royalblue4",main="TauxFaux1 et TauxFaux0",xlab="seuils",ylab=NA,yaxt="n")
axis(side=4,col="royalblue4",lwd=2)
legend("top",c("TauxFaux1","TauxFaux0"),col=c("orange","royalblue4"),lty=c(2,1,1,1,1),pch=20)

x11()
plot(seuils,OR,pch=20,col="orange",las=1,main="OR",xlab="seuils",ylab=NA)

x11()
plot(seuils,OR,pch=20,col="orange",las=1,main="OR",xlab="seuils",ylab=NA)

x11()
plot(seuils,IndYouden,pch=20,col="orange",las=1,main="Indice de Youden et indice de Youden pondéré",
xlab="seuils",ylab=NA,yaxt="n")
axis(side=2,col="orange",lwd=2)
par(new=T)
plot(seuils,IndYoudenPond,pch=20,col="royalblue4",main="Indice de Youden et indice de Youden pondéré",
xlab="seuils",ylab=NA,yaxt="n")
axis(side=4,col="royalblue4",lwd=2)
legend("topright",c("IndYouden","IndYoudenPond"),col=c("orange","royalblue4"),lty=c(2,1,1,1,1),pch=20)

plot(seuils,c,pch=20,col="royalblue4",las=1,main="coefficient c",xlab="seuils",ylab=NA)
```

Annexe E.4. : Résultats des tests dans le cadre de l'étude des accidents isolés

```
> wilcox.test(donneesACC, donneesNONACC)

        Wilcoxon rank sum test with continuity correction

data:  donneesACC and donneesNONACC
W = 3078560, p-value = 0.3656
alternative hypothesis: true location shift is not equal to 0
```

Résultats du test de Mann-Whitney

```
> summary(modele1)

Call:
glm(formula = IndAcc ~ INDvitessemoy, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.0034  -0.0034  -0.0017  -0.0017   5.1879

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -13.457      1.000 -13.457   <2e-16 ***
INDvitessemoy   1.391      1.155   1.205    0.228
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109.03  on 1220512  degrees of freedom
Residual deviance: 107.31  on 1220511  degrees of freedom
  (88916 observations deleted due to missingness)
AIC: 111.31

Number of Fisher Scoring iterations: 16
```

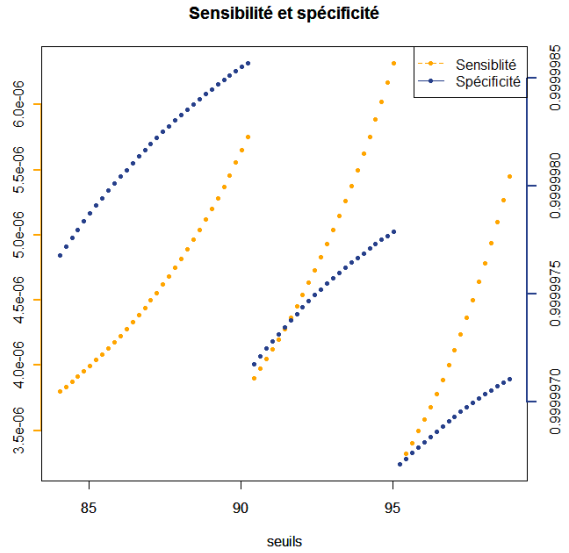
Résultats du test de Wald

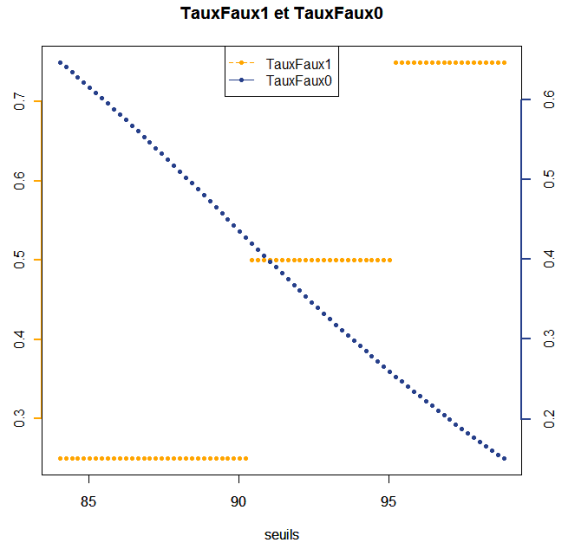
Annexe E.5. : Tableau des seuils pour l'étude des accidents isolés

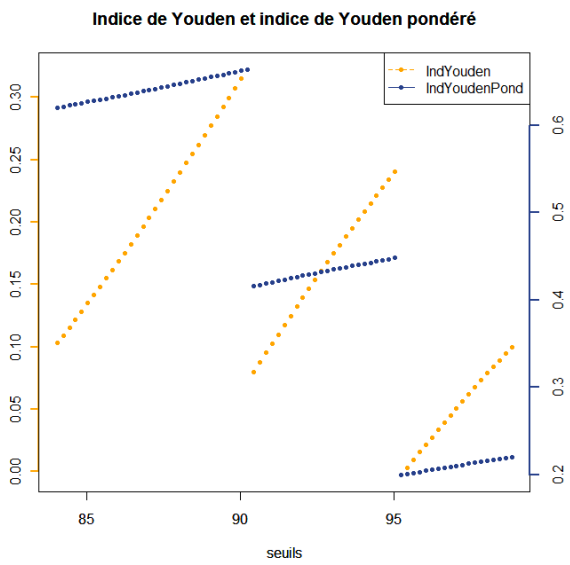
	seuils	Vrai1	Vrai0	Faux1	Faux0	Sensibilite	Specificite
[1,]	84.03371	3	430830	1	789679	3.798998e-06	0.9999977
[2,]	84.23371	3	438330	1	782179	3.835424e-06	0.9999977
[3,]	84.43371	3	446033	1	774476	3.873572e-06	0.9999978
[4,]	84.63371	3	453833	1	766676	3.912981e-06	0.9999978
[5,]	84.83371	3	461877	1	758632	3.954471e-06	0.9999978
[6,]	85.03371	3	469894	1	750615	3.996707e-06	0.9999979
[7,]	85.23371	3	478084	1	742425	4.040796e-06	0.9999979
[8,]	85.43371	3	486245	1	734264	4.085707e-06	0.9999979
[9,]	85.63371	3	494329	1	726180	4.131190e-06	0.9999980
[10,]	85.83371	3	502596	1	717913	4.178762e-06	0.9999980
[11,]	86.03371	3	510922	1	709587	4.227794e-06	0.9999980
[12,]	86.23371	3	519157	1	701352	4.277434e-06	0.9999981
[13,]	86.43371	3	527623	1	692886	4.329698e-06	0.9999981
[14,]	86.63371	3	536148	1	684361	4.383632e-06	0.9999981
[15,]	86.83371	3	544869	1	675640	4.440215e-06	0.9999982
[16,]	87.03371	3	553630	1	666879	4.498547e-06	0.9999982
[17,]	87.23371	3	562288	1	658221	4.557719e-06	0.9999982
[18,]	87.43371	3	571160	1	649349	4.619990e-06	0.9999982
[19,]	87.63371	3	580025	1	640484	4.683936e-06	0.9999983
[20,]	87.83371	3	588921	1	631588	4.749909e-06	0.9999983
[21,]	88.03371	3	598039	1	622470	4.819486e-06	0.9999983
[22,]	88.23371	3	607017	1	613492	4.890015e-06	0.9999984
[23,]	88.43371	3	616159	1	604350	4.963986e-06	0.9999984
[24,]	88.63371	3	625107	1	595402	5.038587e-06	0.9999984
[25,]	88.83371	3	634330	1	586179	5.117864e-06	0.9999984
[26,]	89.03371	3	643575	1	576934	5.199875e-06	0.9999984
[27,]	89.23371	3	652719	1	567790	5.283616e-06	0.9999985
[28,]	89.43371	3	661900	1	558609	5.370454e-06	0.9999985
[29,]	89.63371	3	671006	1	549503	5.459449e-06	0.9999985
[30,]	89.83371	3	680442	1	540067	5.554835e-06	0.9999985
[31,]	90.03371	3	689914	1	530595	5.653998e-06	0.9999986
[32,]	90.23371	3	698940	1	521569	5.751843e-06	0.9999986
[33,]	90.43371	2	708082	2	512427	3.902980e-06	0.9999972
[34,]	90.63371	2	717118	2	503391	3.973039e-06	0.9999972
[35,]	90.83371	2	726545	2	493964	4.048862e-06	0.9999972
[36,]	91.03371	2	735571	2	484938	4.124222e-06	0.9999973
[37,]	91.23371	2	744434	2	476075	4.201001e-06	0.9999973
[38,]	91.43371	2	753271	2	467238	4.280455e-06	0.9999973
[39,]	91.63371	2	762349	2	458160	4.365268e-06	0.9999974
[40,]	91.83371	2	771518	2	448991	4.454412e-06	0.9999974
[41,]	92.03371	2	780348	2	440161	4.543771e-06	0.9999974
[42,]	92.23371	2	789075	2	431434	4.635682e-06	0.9999975
[43,]	92.43371	2	797882	2	422627	4.732283e-06	0.9999975
[44,]	92.63371	2	806512	2	413997	4.830930e-06	0.9999975
[45,]	92.83371	2	815196	2	405313	4.934434e-06	0.9999975
[46,]	93.03371	2	823713	2	396796	5.040348e-06	0.9999976
[47,]	93.23371	2	832122	2	388387	5.149476e-06	0.9999976
[48,]	93.43371	2	840382	2	380127	5.261372e-06	0.9999976
[49,]	93.63371	2	848679	2	371830	5.378773e-06	0.9999976
[50,]	93.83371	2	856823	2	363686	5.499219e-06	0.9999977
[51,]	94.03371	2	864783	2	355726	5.622273e-06	0.9999977
[52,]	94.23371	2	872870	2	347639	5.753061e-06	0.9999977
[53,]	94.43371	2	880705	2	339804	5.885711e-06	0.9999977
[54,]	94.63371	2	888523	2	331986	6.024314e-06	0.9999977
[55,]	94.83371	2	896297	2	324212	6.168765e-06	0.9999978
[56,]	95.03371	2	904118	2	316391	6.321252e-06	0.9999978
[57,]	95.23371	1	911866	3	308643	3.239979e-06	0.9999967
[58,]	95.43371	1	919353	3	301156	3.320527e-06	0.9999967
[59,]	95.63371	1	926929	3	293580	3.406215e-06	0.9999968
[60,]	95.83371	1	934355	3	286154	3.494610e-06	0.9999968
[61,]	96.03371	1	941546	3	278963	3.584692e-06	0.9999968
[62,]	96.23371	1	948797	3	271712	3.680354e-06	0.9999968
[63,]	96.43371	1	956043	3	264466	3.781190e-06	0.9999969
[64,]	96.63371	1	963380	3	257129	3.889083e-06	0.9999969
[65,]	96.83371	1	970486	3	250023	3.999616e-06	0.9999969
[66,]	97.03371	1	977616	3	242893	4.117022e-06	0.9999969
[67,]	97.23371	1	984562	3	235947	4.238222e-06	0.9999970
[68,]	97.43371	1	991423	3	229086	4.365154e-06	0.9999970
[69,]	97.63371	1	998209	3	222300	4.498405e-06	0.9999970
[70,]	97.83371	1	1004938	3	215571	4.638821e-06	0.9999970
[71,]	98.03371	1	1011499	3	209010	4.784437e-06	0.9999970
[72,]	98.23371	1	1017955	3	202554	4.936931e-06	0.9999971
[73,]	98.43371	1	1024390	3	196119	5.098919e-06	0.9999971
[74,]	98.63371	1	1030798	3	189711	5.271148e-06	0.9999971
[75,]	98.83371	1	1037000	3	183509	5.449294e-06	0.9999971

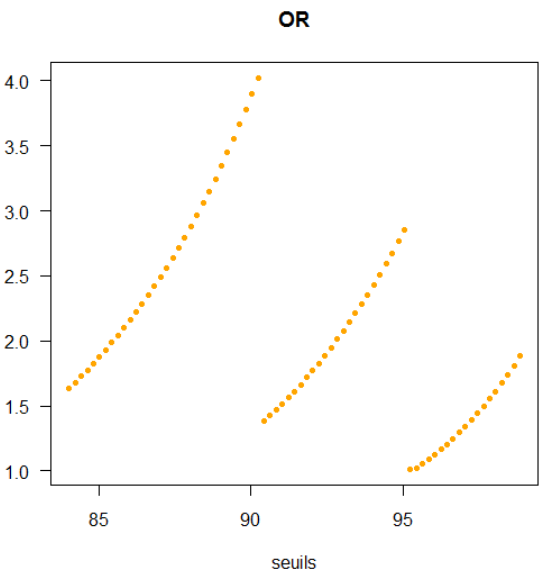
	TauxFaux1	TauxFaux0	OR	IndYouden	IndYoudenPond	c
[1,]	0.25	0.6470079	1.636728	0.102992071	0.6205984	0.5000007
[2,]	0.25	0.6408630	1.681188	0.109137049	0.6218274	0.5000008
[3,]	0.25	0.6345517	1.727748	0.115448350	0.6230897	0.5000008
[4,]	0.25	0.6281609	1.775847	0.121839126	0.6243678	0.5000009
[5,]	0.25	0.6215702	1.826486	0.128429819	0.6256860	0.5000009
[6,]	0.25	0.6150016	1.878036	0.134998390	0.6269997	0.5000009
[7,]	0.25	0.6082913	1.931848	0.141708705	0.6283417	0.5000010
[8,]	0.25	0.6016047	1.986663	0.148395260	0.6296791	0.5000010
[9,]	0.25	0.5949813	2.042175	0.155018726	0.6310037	0.5000011
[10,]	0.25	0.5882079	2.100238	0.161792129	0.6323584	0.5000011
[11,]	0.25	0.5813861	2.160082	0.168613873	0.6337228	0.5000011
[12,]	0.25	0.5746389	2.220670	0.175361058	0.6350722	0.5000012
[13,]	0.25	0.5677025	2.284458	0.182297509	0.6364595	0.5000012
[14,]	0.25	0.5607177	2.350286	0.189282299	0.6378565	0.5000013
[15,]	0.25	0.5535723	2.419346	0.196427679	0.6392855	0.5000013
[16,]	0.25	0.5463942	2.490542	0.203605832	0.6407212	0.5000013
[17,]	0.25	0.5393004	2.562762	0.210699593	0.6421399	0.5000014
[18,]	0.25	0.5320313	2.638766	0.217968692	0.6435937	0.5000014
[19,]	0.25	0.5247679	2.716813	0.225232055	0.6450464	0.5000015
[20,]	0.25	0.5174792	2.797335	0.232520817	0.6465042	0.5000015
[21,]	0.25	0.5100085	2.882255	0.239991471	0.6479983	0.5000016
[22,]	0.25	0.5026526	2.968337	0.247347418	0.6494695	0.5000016
[23,]	0.25	0.4951623	3.058620	0.254837736	0.6509675	0.5000017
[24,]	0.25	0.4878309	3.149672	0.262169103	0.6524338	0.5000017
[25,]	0.25	0.4802742	3.246432	0.269725787	0.6539452	0.5000018
[26,]	0.25	0.4726995	3.346527	0.277300495	0.6554601	0.5000018
[27,]	0.25	0.4652075	3.448735	0.284792451	0.6569585	0.5000019
[28,]	0.25	0.4576853	3.554723	0.292314723	0.6584629	0.5000019
[29,]	0.25	0.4502245	3.663343	0.299775544	0.6599551	0.5000020
[30,]	0.25	0.4424933	3.779764	0.307506745	0.6615013	0.5000020
[31,]	0.25	0.4347326	3.900794	0.315267442	0.6630535	0.5000021
[32,]	0.25	0.4273373	4.020216	0.322662717	0.6645325	0.5000022
[33,]	0.50	0.4198470	1.381820	0.080153035	0.4160306	0.5000005
[34,]	0.50	0.4124435	1.424575	0.087556503	0.4175113	0.5000006
[35,]	0.50	0.4047197	1.470846	0.095280330	0.4190561	0.5000006
[36,]	0.50	0.3973244	1.516835	0.102675605	0.4205351	0.5000007
[37,]	0.50	0.3900627	1.563691	0.109937329	0.4219875	0.5000008
[38,]	0.50	0.3828222	1.612178	0.117177751	0.4234356	0.5000008
[39,]	0.50	0.3753844	1.663936	0.124615632	0.4249231	0.5000009
[40,]	0.50	0.3678719	1.718337	0.132128071	0.4264256	0.5000009
[41,]	0.50	0.3606372	1.772869	0.139362758	0.4278726	0.5000010
[42,]	0.50	0.3534869	1.828959	0.146513053	0.4293026	0.5000011
[43,]	0.50	0.3462711	1.887911	0.153728895	0.4307458	0.5000011
[44,]	0.50	0.3392003	1.948111	0.160799716	0.4321599	0.5000012
[45,]	0.50	0.3320852	2.011275	0.167914780	0.4335830	0.5000012
[46,]	0.50	0.3251070	2.075911	0.174893016	0.4349786	0.5000013
[47,]	0.50	0.3182172	2.142507	0.181782764	0.4363566	0.5000014
[48,]	0.50	0.3114496	2.210793	0.188550433	0.4377101	0.5000014
[49,]	0.50	0.3046516	2.282438	0.195348416	0.4390697	0.5000015
[50,]	0.50	0.2979790	2.355942	0.202021042	0.4404042	0.5000016
[51,]	0.50	0.2914571	2.431037	0.208542911	0.4417086	0.5000017
[52,]	0.50	0.2848312	2.510852	0.215168835	0.4430338	0.5000017
[53,]	0.50	0.2784117	2.591803	0.221588288	0.4443177	0.5000018
[54,]	0.50	0.2720062	2.676387	0.227993812	0.4455988	0.5000019
[55,]	0.50	0.2656367	2.764540	0.234363286	0.4468727	0.5000020
[56,]	0.50	0.2592287	2.857597	0.240771268	0.4481543	0.5000021
[57,]	0.75	0.2528806	1.015422	-0.002880560	0.1994239	0.5000000
[58,]	0.75	0.2467462	1.017582	0.003253765	0.2006508	0.5000000
[59,]	0.75	0.2405390	1.052443	0.009461012	0.2018922	0.5000001
[60,]	0.75	0.2344546	1.088406	0.015545359	0.2031091	0.5000001
[61,]	0.75	0.2285628	1.125055	0.021437163	0.2042874	0.5000002
[62,]	0.75	0.2226219	1.163974	0.027378127	0.2054756	0.5000003
[63,]	0.75	0.2166850	1.204998	0.033314994	0.2066630	0.5000003
[64,]	0.75	0.2106736	1.248893	0.039326420	0.2078653	0.5000004
[65,]	0.75	0.2048514	1.293862	0.045148581	0.2090297	0.5000005
[66,]	0.75	0.1990096	1.341628	0.050990406	0.2101981	0.5000005
[67,]	0.75	0.1933185	1.390937	0.056681475	0.2113363	0.5000006
[68,]	0.75	0.1876971	1.442578	0.062302900	0.2124606	0.5000007
[69,]	0.75	0.1821371	1.496790	0.067862875	0.2135726	0.5000007
[70,]	0.75	0.1766239	1.553916	0.073376149	0.2146752	0.5000008
[71,]	0.75	0.1712482	1.613159	0.078751775	0.2157504	0.5000009
[72,]	0.75	0.1659586	1.675199	0.084041371	0.2168083	0.5000010
[73,]	0.75	0.1606862	1.741103	0.089313762	0.2178628	0.5000011
[74,]	0.75	0.1554360	1.811172	0.094564030	0.2189128	0.5000012
[75,]	0.75	0.1503545	1.883650	0.099645517	0.2199291	0.5000013

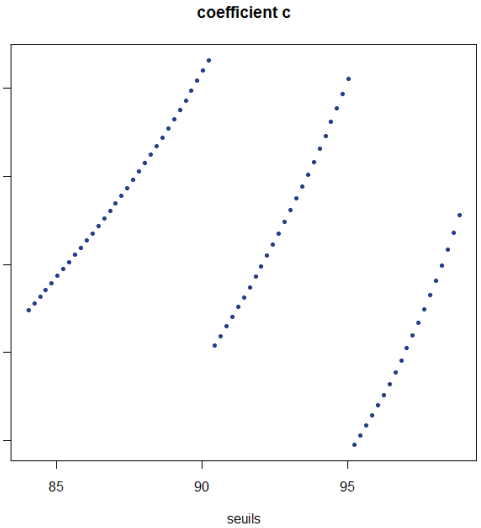
Annexe E.6. : Graphes des performances pour l'étude des accidents isolés











Annexe E.7. : Résultats des tests dans le cadre de l'étude des accidents non-isolés

```
> wilcox.test(donneesACC, donneesNONACC)

        Wilcoxon rank sum test with continuity correction

data:  donneesACC and donneesNONACC
W = 6526360, p-value = 0.7032
alternative hypothesis: true location shift is not equal to 0
```

Résultats du test de Mann-Whitney

```
> summary(modele1)

Call:
glm(formula = IndAcc ~ INDpPICUD0, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.0046  -0.0046  -0.0046  -0.0029   4.9718

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3596     0.7071 -17.479   <2e-16 ***
INDpPICUD0    0.9057     0.7906   1.146    0.252
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 254.24  on 1220518  degrees of freedom
Residual deviance: 252.70  on 1220517  degrees of freedom
  (88916 observations deleted due to missingness)
AIC: 256.7

Number of Fisher Scoring iterations: 15
```

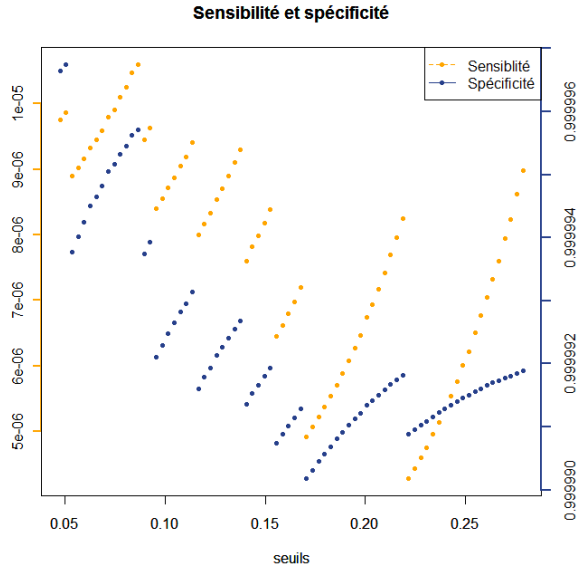
Résultats du test de Wald

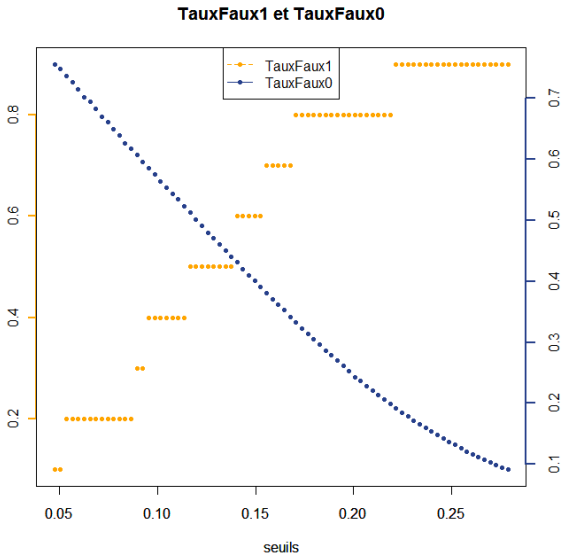
Annexe E.8. : Tableau des seuils pour l'étude des accidents non-isolés

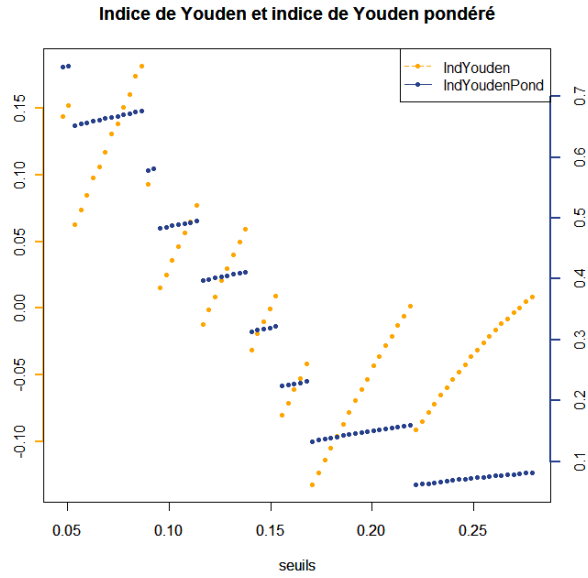
	seuils	Vrai1	Vrai0	Faux1	Faux0	Sensibilite	Specificite
[1,]	0.04761905	9	297702	1	922807	9.752757e-06	0.9999966
[2,]	0.05061905	9	308040	1	912469	9.863251e-06	0.9999968
[3,]	0.05361905	8	321146	2	899363	8.895106e-06	0.9999938
[4,]	0.05661905	8	334595	2	885914	9.030140e-06	0.9999940
[5,]	0.05961905	8	347322	2	873187	9.161757e-06	0.9999942
[6,]	0.06261905	8	363540	2	856969	9.335140e-06	0.9999945
[7,]	0.06561905	8	373724	2	846785	9.447409e-06	0.9999946
[8,]	0.06861905	8	386759	2	833750	9.595110e-06	0.9999948
[9,]	0.07161905	8	403606	2	816903	9.792988e-06	0.9999950
[10,]	0.07461905	8	413192	2	807317	9.909268e-06	0.9999952
[11,]	0.07761905	8	428216	2	792293	1.009717e-05	0.9999953
[12,]	0.08061905	8	439967	2	780542	1.024918e-05	0.9999955
[13,]	0.08361905	8	456991	2	763518	1.047770e-05	0.9999956
[14,]	0.08661905	8	466361	2	754148	1.060788e-05	0.9999957
[15,]	0.08961905	7	479748	3	740761	9.449652e-06	0.9999937
[16,]	0.09261905	7	494217	3	726292	9.637904e-06	0.9999939
[17,]	0.09561905	6	506672	4	713837	8.405210e-06	0.9999921
[18,]	0.09861905	6	518886	4	701623	8.551528e-06	0.9999923
[19,]	0.10161905	6	532268	4	688241	8.717800e-06	0.9999925
[20,]	0.10461905	6	544609	4	675900	8.876974e-06	0.9999927
[21,]	0.10761905	6	557287	4	663222	9.046663e-06	0.9999928
[22,]	0.11061905	6	567728	4	652781	9.191360e-06	0.9999930
[23,]	0.11361905	6	582881	4	637628	9.409787e-06	0.9999931
[24,]	0.11661905	5	595372	5	625137	7.998183e-06	0.9999916
[25,]	0.11961905	5	608572	5	611937	8.170709e-06	0.9999918
[26,]	0.12261905	5	620439	5	600070	8.332292e-06	0.9999919
[27,]	0.12561905	5	635397	5	585112	8.545299e-06	0.9999921
[28,]	0.12861905	5	646242	5	574267	8.706676e-06	0.9999923
[29,]	0.13161905	5	658874	5	561635	8.902500e-06	0.9999924
[30,]	0.13461905	5	671226	5	549283	9.102693e-06	0.9999926
[31,]	0.13761905	5	682933	5	537576	9.300924e-06	0.9999927
[32,]	0.14061905	4	694079	6	526430	7.598293e-06	0.9999914
[33,]	0.14361905	4	708913	6	511596	7.818608e-06	0.9999915
[34,]	0.14661905	4	719717	6	500792	7.987284e-06	0.9999917
[35,]	0.14961905	4	731966	6	488543	8.187544e-06	0.9999918
[36,]	0.15261905	4	743837	6	476672	8.391444e-06	0.9999919
[37,]	0.15561905	3	756074	7	464435	6.459420e-06	0.9999907
[38,]	0.15861905	3	767717	7	452792	6.625515e-06	0.9999909
[39,]	0.16161905	3	779541	7	440968	6.803168e-06	0.9999910
[40,]	0.16461905	3	790269	7	430240	6.972804e-06	0.9999911
[41,]	0.16761905	3	803493	7	417016	7.193917e-06	0.9999913
[42,]	0.17061905	2	814453	8	406056	4.925405e-06	0.9999902
[43,]	0.17361905	2	825974	8	394535	5.069233e-06	0.9999903
[44,]	0.17661905	2	837713	8	382796	5.224688e-06	0.9999905
[45,]	0.17961905	2	848127	8	372382	5.370800e-06	0.9999906
[46,]	0.18261905	2	859363	8	361146	5.537896e-06	0.9999907
[47,]	0.18561905	2	870461	8	350048	5.713470e-06	0.9999908
[48,]	0.18861905	2	880805	8	339704	5.887444e-06	0.9999909
[49,]	0.19161905	2	891487	8	329022	6.078584e-06	0.9999910
[50,]	0.19461905	2	901649	8	318860	6.272306e-06	0.9999911
[51,]	0.19761905	2	911202	8	309307	6.466026e-06	0.9999912
[52,]	0.20061905	2	923754	8	296755	6.739521e-06	0.9999913
[53,]	0.20361905	2	932092	8	288417	6.934356e-06	0.9999914
[54,]	0.20661905	2	941938	8	278571	7.179447e-06	0.9999915
[55,]	0.20961905	2	951042	8	269467	7.422004e-06	0.9999916
[56,]	0.21261905	2	960541	8	259968	7.693195e-06	0.9999917
[57,]	0.21561905	2	969313	8	251196	7.961847e-06	0.9999917
[58,]	0.21861905	2	977973	8	242536	8.246131e-06	0.9999918
[59,]	0.22161905	1	986825	9	233684	4.279265e-06	0.9999909
[60,]	0.22461905	1	995045	9	225464	4.435278e-06	0.9999910
[61,]	0.22761905	1	1002956	9	217553	4.596560e-06	0.9999910
[62,]	0.23061905	1	1010334	9	210175	4.757917e-06	0.9999911
[63,]	0.23361905	1	1018707	9	201802	4.955328e-06	0.9999912
[64,]	0.23661905	1	1026021	9	194488	5.141679e-06	0.9999912
[65,]	0.23961905	1	1033274	9	187235	5.340853e-06	0.9999913
[66,]	0.24261905	1	1040249	9	180260	5.547512e-06	0.9999913
[67,]	0.24561905	1	1047084	9	173425	5.766148e-06	0.9999914
[68,]	0.24861905	1	1054004	9	166505	6.005790e-06	0.9999915
[69,]	0.25161905	1	1059857	9	160652	6.224596e-06	0.9999915
[70,]	0.25461905	1	1066874	9	153635	6.508891e-06	0.9999916
[71,]	0.25761905	1	1072791	9	147718	6.769610e-06	0.9999916
[72,]	0.26061905	1	1078511	9	141998	7.042303e-06	0.9999917
[73,]	0.26361905	1	1084067	9	136442	7.329068e-06	0.9999917
[74,]	0.26661905	1	1088966	9	131543	7.602019e-06	0.9999917
[75,]	0.26961905	1	1094578	9	125931	7.940793e-06	0.9999918

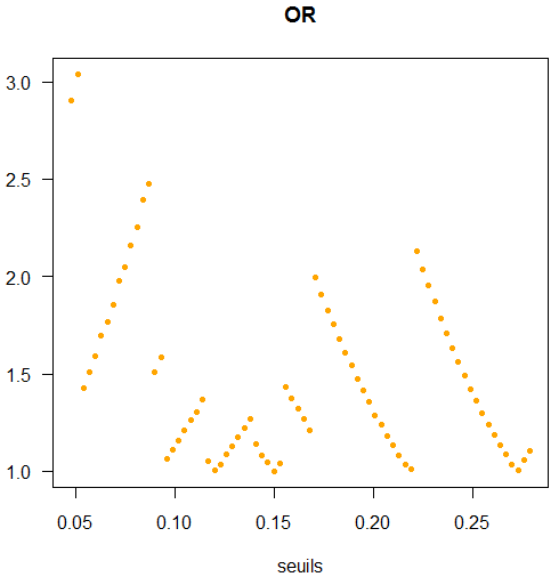
	TauxFaux1	TauxFaux0	OR	IndYouden	IndYoudenPond	c
[1,]	0.1	0.75608373	2.903444	0.1439162677	0.74878325	0.5000032
[2,]	0.1	0.74761350	3.038306	0.1523865043	0.75047730	0.5000033
[3,]	0.2	0.73687535	1.428326	0.0631246472	0.65262493	0.5000013
[4,]	0.2	0.72585618	1.510734	0.0741438203	0.65482876	0.5000015
[5,]	0.2	0.71542856	1.591054	0.0845714370	0.65691429	0.5000017
[6,]	0.2	0.70214066	1.696864	0.0978593357	0.65957187	0.5000019
[7,]	0.2	0.69379660	1.765378	0.1062033955	0.66124068	0.5000020
[8,]	0.2	0.68311663	1.855515	0.1168833659	0.66337667	0.5000022
[9,]	0.2	0.66931338	1.976274	0.1306866234	0.66613732	0.5000024
[10,]	0.2	0.66145928	2.047235	0.1385407236	0.66770814	0.5000025
[11,]	0.2	0.64914966	2.161907	0.1508503419	0.67017007	0.5000027
[12,]	0.2	0.63952171	2.254674	0.1604782923	0.67209566	0.5000029
[13,]	0.2	0.62557343	2.394133	0.1744265712	0.67488531	0.5000031
[14,]	0.2	0.61789630	2.473578	0.1821036961	0.67642074	0.5000032
[15,]	0.3	0.60692793	1.511165	0.0930720708	0.57861441	0.5000016
[16,]	0.3	0.59507304	1.587754	0.1049269608	0.58098539	0.5000018
[17,]	0.4	0.58486828	1.064680	0.0151317196	0.48302634	0.5000003
[18,]	0.4	0.57486098	1.109327	0.0251390199	0.48502780	0.5000004
[19,]	0.4	0.56389670	1.160062	0.0361032979	0.48722066	0.5000006
[20,]	0.4	0.55378535	1.208631	0.0462146531	0.48924293	0.5000008
[21,]	0.4	0.54339788	1.260408	0.0566021226	0.49132042	0.5000009
[22,]	0.4	0.53484325	1.304560	0.0651567502	0.49303135	0.5000011
[23,]	0.4	0.52242794	1.371209	0.0775720621	0.49551441	0.5000013
[24,]	0.5	0.51219368	1.049994	-0.0121936831	0.39756126	0.4999998
[25,]	0.5	0.50137852	1.005529	-0.0013785232	0.39972430	0.5000000
[26,]	0.5	0.49165553	1.033944	0.0083444694	0.40166889	0.5000001
[27,]	0.5	0.47939999	1.085941	0.0206000120	0.40412000	0.5000003
[28,]	0.5	0.47051435	1.125334	0.0294856490	0.40589713	0.5000005
[29,]	0.5	0.46016457	1.173136	0.0398354293	0.40796709	0.5000007
[30,]	0.5	0.45004420	1.222004	0.0499557971	0.40999116	0.5000008
[31,]	0.5	0.44045230	1.270393	0.0595476969	0.41190954	0.5000010
[32,]	0.6	0.43132005	1.137687	-0.0313200476	0.31373599	0.4999995
[33,]	0.6	0.41916610	1.082494	-0.0191661020	0.31616678	0.4999997
[34,]	0.6	0.41031406	1.043727	-0.0103140575	0.31793719	0.4999998
[35,]	0.6	0.40027808	1.001159	-0.0002780807	0.31994438	0.5000000
[36,]	0.6	0.39055181	1.040320	0.0094481892	0.32188964	0.5000002
[37,]	0.7	0.38052567	1.433301	-0.0805256659	0.22389487	0.4999986
[38,]	0.7	0.37098620	1.376177	-0.0709862033	0.22580276	0.4999988
[39,]	0.7	0.36129844	1.319912	-0.0612984419	0.22774031	0.4999989
[40,]	0.7	0.35250867	1.270319	-0.0525086665	0.22949827	0.4999991
[41,]	0.7	0.34167384	1.211009	-0.0416738426	0.23166523	0.4999992
[42,]	0.8	0.33269398	1.994251	-0.1326939826	0.13346120	0.4999976
[43,]	0.8	0.32325448	1.910641	-0.1232544783	0.13534910	0.4999977
[44,]	0.8	0.31363636	1.827815	-0.1136363599	0.13727273	0.4999978
[45,]	0.8	0.30510385	1.756256	-0.1051038542	0.13897923	0.4999980
[46,]	0.8	0.29589786	1.680994	-0.0958978590	0.14082043	0.4999981
[47,]	0.8	0.28680493	1.608564	-0.0868049314	0.14263901	0.4999983
[48,]	0.8	0.27832978	1.542698	-0.0783297788	0.14433404	0.4999984
[49,]	0.8	0.26957769	1.476284	-0.0695776926	0.14608446	0.4999986
[50,]	0.8	0.26125166	1.414564	-0.0612516581	0.14774967	0.4999987
[51,]	0.8	0.25342460	1.357798	-0.0534245958	0.14931508	0.4999988
[52,]	0.8	0.24314036	1.284996	-0.0431403619	0.15137193	0.4999990
[53,]	0.8	0.23630879	1.237719	-0.0363087859	0.15273824	0.4999992
[54,]	0.8	0.22824166	1.182970	-0.0282416598	0.15435167	0.4999993
[55,]	0.8	0.22078248	1.133355	-0.0207824768	0.15584350	0.4999995
[56,]	0.8	0.21299966	1.082590	-0.0129996583	0.15740007	0.4999997
[57,]	0.8	0.20581249	1.036594	-0.0058124930	0.15883750	0.4999999
[58,]	0.8	0.19871709	1.008070	0.0012829074	0.16025658	0.5000000
[59,]	0.9	0.19146438	2.131235	-0.0914643808	0.06170712	0.4999976
[60,]	0.9	0.18472949	2.039281	-0.0847294858	0.06305410	0.4999977
[61,]	0.9	0.17824776	1.952206	-0.0782477638	0.06435045	0.4999978
[62,]	0.9	0.17220274	1.872227	-0.0722027449	0.06555945	0.4999979
[63,]	0.9	0.16534249	1.782866	-0.0653424924	0.06693150	0.4999981
[64,]	0.9	0.15934991	1.706000	-0.0593499106	0.06813002	0.4999982
[65,]	0.9	0.15340731	1.630850	-0.0534073079	0.06931854	0.4999983
[66,]	0.9	0.14769248	1.559569	-0.0476924791	0.07046150	0.4999984
[67,]	0.9	0.14209236	1.490640	-0.0420923565	0.07158153	0.4999986
[68,]	0.9	0.13642259	1.421764	-0.0364225909	0.07271548	0.4999987
[69,]	0.9	0.13162705	1.364210	-0.0316270507	0.07367459	0.4999989
[70,]	0.9	0.12587781	1.296043	-0.0258778100	0.07482444	0.4999990
[71,]	0.9	0.12102983	1.239255	-0.0210298326	0.07579403	0.4999992
[72,]	0.9	0.11634326	1.184950	-0.0163432633	0.07673135	0.4999993
[73,]	0.9	0.11179106	1.132751	-0.0117910642	0.07764179	0.4999995
[74,]	0.9	0.10777717	1.087166	-0.0077771651	0.07844457	0.4999997
[75,]	0.9	0.10317908	1.035448	-0.0031790835	0.07936418	0.4999999

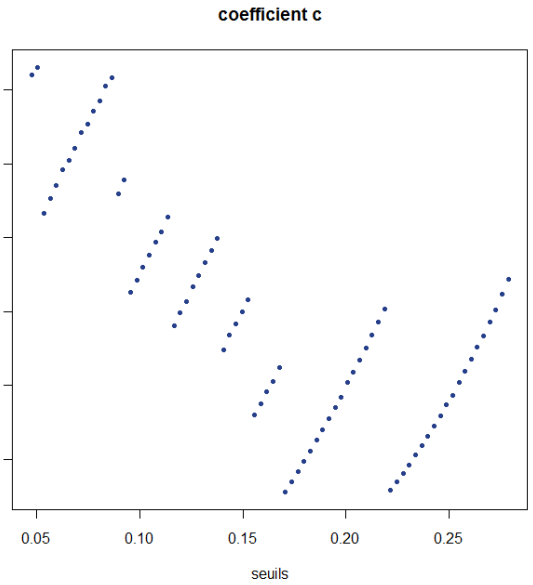
Annexe E.9. : Graphes des performances pour l'étude des accidents non-isolés











Annexe E.10. : Résultats des tests dans le cadre de l'étude croisée entre pPICUD0 et pTIV2

```
> wilcox.test(donneesACC, donneesNONACC)

        Wilcoxon rank sum test with continuity correction

data:  donneesACC and donneesNONACC
W = 214288277, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Résultats du test de Mann-Whitney

```
> summary(modele1)

Call:
glm(formula = INDpPICUD0 ~ INDpTIV2, family = binomial, data = donnees)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.575  -0.013  -0.013  -0.013   4.331

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.37981    0.09854  -95.19   <2e-16 ***
INDpTIV2    10.27901    0.18318   56.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4712.6  on 1220522  degrees of freedom
Residual deviance: 2383.6  on 1220521  degrees of freedom
  (88917 observations deleted due to missingness)
AIC: 2387.6

Number of Fisher Scoring iterations: 12
```

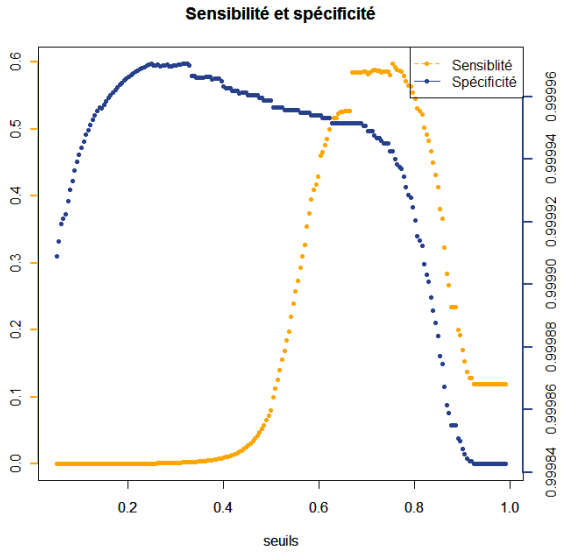
Résultats du test de Wald

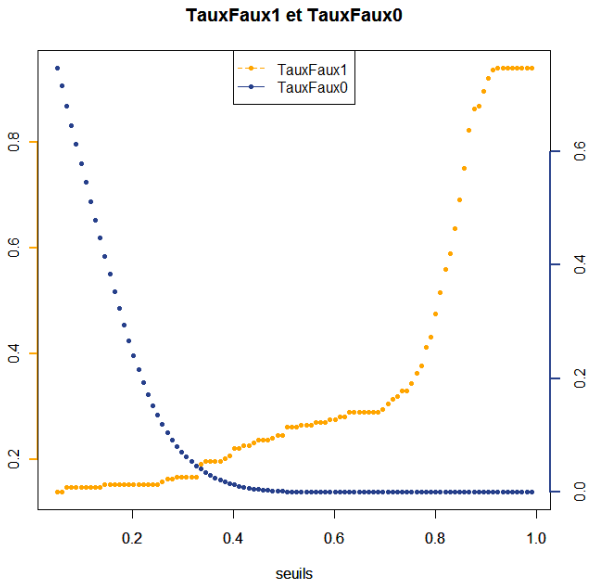
Annexe E.11. : Tableau des seuils pour l'étude croisée entre pPICUD0 et pTIV2

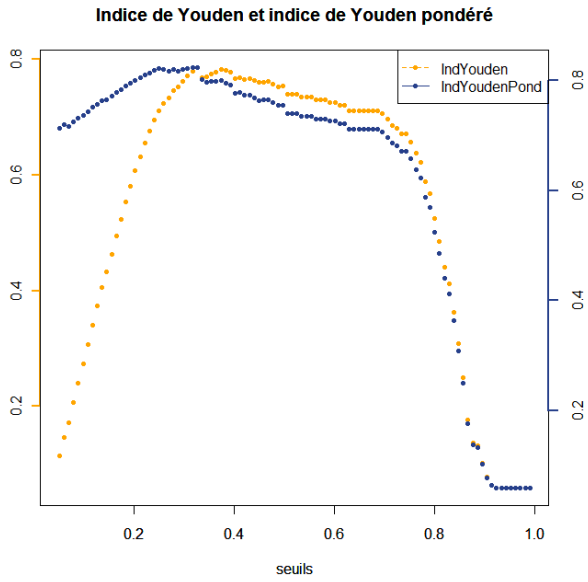
	seuils	Vrai1	Vrai0	Faux1	Faux0	Sensibilite	Specificite	TauxFaux1
[1,]	0.0500000	176	308008	28	912311	0.0001928795	0.9999091	0.1372549
[2,]	0.0627027	175	363516	29	856803	0.0002042059	0.9999202	0.1421569
[3,]	0.0754054	174	416252	30	804067	0.0002163531	0.9999279	0.1470588
[4,]	0.0881081	174	472623	30	747696	0.0002326608	0.9999365	0.1470588
[5,]	0.1008108	174	529495	30	690824	0.0002518097	0.9999433	0.1470588
[6,]	0.1135135	174	582859	30	637460	0.0002728838	0.9999485	0.1470588
[7,]	0.1262162	174	636075	30	584244	0.0002977321	0.9999528	0.1470588
[8,]	0.1389189	174	688543	30	531776	0.0003270984	0.9999564	0.1470588
[9,]	0.1516216	173	740098	31	480221	0.0003601211	0.9999581	0.1519608
[10,]	0.1643243	173	789234	31	431085	0.0004011520	0.9999607	0.1519608
[11,]	0.1770270	173	838472	31	381847	0.0004528559	0.9999630	0.1519608
[12,]	0.1897297	173	885048	31	335271	0.0005157344	0.9999650	0.1519608
[13,]	0.2024324	173	928640	31	291679	0.0005927662	0.9999666	0.1519608
[14,]	0.2151351	173	967989	31	252330	0.0006851404	0.9999680	0.1519608
[15,]	0.2278378	173	1003633	31	216686	0.0007977534	0.9999691	0.1519608
[16,]	0.2405405	173	1035446	31	184873	0.0009349027	0.9999701	0.1519608
[17,]	0.2532432	172	1063685	32	156634	0.0010968968	0.9999699	0.1568627
[18,]	0.2659459	171	1088099	33	132220	0.0012916286	0.9999697	0.1617647
[19,]	0.2786486	171	1109147	33	111172	0.0015357948	0.9999702	0.1617647
[20,]	0.2913514	170	1126986	34	93333	0.0018181235	0.9999698	0.1666667
[21,]	0.3040541	170	1142462	34	77857	0.0021787330	0.9999702	0.1666667
[22,]	0.3167568	170	1155522	34	64797	0.0026167131	0.9999706	0.1666667
[23,]	0.3294595	169	1166620	35	53699	0.0031372986	0.9999700	0.1715686
[24,]	0.3421622	164	1176357	40	43962	0.0037166297	0.9999660	0.1960784
[25,]	0.3548649	164	1184328	40	35991	0.0045360254	0.9999662	0.1960784
[26,]	0.3675676	164	1191046	40	29273	0.0055712199	0.9999664	0.1960784
[27,]	0.3802703	163	1196669	41	23650	0.0068450006	0.9999657	0.2009804
[28,]	0.3929730	162	1201590	42	18729	0.0085755121	0.9999650	0.2058824
[29,]	0.4056757	159	1205635	45	14684	0.0107121202	0.9999627	0.2205882
[30,]	0.4183784	158	1208847	46	11472	0.0135855546	0.9999619	0.2254902
[31,]	0.4310811	158	1211568	46	8751	0.0177348748	0.9999620	0.2254902
[32,]	0.4437838	157	1213699	47	6620	0.0231665929	0.9999613	0.2303922
[33,]	0.4564865	156	1215462	48	4857	0.0311190904	0.9999605	0.2352941
[34,]	0.4691892	156	1216722	48	3597	0.0415667466	0.9999606	0.2352941
[35,]	0.4818919	155	1217694	49	2625	0.0557553957	0.9999598	0.2401961
[36,]	0.4945946	154	1218365	50	1954	0.0730550285	0.9999590	0.2450980
[37,]	0.5072973	151	1219110	53	1209	0.1110294118	0.9999565	0.2598039
[38,]	0.5200000	151	1219460	53	859	0.1495049505	0.9999565	0.2598039
[39,]	0.5327027	150	1219670	54	649	0.1877346683	0.9999557	0.2647059
[40,]	0.5454054	150	1219837	54	482	0.2373417722	0.9999557	0.2647059
[41,]	0.5581081	149	1219944	55	375	0.2843511450	0.9999549	0.2696078
[42,]	0.5708108	149	1220021	55	298	0.3333333333	0.9999549	0.2696078
[43,]	0.5835135	148	1220091	56	228	0.3936170213	0.9999541	0.2745098
[44,]	0.5962162	148	1220121	56	198	0.4277456647	0.9999541	0.2745098
[45,]	0.6089189	147	1220153	57	166	0.4696485623	0.9999533	0.2794118
[46,]	0.6216216	147	1220172	57	147	0.5000000000	0.9999533	0.2794118
[47,]	0.6343243	145	1220183	59	136	0.5160142349	0.9999516	0.2892157
[48,]	0.6470270	145	1220188	59	131	0.5253623188	0.9999516	0.2892157
[49,]	0.6597297	145	1220189	59	130	0.5272727273	0.9999516	0.2892157
[50,]	0.6724324	145	1220216	59	103	0.5846774194	0.9999517	0.2892157
[51,]	0.6851351	145	1220216	59	103	0.5846774194	0.9999517	0.2892157
[52,]	0.6978378	144	1220217	60	102	0.5853658537	0.9999508	0.2941176
[53,]	0.7105405	142	1220219	62	100	0.5867768595	0.9999492	0.3039216
[54,]	0.7232432	139	1220221	65	98	0.5864978903	0.9999467	0.3186275
[55,]	0.7359459	137	1220222	67	97	0.5854700855	0.9999451	0.3284314
[56,]	0.7486486	134	1220222	70	97	0.5800865801	0.9999426	0.3431373
[57,]	0.7613514	130	1220229	74	90	0.5909090909	0.9999394	0.3627451
[58,]	0.7740541	125	1220229	79	90	0.5813953488	0.9999353	0.3872549
[59,]	0.7867568	117	1220229	87	90	0.5652173913	0.9999287	0.4264706
[60,]	0.7994595	110	1220229	94	90	0.5500000000	0.9999230	0.4607843
[61,]	0.8121622	99	1220230	105	89	0.5265957447	0.9999140	0.5147059
[62,]	0.8248649	86	1220230	118	89	0.4914285714	0.9999033	0.5784314
[63,]	0.8375676	74	1220231	130	88	0.4567901235	0.9998935	0.6372549
[64,]	0.8502703	56	1220231	148	88	0.3888888889	0.9998787	0.7254902
[65,]	0.8629730	41	1220231	163	88	0.3178294574	0.9998664	0.7990196
[66,]	0.8756757	29	1220231	175	88	0.2478632479	0.9998566	0.8578431
[67,]	0.8883784	23	1220231	181	88	0.2072072072	0.9998517	0.8872549
[68,]	0.9010811	18	1220231	186	88	0.1698113208	0.9998476	0.9117647
[69,]	0.9137838	13	1220231	191	88	0.1287128713	0.9998435	0.9362745
[70,]	0.9264865	12	1220231	192	88	0.1200000000	0.9998427	0.9411765
[71,]	0.9391892	12	1220231	192	88	0.1200000000	0.9998427	0.9411765
[72,]	0.9518919	12	1220231	192	88	0.1200000000	0.9998427	0.9411765
[73,]	0.9645946	12	1220231	192	88	0.1200000000	0.9998427	0.9411765
[74,]	0.9772973	12	1220231	192	88	0.1200000000	0.9998427	0.9411765
[75,]	0.9900000	12	1220231	192	88	0.1200000000	0.9998427	0.9411765

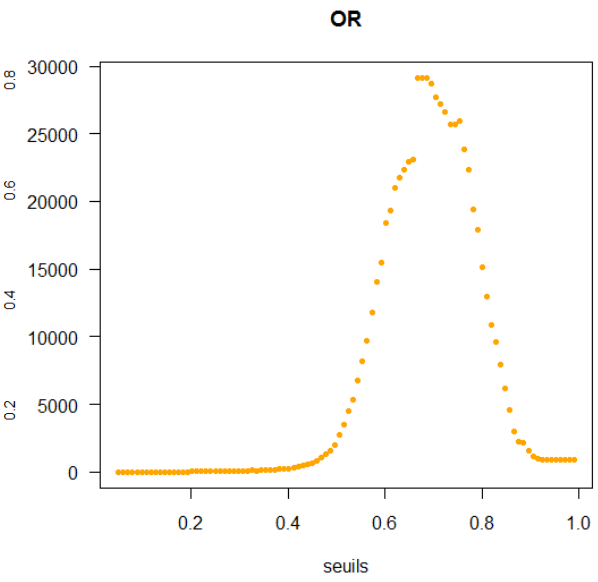
	TauxFaux0	OR	IndYouden	IndYoudenPond	c
[1,]	7.476004e-01	2.122138	0.11514468	0.71322501	0.5000510
[2,]	7.021140e-01	2.560251	0.15572918	0.71742035	0.5000622
[3,]	6.588990e-01	3.002563	0.19404215	0.72116137	0.5000721
[4,]	6.127054e-01	3.666214	0.24023581	0.73040010	0.5000846
[5,]	5.661012e-01	4.445519	0.28684002	0.73972094	0.5000976
[6,]	5.223716e-01	5.303207	0.33056957	0.74846685	0.5001107
[7,]	4.787633e-01	6.314545	0.37417784	0.75718851	0.5001253
[8,]	4.357680e-01	7.509834	0.41717315	0.76578757	0.5001418
[9,]	3.935209e-01	8.600674	0.45451834	0.76933504	0.5001591
[10,]	3.532560e-01	10.217092	0.49478322	0.77738802	0.5001809
[11,]	3.129075e-01	12.254161	0.53513169	0.78545771	0.5002079
[12,]	2.747405e-01	14.731781	0.57329876	0.79309112	0.5002404
[13,]	2.390187e-01	17.767513	0.60902057	0.80023549	0.5002797
[14,]	2.067738e-01	21.408485	0.64126541	0.80668446	0.5003266
[15,]	1.775650e-01	25.848092	0.67047417	0.81252621	0.5003834
[16,]	1.514956e-01	31.256358	0.69654358	0.81774009	0.5004525
[17,]	1.283550e-01	36.501059	0.71478229	0.81746626	0.5005334
[18,]	1.083487e-01	42.643558	0.72988658	0.81656555	0.5006307
[19,]	9.110077e-02	51.698252	0.74713452	0.82001514	0.5007530
[20,]	7.648246e-02	60.374466	0.75685087	0.81803684	0.5008940
[21,]	6.380053e-02	73.369254	0.76953280	0.82057323	0.5010745
[22,]	5.309841e-02	89.164776	0.78023492	0.82271365	0.5012936
[23,]	4.400407e-02	104.901544	0.78442730	0.81963056	0.5015536
[24,]	3.602501e-02	109.709833	0.76789656	0.79671657	0.5018413
[25,]	2.949311e-02	134.915529	0.77442846	0.79802295	0.5022511
[26,]	2.398799e-02	166.818864	0.77993358	0.79912397	0.5027688
[27,]	1.938018e-02	201.162324	0.77963943	0.79514357	0.5034054
[28,]	1.534763e-02	247.461385	0.77877002	0.79104812	0.5042703
[29,]	1.203292e-02	290.105580	0.76737885	0.77700518	0.5053374
[30,]	9.400821e-03	361.935726	0.76510898	0.77262964	0.5067738
[31,]	7.171076e-03	475.542532	0.76733873	0.77307559	0.5088485
[32,]	5.424811e-03	612.427663	0.76418303	0.76852288	0.5115639
[33,]	3.980107e-03	813.310994	0.76072578	0.76390986	0.5155398
[34,]	2.947590e-03	1099.345705	0.76175829	0.76411636	0.5207636
[35,]	2.151077e-03	1467.386356	0.75765284	0.75937371	0.5278576
[36,]	1.601221e-03	1920.452508	0.75330074	0.75458172	0.5365070
[37,]	9.907246e-04	2872.881221	0.73920535	0.73999793	0.5554930
[38,]	7.039143e-04	4044.599029	0.73949216	0.74005530	0.5747307
[39,]	5.318282e-04	5220.296182	0.73476229	0.73518775	0.5938452
[40,]	3.949787e-04	7029.950438	0.73489914	0.73521512	0.6186488
[41,]	3.072967e-04	8813.171200	0.73008486	0.73033070	0.6421530
[42,]	2.441984e-04	11091.100000	0.73014796	0.73034332	0.6666441
[43,]	1.868364e-04	14142.658835	0.72530336	0.72545283	0.6967856
[44,]	1.622527e-04	16285.886364	0.72532794	0.72545775	0.7138499
[45,]	1.360300e-04	18956.086557	0.72045221	0.72056103	0.7348009
[46,]	1.204603e-04	21406.526316	0.72046777	0.72056414	0.7499766
[47,]	1.114463e-04	22049.667871	0.71067287	0.71076202	0.7579829
[48,]	1.073490e-04	22891.352051	0.71067696	0.71076284	0.7626570
[49,]	1.065295e-04	23067.458279	0.71067778	0.71076301	0.7636122
[50,]	8.440416e-05	29114.911963	0.71069991	0.71076743	0.7923145
[51,]	8.440416e-05	29114.911963	0.71069991	0.71076743	0.7923145
[52,]	8.358470e-05	28710.988235	0.70579877	0.70586564	0.7926583
[53,]	8.194579e-05	27946.951290	0.69599649	0.69606204	0.7933630
[54,]	8.030687e-05	26626.486499	0.68129224	0.68135649	0.7932223
[55,]	7.948741e-05	25722.482536	0.67148914	0.67155273	0.7927076
[56,]	7.948741e-05	24080.964359	0.65678326	0.65684685	0.7900146
[57,]	7.375121e-05	23818.283784	0.63718115	0.63724015	0.7954242
[58,]	7.375121e-05	21452.689873	0.61267135	0.61273035	0.7906653
[59,]	7.375121e-05	18233.306897	0.57345566	0.57351466	0.7825730
[60,]	7.375121e-05	15865.861702	0.53914194	0.53920094	0.7749615
[61,]	7.293175e-05	12926.995185	0.48522119	0.48527953	0.7632549
[62,]	7.293175e-05	9992.361455	0.42149570	0.42155404	0.7456659
[63,]	7.211229e-05	7893.102622	0.36267299	0.36273068	0.7283418
[64,]	7.211229e-05	5246.693489	0.27443769	0.27449538	0.6943838
[65,]	7.211229e-05	3487.832613	0.20090828	0.20096597	0.6588479
[66,]	7.211229e-05	2297.837597	0.14208475	0.14214244	0.6238599
[67,]	7.211229e-05	1762.011113	0.11267299	0.11273068	0.6035294
[68,]	7.211229e-05	1341.896261	0.08816318	0.08822087	0.5848295
[69,]	7.211229e-05	943.776951	0.06365338	0.06371107	0.5642782
[70,]	7.211229e-05	866.641335	0.05875142	0.05880911	0.5599213
[71,]	7.211229e-05	866.641335	0.05875142	0.05880911	0.5599213
[72,]	7.211229e-05	866.641335	0.05875142	0.05880911	0.5599213
[73,]	7.211229e-05	866.641335	0.05875142	0.05880911	0.5599213
[74,]	7.211229e-05	866.641335	0.05875142	0.05880911	0.5599213
[75,]	7.211229e-05	866.641335	0.05875142	0.05880911	0.5599213

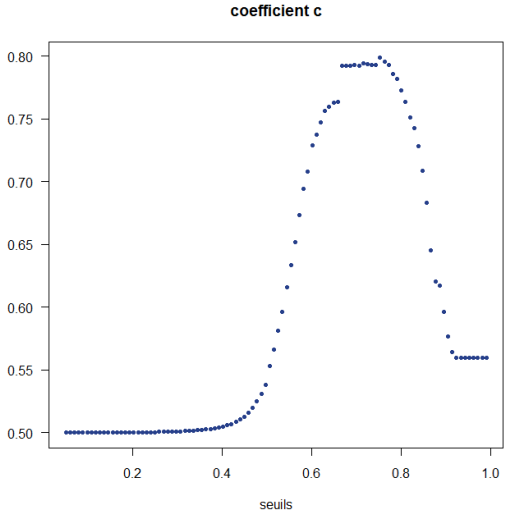
Annexe E.12. : Graphes des performances pour l'étude croisée entre pPICUD0 et pTIV2











Références

Aron M. (2011). *Indicateurs visant à décrire le risque routier selon les conditions de trafic*, Actes Inrets

Cohen S. (1993). *Ingénierie du trafic routier,* Presses de l'école nationale des Ponts et chaussées

Rouvière L. (2008). *Régression sur variables catégorielles.* Polycopié de cours, 71 pages

Taffé P. (2004). *Cours de Régression logistique Appliquée.* Polycopié de cours, 60 pages