Méthodes statistiques appliquées aux mesures du bruit de roulement

Nelly Winzenrieth

Méthodes statistiques appliquées aux mesures du bruit de roulement

Rapport de stage

Laboratoire Régional des Ponts et Chaussées de Strasbourg

Groupe « Acoustique »
Sous la direction de Guillaume Dutilleux et Loïc Toussaint

Nelly Winzenrieth
M1 Statistique
Année 2012-2013

Remerciements

Tout d'abord, je souhaiterais remercier M. George Kuntz, directeur du Laboratoire Régional des Ponts et Chaussées de Strasbourg, de m'avoir accueillie et permis d'effectuer mon stage au sein de son laboratoire.

Je tiens à remercier Guillaume Dutilleux et Loïc Toussaint, mes maîtres de stage, pour leurs conseils tout au long du stage.

Je remercie également Jonas Bauche pour le temps qu'il a consacré à la réalisation de la campagne de mesures ainsi que pour son soutien durant le stage.

Je remercie enfin les membres de l'équipe « Acoustique » qui m'ont accueillie au sein de leur groupe et qui m'ont fait découvrir les différents travaux effectués dans le laboratoire.

Table des matières

Chapitre 1

Présentation de l'entreprise et objectif du stage

1.1 Organisme d'accueil

Mon stage s'est déroulé au sein du groupe acoustique du Laboratoire Régional des Ponts et Chaussées (LRPC) de Strasbourg qui fait partie du Centres d'Etude Techniques de l'Equipement de l'Est (CETE). Il a été effectué sous la direction de M. Guillaume Dutilleux et M. Loïc Toussaint durant la période du 3 juin au 9 août 2013.

1.1.1 Le Centre d'Etudes Techniques de l'Equipement (CETE) de l'Est

Le CETE de l'Est est un service extérieur du Ministère de l'écologie, du développement durable et de l'énergie, apportant des prestations d'ingénierie dans les domaines touchant aux infrastructures et à leur exploitation, à l'équipement, à l'aménagement du territoire ainsi qu'à l'environnement.

Des équipes interdisciplinaires, tant dans les départements d'études que dans les laboratoires, offrent un large éventail de prestations : Etudes, Recherche, Assistance, Expertises, Méthodologie, Contrôle.

Le CETE agit pour le compte de l'Etat, mais se met aussi au service des villes, des collectivités territoriales, et de divers organismes publics, para-publics, privés et étrangers. Il travaille quotidiennement au plus près des réalités du terrain et au service des collectivités locales, toujours en étroit partenariat avec les professionnels des différents secteurs concernés. Cet ancrage lui permet de bénéficier de l'expérience et de la proximité indispensables pour écouter et comprendre les préoccupations des concitoyens, répondre à leurs attentes actuelles et anticiper leurs évolutions futures.

La zone de compétence du CETE couvre 3 régions : la Lorraine, l'Alsace, la Champagne-Ardenne, soit 10 départements. Le laboratoire de Strasbourg mobilise ses compétences prioritairement en Alsace, sur le territoire de Belfort et l'arrondissement de Saint-Dié des Vosges.

1.1.2 Le Laboratoire Régional des Ponts et Chaussées de Strasbourg (LRPC)

Le laboratoire de Strasbourg est dirigé par M. Georges Kuntz et est constitué de cinq groupes techniques qui sont « Géotechnique, Terrassement, Chaussées », « Ouvrages d'art », « Construction », « Acoustique » et « Méthodes Physiques ».

Le groupe acoustique

Le groupe acoustique est dirigé par M. Guillaume Dutilleux et est composé d'une dizaine de personnes. Il est spécialisé dans le domaine du bruit des transports terrestres. L'équipe participe très activement aux recherches portant sur les sources de bruit routier, les ambiances sonores urbaines et la propagation acoustique en site complexe.

1.2 Contexte du stage

L'émission et la propagation du bruit émis par la circulation dépend dans une large mesure des caractéristiques du revêtement de la chaussée, et notamment de sa texture et de sa porosité. Ces deux paramètres exercent une influence notable sur la génération du bruit résultant de l'interaction entre le pneumatique et la chaussée et, de plus, le facteur porosité peut influer sur la propagation du son, en particulier lorsque celle-ci s'effectue de façon proche de la surface du revêtement. Le bruit du groupe motopropulseur, qui est généralement produit à une hauteur plus élevée, au-dessus de la surface du revêtement que le bruit de contact pneumatique/chaussée, peut aussi être affecté durant sa propagation par les caractéristiques de porosité du revêtement de la chaussée. Par conséquent, en fonction des revêtements de chaussée, on relève des variations du niveau sonore pour un même trafic d'un débit et d'une composition donnée. Celles-ci peuvent atteindre jusqu'à 15 dB, ce qui n'est pas sans répercussions sur la qualité de l'environnement le long d'une route. Il est donc important de disposer d'une méthode qui permette de mesurer cette influence et d'établir un classement quantitatif des revêtements de chaussée en fonction du bruit émis par la circulation. L'une des missions du groupe acoustique est de classifier les revêtements selon le niveau de bruit de roulement. Dans cet objectif, des campagnes de mesures de bruit de roulement sont réalisées selon la norme AFNOR ISO 11819.

On distingue deux catégories de véhicules, les véhicules légers et les trains routiers. Un véhicule léger est un véhicule particulier ou utilitaire léger et un train routier est un ensemble routier constitué d'un véhicule tracteur suivi de véhicules remorqués.

On va s'intéresser aux campagnes de mesures qui suivent la procédure dite VI (Véhicules Isolés) ainsi que celles qui suivent la procédure dite VM (Véhicules Maîtrisés). Durant une procédure dite VI, on effectue les mesures selon le trafic en enregistrant tous les véhicules qui passent (selon certains critères explicités ci-dessous). Tandis que durant une procédure dite VM, on effectue les mesures sur des véhicules du laboratoire conduit par des techniciens à des vitesses définies au préalable.

Ces mesures sont réalisées grâce à différents instruments de mesure. On mesure le niveau sonore à l'aide d'un sonomètre et la vitesse à l'aide d'un radar. Des mesures de la température et de la vitesse du vent sont également effectuées car elles peuvent influer sur le bruit.

Un premier microphone est relié à un enregistreur qui servira à mesurer le bruit, il doit être placé à 7,5m du milieu de la voie sur laquelle se déplacent les véhicules dont on veut effectuer des mesures et à 1,2m de hauteur. Le radar est placé de manière à ce que la vitesse du véhicule soit mesurée au moment où le point central du véhicule passe devant le microphone. Un deuxième microphone permet d'enregistrer les commentaires de l'opérateur qui donne la catégorie du véhicule ainsi que sa vitesse qu'il lit sur l'écran d'affichage du radar. Avant de

commencer les mesures, il faut calibrer l'enregistreur par rapport à une source de bruit de référence. Ceci dans le but de rendre comparable les niveaux sonores issus de deux campagnes différentes.

Les deux figures 1.1 et 1.2 montrent un exemple d'installation sur une route deux fois deux voies, en vue de dessus (« plan ») et en vue de profil (« travers »), permettant de réaliser de telles campagnes.

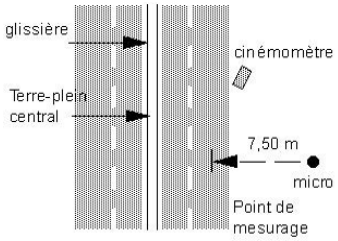


FIGURE 1.1 – Vue en plan

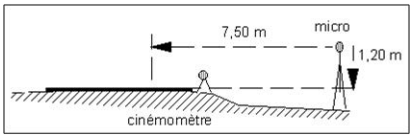


FIGURE 1.2 – Vue en travers

Les mesures doivent être effectuées uniquement sur le passage individuel de véhicules qui peuvent être clairement distingués d'un point de vue acoustique des autres véhicules se déplaçant sur la route.

Les critères suivants permettent de qualifier de valide un passage de véhicule :
- Immédiatement avant et après le passage d'un véhicule dont le niveau acoustique doit être mesuré, le niveau sonore doit être inférieur de 6 dB au minimum par rapport au niveau sonore maximal correspondant à celui du passage de ce véhicule. On s'assure également qu'au moment où le niveau sonore maximal est observé, le bruit collectif émis par la reste de la circulation sera inférieur de 10 dB au minimum par rapport au niveau maximal enregistré et qu'il aura une incidence négligeable sur la niveau mesuré.
- Lors de la sélection de véhicules pour le mesurage, il est recommandé de prêter une grande attention afin de s'assurer que le bruit émis par d'autres véhicules dépassant le véhicule cible ou se déplaçant sur l'autre voie n'exerce aucune influence sur le résultat du bruit mesuré. Dans de tels cas, il peut arriver que les bruits maximaux émis par le véhicule cible et par le reste de la circulation soient générés presque simultanément, si bien que les pics relevés ne sont pas identifiables. Ces mesurages ne doivent pas être pris en compte.
- De plus, les véhicules manifestant des caractéristiques acoustiques anormales ou atypiques, par exemple celles résultant d'un système d'échappement défaillant ou de craquements dans le corps du véhicule, ne doivent pas être pris en considération dans les mesures. Les véhicules dotés d'équipements auxiliaires émettant un son audible doivent également être écartés.

- Les niveaux sonores doivent être mesurés uniquement sur des véhicules se déplaçant à vitesse constante. Les véhicules particuliers jugés significativement écartés de l'axe central de la voie d'essai ne doivent pas être retenus pour l'analyse.

Une fois les mesures effectuées, l'importation et le dépouillement de celles-ci se font avec le logiciel *dBEuler*[1]. L'opérateur commence par créer une campagne. Il entre la température, qui a un impact sur le bruit, et il écoute en premier les calibrages de début et de fin de mesures afin de donner une base de niveau de bruit au logiciel. Ensuite l'opérateur écoute chaque enregistrement (qui correspond aux deux microphones). Pour chaque enregistrement, le logiciel affiche un spectre et l'opérateur valide ou non chaque passage de véhicules en respectant les règles décrites ci-dessus. Il faut aussi délimiter les passages au sein de chaque enregistrement car plusieurs passages de véhicules peuvent être effectués sur le même fichier. Si le passage est validé il reste à enregistrer les informations telles que la vitesse et la catégorie du véhicule. *dBEuler* permet la transformation des fichiers audio en niveau de bruit maximum, noté L_{Amax} (en décibels (dB)), pour chaque véhicule.

Un général, lors d'une campagne de mesure, on réalise des mesures sur environ 120 véhicules dans le but d'en avoir environ 100 après dépouillement puisque on va éliminer ceux qui ne respectent pas les critères décrits ci-dessus.

Avec *dBEuler* on peut, une fois tous les passages dépouillés, effectuer une analyse statistique. Un paramètre adimentionnel de vitesse est utilisé, on utilise comme variable explicative :

$$\mathrm{X} = \log_{10}\left(\frac{vitesse}{vitesse_{ref}}\right)$$

où *vitesse* est la vitesse du véhicule et $vitesse_{ref}$ est la vitesse de référence c'est-à-dire la limitation de vitesse de la route sur laquelle ont été effectuées les mesures.

On recherche des paramètres a et b tels que :

$$L_{Amax} = a \times \log_{10}(\frac{vitesse}{vitesse_{ref}}) + b \tag{1.1}$$

Le but est d'obtenir un niveau de bruit de référence pour une vitesse donnée. Cependant comme les mesures sont effectuées sur tous les véhicules du trafic (selon les critères ci-dessus) on observe des valeurs aberrantes. Le problème est que lors d'une régression linéaire simple, ces valeurs faussent l'analyse. Les valeurs aberrantes que l'on obtient sont souvent des véhicules dont le niveau de bruit est plus important que la moyenne. Elles peuvent être dues à des erreurs de mesure qui peuvent être de différentes nature. Par exemple lors des mesures sur le terrain, durant le passage d'un véhicule, il peut y avoir une erreur de mesure où de l'opérateur sur la vitesse ou la catégorie du véhicule. Il peut aussi y avoir un bruit parasite qui accroit donc le niveau de bruit mesuré. Une erreur peut aussi survenir au moment du dépouillement des données, lors de la saisie de la vitesse et de la catégorie du véhicule par l'opérateur.
Le logiciel *dBEuler* permet ensuite à l'opérateur de supprimer manuellement les valeurs qu'il juge aberrantes. *dBEuler* produit ensuite un rapport d'analyse, on obtient deux valeurs la pente et l'ordonnée à l'origine de la droite. Cette droite constitue le niveau de bruit de référence pour la route étudiée.

1.3 Objectifs du stage

La mesure du bruit de roulement au passage s'appuie sur un échantillon de couples, vitesse et niveau de bruit du véhicule, et mène à la détermination d'une droite de régression. L'objectif est de sélectionner une méthode d'estimation robuste parmi la famille des M-estimateurs afin de remplacer la méthode des moindres carrés classiques qui est trop sensible aux points aberrants. En effet comme expliqué ci-dessus, dans le logiciel *dBEuler* l'analyse statistique passe par une régression linéaire au sens des moindres carrés. Le problème est que cette méthode est sensible aux données aberrantes et que leurs suppressions nécessitent l'intervention de l'opérateur et de son avis pour déterminer quelles données sont aberrantes, ce qui n'est pas souhaitable car cela présente un risque de biais. Une procédure réalisée par Marie-Paule Ehrhart, stagiaire au LRPC en 2011, a permis d'analyser des données issues de différentes campagnes de mesure du bruit de roulement en estimant les paramètres de la droite en utilisant différentes techniques d'estimation alternatives à la régression linéaire[2], cependant il n'a pas été possible de hiérarchiser les méthodes. Une deuxième procédure réalisée par Makarim Ghazza, stagiaire au LRPC en 2012, a permis de calculer des intervalles de confiance sur une petite tranche de vitesse en utilisant différentes méthodes d'estimations [5]. Cependant les intervalles de confiance obtenus étaient trop larges et se chevauchaient selon les méthodes, on ne pouvait émettre aucune conclusion.

Le premier objectif du stage est de comparer des méthodes d'estimations. Dans ce but, une première étape du stage consistera à concevoir une campagne de mesure qui permette d'obtenir le niveau de bruit dans un intervalle de confiance étroit pour une plage de vitesse restreinte. La deuxième étape sera de réaliser les campagnes de mesures sur le terrain avec l'aide d'un technicien spécialiste de la mesure. Par la suite il faudra dépouiller les données obtenues sur la terrain. La fusion de plusieurs jours de mesures en une seule campagne nécessitera ensuite la vérification de l'homogénéité statistiques des campagnes. Une troisième étape permettra d'analyser les M-estimateurs et de les évaluer en situant leur bande de confiance des données brutes sur des plages de vitesses étroites aux extrémités de l'intervalle des vitesses observées.

Le deuxième objectif du stage est le classement automatique des véhicules, selon leurs catégories, à partir d'informations tels que leurs vitesses et leurs niveaux de bruit. Lors d'une campagne de mesures de véhicule léger et de train routier, on voudrait que l'opérateur ai besoin de rentrer dans le logiciel le minimum de catégorie des véhicules et qu'à partir de ces catégories les véhicules suivants soient attribués automatiquement à une catégorie. Pour cela on va utiliser l'analyse discriminante.

Durant mon stage j'ai utilisé la version 3.0.1 de *R* et la version 5.4.1 de *Scilab*.

Chapitre 2

Mise en place et réalisation de campagnes de mesures

2.1 Conception d'une campagne de mesure

2.1.1 Principe et but

Des intervalles de confiance, pour le niveau sonore noté L_{Amax} (en décibels (dB)), ont été calculés sur la tranche de vitesse $[v_{min}\,;\, v_{min} + 10]$ sur plusieurs campagnes de mesures par Makarim Ghazza[5], stagiaire au LRPC en 2012, cependant ils étaient trop larges. Notre objectif est d'obtenir un niveau sonore dans un intervalle de confiance étroit pour une plage de vitesse restreinte. Dans ce but, il va falloir concevoir une campagne de mesure afin d'obtenir un nombre de données suffisant pour atteindre notre objectif. Durant cette campagne, on va uniquement s'intéresser aux véhicules légers.

2.1.2 Calcul de la taille de l'échantillon

La taille de l'échantillon a une influence fondamentale sur la précision des estimations réalisées. Pour des raisons économiques il est nécessaire d'utiliser une taille d'échantillon la plus réduite possible tout en obtenant un taux de confiance suffisant. Le problème d'estimation de la taille d'un échantillon est en toute rigueur impossible à résoudre. En effet, sa résolution suppose connues des valeurs qui sont l'enjeu de l'expérience. Dans tous les cas, on est amené à faire des hypothèses sur l'ordre de grandeur des résultats que l'on veut obtenir. On n'est jamais assuré a priori d'avoir correctement dimensionné un échantillon. Ce n'est qu'a posteriori que l'on pourra vérifier l'adéquation des hypothèses sur les ordres de grandeur.

La fiabilité des données n'est pas absolue mais se situe plutôt dans un intervalle de confiance. Plus cet intervalle doit être petit ou plus la marge d'erreur doit être petite, plus la taille de l'échantillon devra être grande. Pour un intervalle de confiance, le taux de précision est inversement proportionnel au quadruple de la taille de l'échantillon.

On appelle généralement population ou population-mère l'ensemble des mesures faites auxquelles on s'intéresse dans le cadre d'une étude donnée. Un échantillon est une fraction de cette population.

Plusieurs paramètres doivent être pris en compte pour la détermination d'une taille minimum d'échantillon :
- la marge d'erreur que l'on se donne pour la grandeur que l'on veut estimer
- le taux de confiance que l'on souhaite garantir sur la mesure
- la proportion connue ou supposée de la population-mère présentant la caractéristique voulue

La marge d'erreur est l'erreur d'estimation qu'on est disposé à accepter. C'est un pourcentage de la valeur du paramètre étudié que nous fixons comme une erreur d'estimation raisonnable. Pour contrôler la précision nous devons aussi contrôler le niveau de confiance, c'est-à-dire fixer la probabilité que la marge d'erreur soit supérieure à celle fixée. Par exemple, un échantillon défini à un seuil de confiance de 95% avec une marge d'erreur de 3% permet d'extrapoler chaque résultat issu des mesures avec 5% de risque de se tromper de plus ou moins 3%.

Pour définir la taille de l'échantillon, on doit supposer connaître la proportion d'éléments de la population-mère sur laquelle porte l'étude. Afin d'avoir une approximation de cette proportion, on étudie des données obtenues lors d'une précédente campagne de mesures réalisée sur la D41 entre Stutzheim et Oberhausbergen où l'on va réaliser cette nouvelle campagne. On utilise l'intervalle de vitesse $[v_{min}\,;\, v_{min}{+}10]$ soit [58 ;68]. Ce jeu de données comporte 116 données dont 9 qui appartiennent à l'intervalle donné. On estime donc la proportion des mesures de la population qui se trouve dans l'intervalle $[v_{min}\,;\, v_{min} + 10]$ par :

$$p = \frac{9}{116} = 7.7\%$$

En ce qui concerne l'exhaustivité de l'échantillon, si la population mère est finie et que la taille de l'échantillon est supérieure à 1/7 (environ 14,3%) de la taille de la population-mère, on est en présence d'un échantillon exhaustif, la taille de l'échantillon devra être corrigée. En revanche si la population-mère est finie mais que la taille de l'échantillon est inférieure à 1/7 (environ 14,3%) de la taille de la population-mère, on ne considère pas l'échantillon comme étant exhaustif et on peut considérer la population-mère comme infinie.

Dans notre cas, la population mère est finie (116) mais la taille de l'échantillon (9) est inférieure à 1/7 de la population mère (17). On peut ainsi considérer que la population-mère est infinie et l'échantillon comme étant non exhaustif.

Le calcul de la taille d'échantillon se fait donc grâce à la formule :

$$n = t^2 \times \frac{p(1-p)}{e^2} \tag{2.1}$$

avec comme notation :
- p : proportion estimative des éléments de la population mère présentant la caractéristique
- e : marge d'erreur
- t : le coefficient de marge déduit du niveau de confiance (le coefficient de marge associé au taux de confiance 95% est 1.96, celui associé au taux de confiance 90% est 1.65)
- n : taille de l'échantillon
- N : taille de la population mère (ou population de référence, population d'origine) étudiée

En utilisant la proportion p calculée précédemment on obtient la taille d'échantillon pour différentes marges d'erreurs et différents niveaux de confiance :

- pour une marge d'erreur de 5% avec un taux de confiance de 95% on obtient n=109 ce qui correspond à N=1418
- pour une marge d'erreur de 5% avec un taux de confiance de 90% on obtient n=78 ce qui correspond à N=1005
- pour une marge d'erreur de 6% avec un taux de confiance de 95% on obtient n=76 ce qui correspond à N=985

Après discussion avec mes maîtres de stage, on a décidé de récolter 1000 données dans le but d'en obtenir environ 75-80 dans la tranche de vitesse $[v_{min}\,;\, v_{min} + 10]$.

2.2 Réalisation de la campagne de mesures

2.2.1 Récolte des données sur le terrain

Les campagnes de mesures qui suivent la procédure dite VI (Véhicules Isolés) dépendent du trafic. Lors d'une précédente campagne de mesure sur la D41, 120 données ont été récoltées en environ 1h30. Dépendant aussi des conditions météorologiques, on prévoit de réaliser 1000 mesures en deux jours similaires du point de vue climatique.

Afin d'obtenir des résultats de référence, on réalise aussi une campagne de mesure qui suit la procédure dite VM (Véhicules Maîtrisés). Pour cela un technicien du laboratoire circule sur la route à des vitesses prédéfinies avec différents véhicules. On utilise trois véhicules différents, une Clio II essence, une Clio III diesel et un Scenic essence, qui circulent à 50km/h, 70km/h et 90km/h.

Avec l'aide d'un technicien spécialiste de la mesure, on a réalisé cette campagne. Après avoir installé le radar et le microphone comme décrit dans la section 1.2, on enregistre le niveau sonore et la vitesse à chaque passage de véhicule conforme.

FIGURE 2.1 – Photo du site où les mesures ont été réalisées

2.2.2 Dépouillement des données

Après avoir enregistré 1000 passages de véhicules, il faut les importer et les dépouiller avec le logiciel *dBEuler* [1]. On écoute les enregistrements réalisés et le logiciel nous donne une interface graphique pour le découpage d'un enregistrement. A ce stade l'utilisateur peut alors écouter le commentaire et délimiter un passage à l'aide de la souris.

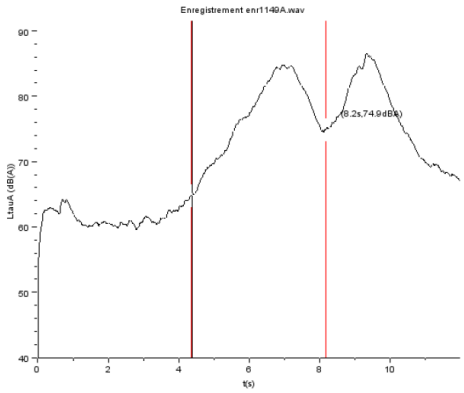


FIGURE 2.2 – Interface graphique pour le découpage d'un enregistrement en plusieurs passages

L'intervalle ainsi défini sert d'indication à *dBEuler* pour extraire un passage de l'enregistrement en cours de dépouillement, typiquement selon le critère "des 10 dB" vis-à-vis du bruit de fond et selon le critère "des 6dB" vis-à-vis des véhicules avant et après le passage considéré [S31-119][NFS31-119-2]. L'important étant que les marqueurs entourent un passage et un seul. On obtient ensuite une interface graphique de visualisation d'un passage, il reste à saisir les informations sur le véhicule à savoir la vitesse et la catégorie, ici se sont tous des véhicules légers.

Le technicien avec lequel j'ai effectué les mesures a réalisé la moitié du dépouillement et moi l'autre moitié. En respectant les critères pour la validation d'un passage on élimine des mesures et on se trouve finalement en possession de 882 couples de données, vitesse et niveau de bruit du véhicule.

2.3 Fusion des campagnes

2.3.1 Introduction

Les mesures ont été effectuées au même endroit avec les mêmes appareils de mesure par les mêmes opérateurs et ont été réalisées pendant deux jours similaires du point de vue météorologique. Cependant il y a des différences de température entre les deux jours, durant le premier jour la moyenne des températures était de 22° et durant le deuxième jour la moyenne des températures était de 24°. D'après la norme, on peut corriger cette différence de température en augmentant le niveau de bruit de 0.10dB par degrés en plus. Par conséquent, on augmente le niveau de bruit L_{Amax}, obtenu après dépouillement, du deuxième jour de 0.2dB. Une fois cette correction de température effectuée, afin de pouvoir fusionner ces deux jours de mesures en un seul jeu de données, il faut vérifier que le jour où les mesures ont été réalisées n'a pas d'influence sur les données.

2.3.2 Représentation graphique

On commence par représenter le niveau de bruit L_{Amax} en fonction du jour où on a effectué les mesures.

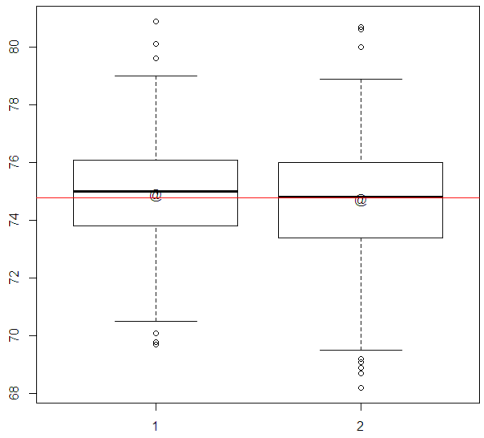


FIGURE 2.3 – Représentation du niveau de bruit en fonction du jour 1 et 2

On voit que la répartition des niveaux de bruit semble similaire pour les deux jour où les mesures ont été réalisées.

2.3.3 Test pour la fusion de la campagne

La variabilité existe toujours d'un jour à l'autre mais elle peut être due au hasard. L'analyse de la variance permet de montrer si la variabilité de la mesure du bruit entre les deux jours est due au hasard ou si le facteur jour intervient dans cette variabilité.

On est en présence d'un modèle aléatoire d'analyse de la variance [3] à deux facteurs avec répétition. Le facteur « Jour » est de nature qualitatif à deux modalités. La première modalité signifie que le véhicule a été enregistré le premier jour des mesures et la seconde que le véhicule a été enregistré le deuxième jour des mesures. Le facteur « Vitesse » peut être considéré de nature qualitatif en prenant comme première modalité une vitesse comprise entre 55 et 65 km/h, comme deuxième une vitesse comprise entre 65 et 75 km/h, comme troisième une vitesse comprise entre 75 et 85 km/h, comme quatrième une vitesse comprise entre 85 et 95 km/h et enfin une dernière avec une vitesse supérieure à 95km/h.
La réponse est le niveau de bruit L_{Amax} qui est de nature quantitative.

Le modèle statistique s'écrit :

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk}$$

où $i = 1, ..., I$, $j = 1, ..., J$ et $k = 1, ..., K$
où Y_{ijk} est la valeur prise par la réponse Y dans les conditions (A_i,B_j) lors du k-ème essai.

Il y a plusieurs conditions liées à ce type d'analyse.

Nous supposons que :
$\mathcal{L}\ (A_j) = \mathcal{N}\ (0,\ {\sigma^2}_A)$, pour tout $i, 1 \leq i \leq I$
$\mathcal{L}\ (B_j) = \mathcal{N}\ (0,\ {\sigma^2}_B)$, pour tout $j, 1 \leq j \leq J$
$\mathcal{L}\ ((AB)_{ij}) = \mathcal{N}(0,\ {\sigma^2}_{AB})$, pour tout $(i,j), 1 \leq i \leq I, 1 \leq j \leq J$

ainsi que l'indépendance des effets aléatoires :
- les effets aléatoires A_i sont indépendants
- les effets aléatoires B_j sont indépendants
- les effets aléatoires $(AB)_{ij}$ sont indépendants
- les effets aléatoires A_i et B_j sont indépendants
- les effets aléatoires A_i et $(AB)_{ij}$ sont indépendants
- les effets aléatoires B_j et $(AB)_{ij}$ sont indépendants

Nous postulons les hypothèses classiques de l'ANOVA (ANalysis Of VAriance) pour les variables d'erreurs ε_{ijk} :
- les erreurs sont indépendantes
- les erreurs ont même variance σ^2
- les erreurs sont de loi gaussienne

Nous ajoutons l'indépendance des effets aléatoires et des erreurs dues à ce type d'analyse :
- les effets aléatoires A_i et les erreurs ε_{ijk} sont indépendants
- les effets aléatoires B_j et les erreurs ε_{ijk} sont indépendants
- les effets aléatoires $(AB)_{ij}$ et les erreurs ε_{ijk} sont indépendants

On commence par vérifier les hypothèses classiques de l'ANOVA sur les variables d'erreurs ε_{ijk}. On veut tester :

H_0 : les résidus suivent une loi normale

contre

H_1 : les résidus ne suivent pas une loi normale

En réalisant ce test avec le logiciel R, on obtient une p-valeur de 0.042 qui est inférieure à 0.05. Nous décidons, au seuil $\alpha = 5\%$, de refuser l'hypothèse nulle H_0. Par conséquent, nous pouvons dire que les résidus ne suivent pas une loi normale. Le risque associé à cette décision est un risque de première espèce qui vaut 5%.

Cette hypothèse n'étant pas vérifiée, on va utiliser une variante non paramétrique de l'analyse de la variance, l'ANOVA de Kruskal-Wallis.

On veut tester :

H_0 : il n'y a pas d'effet du facteur aléatoire « Jour »

contre

H_1 : il y a un effet du facteur aléatoire « Jour »

En réalisant ce test avec le logiciel R, on obtient une p-valeur de 0.061 qui est supérieure à 0.05. Nous décidons de ne pas refuser l'hypothèse nulle H_0. Par conséquent, nous n'avons pas réussi à mettre en évidence d'effet du facteur aléatoire « Jour ».

Le jour n'ayant pas d'influence sur la réponse on peut donc fusionner les mesures obtenues durant les deux jours de mesures, on obtient alors un jeu de données contenant 882 couples de données, vitesse et niveau de bruit du véhicule.

On peut également tester :

H_0 : il n'y a pas d'effet du facteur aléatoire « Vitesse »

contre

H_1 : il y a un effet du facteur aléatoire « Vitesse »

En réalisant ce test avec le logiciel R, on obtient une p-valeur $<$ 2.2e-16 qui est inférieure à 0.05. Nous décidons, au seuil $\alpha = 5\%$, de refuser l'hypothèse nulle H_0. Par conséquent, nous pouvons dire qu'il y a un effet significatif du facteur aléatoire « Vitesse ». Le risque associé à cette décision est un risque de première espèce qui vaut 5%.

2.3.4 Calcul du taux de confiance et de la marge d'erreur

Comme expliqué dans la section 2.1.2, on n'est jamais assuré a priori d'avoir correctement dimensionné la taille de l'échantillon. L'échantillon obtenu, on peut maintenant calculer le taux de confiance et la marge d'erreur que l'on a réellement en fixant l'un des deux paramètres.

Sur ce jeu de données, l'intervalle de vitesse [v_{min} ; v_{min}+10] est [55 ;65]. Sur les 882 données, 55 appartiennent à cet intervalle. On estime la proportion p par :

$$p = \frac{55}{882} = 6.2\%$$

On obtient donc une proportion p de données qui appartiennent à l'intervalle de vitesse [v_{min} ; v_{min}+10] moins importante que l'approximation qu'on en a faite pour le calcul de la taille de l'échantillon.

Avec la proportion p qu'on vient de calculer et $n = 55$ on obtient différents résultats à l'aide de la formule (2.1) :

- En fixant la marge d'erreur à 5%, on obtient t = 1.53 qui correspond à un taux de confiance de 88%
- En fixant la marge d'erreur à 6%, on obtient t = 1.84 qui correspond à un taux de confiance de 94%
- En fixant maintenant le taux de confiance à 90%, on obtient une marge d'erreur de 5.4%

En choisissant le dernier résultat, on peut dire que chaque résultat issu des mesures sera extrapolé avec un risque de 10% de se tromper de plus ou moins 5.4%.

Chapitre 3

Evaluation des M-estimateurs

3.1 But et principe

D'après ce qui a déjà été fait lors des précédents stages, des méthodes tels que l'ACP et l'ACP robuste ont été écartées. On va ainsi uniquement s'intéresser aux M-estimateurs Geman et McClure et « Smooth Exponential Family » (SEF) (cf Annexe A). On va analyser ces M-estimateurs et les évaluer en situant leur intervalle de confiance par rapport à un intervalle de confiance des données brutes sur des plages de vitesse étroites. L'intervalle de vitesse choisi, pour le calcul d'un intervalle de confiance du niveau de bruit, est $[v_{min}\,;\,v_{min}+10]$ soit $[55\,;\,65]$. On a un échantillon de 882 données dont 55 dans l'intervalle de vitesse $[55\,;\,65]$ km/h.

Les véhicules maîtrisés vont servir de référence pour le calcul d'un niveau moyen de bruit à 50 km/h. On va commencer par regarder comment se situent les véhicules maîtrisés par rapport aux véhicules isolés.

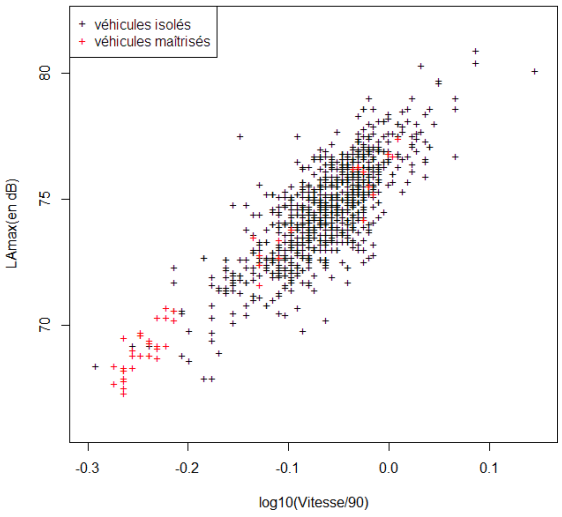


FIGURE 3.1 – Représentation graphique des VI et VM

On remarque sur la figure 3.1 que les véhicules maîtrisés complètent assez bien le nuage de points des véhicules isolés.

En calculant les paramètres de droite des différentes méthodes pour les véhicules isolés et pour les véhicules maîtrisés séparément on obtient :

Méthode d'estimation	Pente	Ordonnée à l'origine
Régression linéaire	30.79	76.50
M-estimation Geman et McClure	30.90	76.50
M-estimation SEF $\alpha = 0.5$	31.00	76.52
M-estimation SEF $\alpha = 0.0$	30.92	76.52
M-estimation SEF α = -0.5	31.00	76.51

TABLE 3.1 – Paramètres de droites pour les 882 véhicules isolés

Méthode d'estimation	Pente	Ordonnée à l'origine
Régression linéaire	31.11	76.61
M-estimation Geman et McClure	31.85	76.77
M-estimation SEF $\alpha = 0.5$	31.69	76.73
M-estimation SEF $\alpha = 0.0$	31.76	76.75
M-estimation SEF α = -0.5	31.69	76.75

TABLE 3.2 – Paramètres de droites pour les 45 véhicules maîtrisés

En comparant les tableaux 3.1 et 3.2, on peut remarquer que les paramètres de droites trouvés pour les véhicules isolés et pour les véhicules maîtrisés sont assez proches.

De plus pour les véhicules isolés les paramètres de droite sont très proches entre les différentes méthodes. On peut faire les mêmes observations pour les véhicules maîtrisés.

En traçant les droites obtenues respectivement pour les véhicules isolés et pour les véhicules maîtrisés selon différentes méthodes, dans l'intervalle de vitesse [55 ; 65], on obtient :

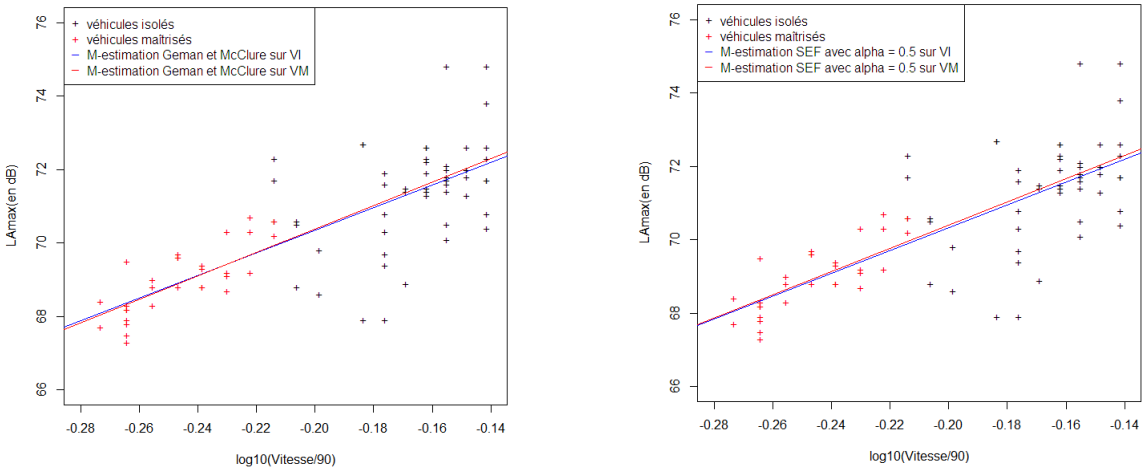


FIGURE 3.2 – M-Estimation de Geman et McClure et SEF avec $\alpha = 0.5$

On remarque que, pour les différentes méthodes, la droite trouvée à partir des véhicules isolés et la droite trouvée à partir des véhicules maîtrisés sont très proches l'une de l'autre sur l'intervalle de vitesse [55 ; 65] km/h. Le fait de considérer les véhicules maîtrisés comme référence pour le calcul d'un niveau moyen de bruit à 50 km/h semble raisonnable.

3.2 Différentes représentations graphiques des données

On commence par représenter les véhicules isolés et les droites obtenues par les différentes méthodes d'estimation sur tout l'intervalle de vitesse ainsi que sur l'intervalle de vitesse réduit à [55 ; 65] km/h.

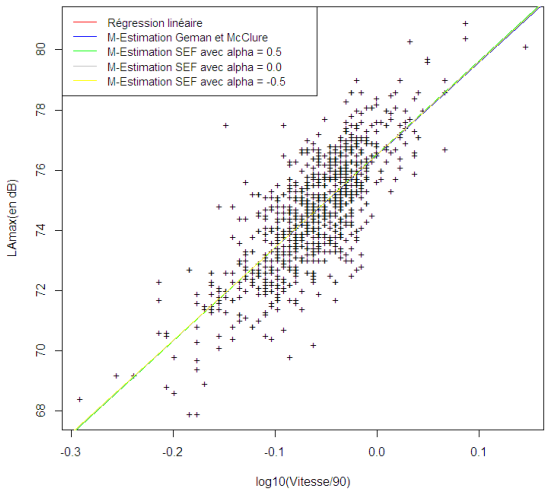


FIGURE 3.3 – Véhicules isolés sur toute la tranche de vitesse

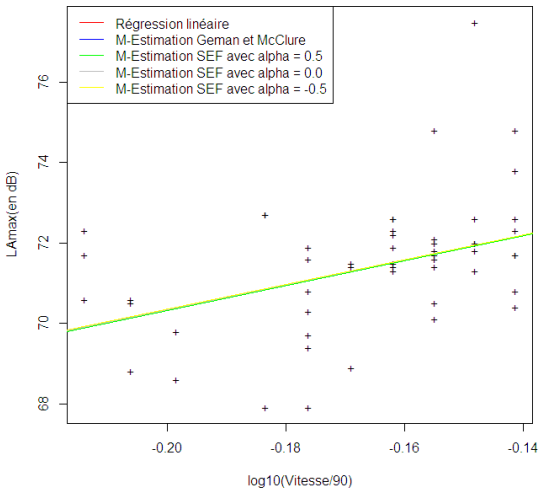


FIGURE 3.4 – Véhicules isolés sur la tranche de vitesse [55 ; 65]

On observe sur les figures 3.3 et 3.4 que toutes les droites, obtenues par différentes méthodes d'estimation, se confondent.

Cela peut s'expliquer par le fait qu'on a un grand nombre de données par rapport au nombre de données que l'on obtient en générale lors d'une campagne de mesure et qu'on observe pratiquement pas de valeurs aberrantes. Afin de me ramener à un jeu de données de 100 véhicules, que l'on obtient typiquement lors d'une campagne de mesure, je réalise de l'échantillonnage parmi les 882 véhicules isolés.

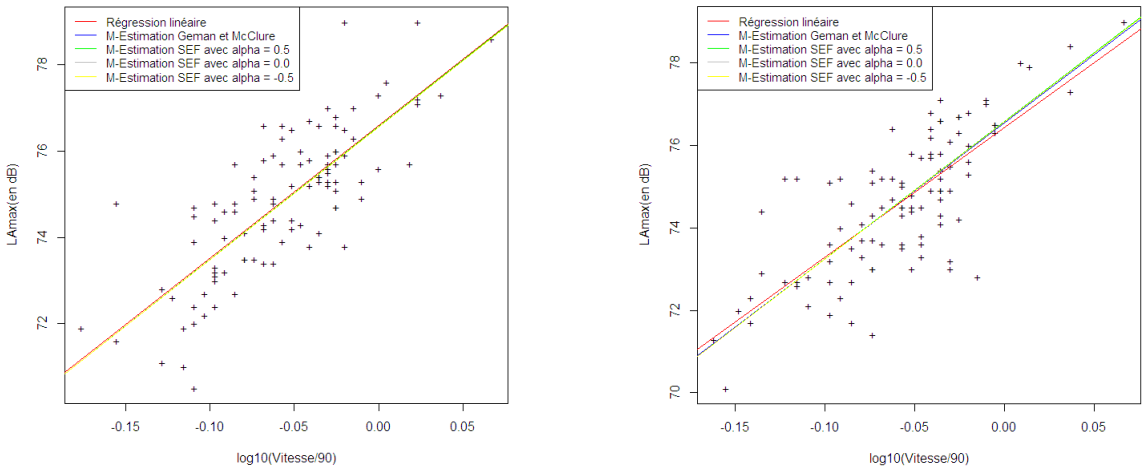


FIGURE 3.5 – Deux échantillons de 100 véhicules isolés avec les différentes droites

On observe sur les figures 3.5, qui sont deux exemples d'échantillonnage de 100 données parmi les 882, que même sur un nombre de données plus restreint, excepté la droite obtenue par régression linéaire, toutes les droites se confondent.

Marie-Paule Ehrhart [2] avait tracé des droites sur d'autres jeux de données et elle obtenait des droites distinctes selon les méthodes utilisées. J'ai aussi refait des représentations graphiques sur d'autres jeux de données et des différences entre les diverses méthodes étaient observables.

On peut expliquer ces variations entre les jeux de données par la qualité du jeu de données obtenu lors de la campagne réalisée durant le stage du point de vue de la répartition de ces observations et de la non présence d'observations jugées aberrantes. En effet sur ce jeu de données on observe au plus un écart de 4dB entre l'une des droites obtenues par M-estimation et une donnée alors que sur d'autres jeux de données, étudiés les années précédentes, on observait des écarts pouvant aller jusqu'à plus de 10dB. Comme on a réalisé une campagne uniquement sur des véhicules légers, il n'y a pas eu d'erreur de la part de l'opérateur sur la catégorie des véhicules lors de la campagne de mesures et encore moins lors de l'attribution d'une catégorie aux véhicules lors du dépouillement.

3.3 Intervalles de confiance sur une tranche de vitesse

L'intervalle de confiance représente la zone d'incertitude quant à l'estimation d'un paramètre. L'utilisation de cet intervalle tient compte du fait qu'une campagne fournit une estimation d'un paramètre parmi les nombreuses estimations qui seraient possibles si la campagne était répétée plusieurs fois. Ainsi, un intervalle de confiance à 90% indique une probabilité de 90% que l'intervalle de confiance calculé à partir d'une campagne donnée contienne la vraie valeur du paramètre. Un intervalle de confiance étroit autour de l'estimation ponctuelle indique une estimation plus précise qu'un intervalle de confiance large.

On ne connaît pas la vraie valeur du niveau de bruit pour une vitesse donnée sur une route donnée. On calcule alors la valeur de l'estimateur du paramètre étudié à partir des mesures effectuées sur les véhicules d'un échantillon. En raison des fluctuations d'échantillonnage, on sait que si l'on avait calculé l'estimateur à partir d'un autre échantillon, on aurait probablement trouvé une valeur différente. De ce fait, les résultats du calcul de cet estimateur doivent être présentés sous la forme d'un intervalle de confiance, construit à partir de l'estimateur produit.

Afin de pouvoir calculer des intervalles de confiance pour le niveau de bruit L_{Amax}, suivant différentes méthodes, on utilise la moyenne énergétique définie par :

$$\overline{L_{Amax}} = 10 \log_{10} \frac{\sum_{i=1}^{n} (10^{\left(\frac{l_{amaxi}}{10}\right)})}{n}$$

Pour utiliser un intervalle de confiance de type normal il faut vérifier la normalité de l'échantillon utilisé. On va réaliser un test de Shapiro-Wilk sur les L_{Amax} des 55 données se situant dans la tranche de vitesse [55 ; 65] km/h.

On veut tester :

H_0 : les L_{Amax} suivent une loi normale

contre

H_1 : les L_{Amax} ne suivent pas une loi normale

En réalisant ce test avec le logiciel R, on obtient une p-valeur de 0.067 qui est supérieure à 0.05. Nous décidons de ne pas refuser l'hypothèse nulle H_0. Par conséquent, nous pouvons dire que les L_{Amax}, dans la tranche de vitesse [55 ; 65] km/h, suivent une loi normale.

On peut donc utiliser un intervalle de confiance de type normal. En calculant la moyenne énergétique $\overline{L_{Amax}}$ et l'écart-type σ des L_{Amax_i}, on obtient l'intervalle de confiance pour le niveau de bruit L_{Amax} :

$$[\overline{L_{Amax}} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \overline{L_{Amax}} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \tag{3.1}$$

avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre 1-$\frac{\alpha}{2}$ de la loi normale.

On choisit ici α=5%, on aura un intervalle de confiance à 95%. Cela signifie qu'on a une probabilité de 95% que l'intervalle de confiance qu'on va calculer à partir de cette campagne contienne la vraie valeur du paramètre.

On obtient, après le calcul des intervalles de confiance pour L_{Amax} selon les différentes méthodes d'estimation, les résultats suivants :

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne Inf	Borne Sup	Amplitude
Données brutes	71.83	71.37	72.28	0.91
Régression linéaire	71.43	71.25	71.61	0.36
M-estimation Geman et McClure	71.43	71.25	71.61	0.36
M-estimation SEF $\alpha = 0.5$	71.41	71.23	71.59	0.36
M-estimation SEF $\alpha = 0.0$	71.42	71.24	71.60	0.36
M-estimation SEF α = -0.5	71.43	71.25	71.61	0.36

TABLE 3.3 – Intervalles de confiance pour L_{Amax} sur la tranche de vitesse [55 ; 65]

On remarque une nette diminution de l'amplitude de l'intervalle de confiance en utilisant une méthode d'estimation par rapport aux données brutes. Les intervalles de confiance sont assez étroits et se ressemblent fortement suivant les différentes méthodes d'estimation. Avec un nombre de données importantes, les intervalles de confiance sont quasiment identiques pour toutes les méthodes.

3.3.1 Méthode du Bootstrap

La méthode du Bootstrap est une méthode statistique basée sur un rééchantillonnage des données à partir d'une distribution estimée de la population réelle (inconnue). Cette distribution est généralement construite à partir de l'échantillon observé. On calcule une statistique (ici $\overline{L_{Amax}}$) à partir d'un échantillon de n individus issus de la population d'étude.

Afin de construire un intervalle de confiance par la méthode du Bootstrap on commence par réaliser une boucle pour i allant de 1 à N. Dans cette boucle on réalise un échantillonnage aléatoire avec remise de n individus dans l'échantillon puis on calcul la statistique $\overline{L_{Amax}}$ sur le i$^{\text{ième}}$ échantillon Bootstrap. On calcul la fonction de distribution à partir des N valeurs de la statistique. Pour terminer on calcul la statistique qui nous intéresse.

Pour chaque échantillon simulé i, on obtient une valeur $\overline{L_{Amax_i}}$. La distribution empirique de ces $\overline{L_{Amax_i}}$ est une approximation de la distribution théorique de $\overline{L_{Amax}}$. On peut alors prendre la moyenne empirique des $\overline{L_{Amax_i}}$ comme une nouvelle estimation de $\overline{L_{Amax}}$.

Cette méthode étant basée sur le rééchantillonnage, les intervalles de confiance obtenus varient d'une exécution à l'autre du code, en voici deux exemples :

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne Inf	Borne Sup	Amplitude
Données brutes	71.83	71.18	72.37	1.19
Régression linéaire	71.43	71.26	71.61	0.35
M-estimation Geman et McClure	71.43	71.26	71.61	0.35
M-estimation SEF $\alpha = 0.5$	71.42	71.26	71.59	0.33
M-estimation SEF $\alpha = 0.0$	71.42	71.26	71.61	0.35
M-estimation SEF α = -0.5	71.42	71.26	71.60	0.34

TABLE 3.4 – Intervalles de confiance Bootstrap pour L_{Amax} sur la tranche de vitesse [55 ; 65]

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne Inf	Borne Sup	Amplitude
Données brutes	71.83	71.25	72.36	1.11
Régression linéaire	71.43	71.27	71.59	0.32
M-estimation Geman et McClure	71.43	71.27	71.61	0.34
M-estimation SEF $\alpha = 0.5$	71.42	71.26	71.59	0.33
M-estimation SEF $\alpha = 0.0$	71.42	71.26	71.60	0.34
M-estimation SEF α = -0.5	71.42	71.26	71.59	0.33

TABLE 3.5 – Intervalles de confiance Bootstrap pour L_{Amax} sur la tranche de vitesse [55 ; 65]

En comparant les tableaux 3.4 et 3.5 on remarque que les intervalles de confiance sont quasiment identiques d'une exécution à l'autre du code. Dans la même exécution du code, les intervalles de confiance obtenus par la méthode du Bootstrap sont plus ou moins équivalents suivant les différentes méthodes. On remarque aussi que les intervalles obtenus par la méthode du Bootstrap sont aussi équivalents avec ceux obtenus au tableau 3.3. On obtient cependant une amplitude des intervalles de confiance légèrement plus restreinte avec cette méthode.

3.3.2 Échantillonnage dans le jeu de données

But de l'échantillonnage

En général, lors d'une campagne de mesure on récolte 100 données. On va donc faire de l'échantillonnage de 100 données parmi les 882 afin de se ramener à un nombre de données que l'on a habituellement et on va pouvoir comparer les résultats obtenus avec ceux obtenus précédemment sur le jeu de données complet.

On réalise également de l'échantillonnage sur 50 données seulement de manière à voir si l'estimation robuste ne permettrait pas de réduire la durée de présence sur le terrain lors d'une campagne de mesure.

Résultats de l'échantillonnage sur 100 données

Comme c'est de l'échantillonnage, les résultats varient d'une exécution à l'autre du code. En voici trois exemples :

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	71.52	70.54	72.50	1.96
Régression linéaire	71.31	70.99	71.64	0.65
M-Estimation Geman McClure	71.30	70.97	71.63	0.66
M-Estimation SEF $\alpha = 0.5$	71.29	70.96	71.62	0.66
M-Estimation SEF $\alpha = 0.0$	71.29	70.96	71.62	0.66
M-Estimation SEF α = -0.5	71.30	70.97	71.63	0.66

TABLE 3.6 – Intervalles de confiance pour L_{Amax} sur la tranche de vitesse [55 ; 65] avec échantillonnage sur 100 données

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	71.25	70.73	71.76	1.03
Régression linéaire	71.49	71.02	71.96	0.95
M-Estimation Geman McClure	71.36	70.87	71.85	0.98
M-Estimation SEF $\alpha = 0.5$	71.37	70.88	71.86	0.98
M-Estimation SEF $\alpha = 0.0$	71.36	70.87	71.85	0.98
M-Estimation SEF α = -0.5	71.36	70.87	71.85	0.98

TABLE 3.7 – Intervalles de confiance pour L_{Amax} sur la tranche de vitesse [55 ; 65] avec échantillonnage sur 100 données

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	71.12	69.82	72.23	2.41
Régression linéaire	70.97	70.04	71.98	1.94
M-Estimation Geman McClure	70.90	69.93	71.87	1.94
M-Estimation SEF $\alpha = 0.5$	70.94	69.97	71.91	1.94
M-Estimation SEF $\alpha = 0.0$	70.92	69.95	71.89	1.94
M-Estimation SEF α = -0.5	70.91	69.94	71.88	1.94

TABLE 3.8 – Intervalles de confiance pour L_{Amax} sur la tranche de vitesse [55 ; 65] avec échantillonnage sur 100 données

L'amplitude des intervalles de confiance varie beaucoup d'une exécution à l'autre. Cependant les intervalles de confiance, obtenus lors d'un échantillonnage, de plus petites amplitudes sont contenus dans les intervalles de confiance, obtenus lors d'un autre échantillonnage, de plus grandes amplitudes.

Dans le même échantillonnage, les intervalles de confiance sont quasiment identiques selon les différentes méthodes d'estimation et l'amplitude de l'intervalle également.

En comparant ces intervalles de confiance avec ceux obtenus au tableau 3.3, on remarque que les intervalles de confiance obtenus à partir du jeu de données complet sont de plus petites amplitudes que ceux obtenus en faisant de l'échantillonnage sur 100 données. De plus les intervalles de confiance obtenus à partir de toutes les données sont contenus dans les intervalles de confiance obtenus à partir de 100 données, pour différents échantillonnage.

Résultats de l'échantillonnage sur 100 données avec la méthode du Bootstrap

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	71.41	70.69	72.82	2.13
Regression linéaire	71.29	71.01	71.59	0.58
M-Estimation Geman McClure	71.43	71.16	71.72	0.56
M-Estimation SEF $\alpha = 0.5$	71.23	71.16	71.72	0.56
M-Estimation SEF $\alpha = 0.0$	71.43	71.16	71.72	0.56
M-Estimation SEF α = -0.5	71.43	71.16	71.67	0.51

TABLE 3.9 – Intervalles de confiance Bootstrap pour L_{Amax} sur la tranche de vitesse [55 ;65] avec échantillonnage sur 100 données

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	71.25	70.48	72.01	1.53
Regression linéaire	71.42	70.93	71.90	0.97
M-Estimation Geman McClure	71.36	70.83	71.82	0.99
M-Estimation SEF $\alpha = 0.5$	71.36	70.87	71.86	0.99
M-Estimation SEF $\alpha = 0.0$	71.36	70.86	71.85	0.99
M-Estimation SEF α = -0.5	71.36	70.86	71.88	1.02

TABLE 3.10 – Intervalles de confiance Bootstrap pour L_{Amax} sur la tranche de vitesse [55 ;65] avec échantillonnage sur 100 données

L'amplitude des intervalles de confiance varie beaucoup d'une exécution à l'autre cependant la méthode du Bootstrap permet de réduire l'amplitude de l'intervalle, qui est au maximum de environ 1dB, alors que sans cette méthode l'amplitude de l'intervalle peut atteindre 2dB. On observe également que les intervalles de plus petites amplitudes sont contenus dans les intervalles de plus grandes amplitudes.

En comparant ces intervalles de confiance avec ceux obtenus au tableau 3.3, 3.4 et 3.5, on remarque que les intervalles de confiance obtenus à partir du jeu de données complet sont de plus petites amplitudes que ceux obtenus en faisant de l'échantillonnage sur 100 données. De plus les intervalles de confiance obtenus à partir de toutes les données sont contenus dans les intervalles de confiance obtenus à partir de 100 données, pour différents échantillonnage.

Résultats de l'échantillonnage sur 50 données

Dans la perspective de rester moins longtemps sur le terrain on effectue de l'échantillonnage sur 50 données. En voici trois exemples :

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	70.64	69.85	71.43	1.58
Régression linéaire	71.15	70.72	71.58	0.86
M-Estimation Geman McClure	71.14	70.72	71.57	0.85
M-Estimation SEF 0.5	71.16	70.73	71.59	0.85
M-Estimation SEF 0.0	71.16	70.73	71.58	0.85
M-Estimation SEF -0.5	71.15	70.73	71.58	0.85

TABLE 3.11 – Intervalles de confiance pour L_{Amax} sur la tranche de vitesse [55 ; 65] avec échantillonnage sur 50 données

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	70.44	68.91	71.97	3.06
Régression linéaire	70.32	69.57	71.08	1.51
M-Estimation Geman McClure	70.32	69.55	71.09	1.54
M-Estimation SEF 0.5	70.36	69.60	71.12	1.53
M-Estimation SEF 0.0	70.35	69.59	71.12	1.53
M-Estimation SEF -0.5	70.34	69.57	71.11	1.53

TABLE 3.12 – Intervalles de confiance pour L_{Amax} sur la tranche de vitesse [55 ; 65] avec échantillonnage sur 50 données

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne sup	Amplitude
Données brutes	70.86	66.25	75.46	9.21
Régression linéaire	70.70	69.30	72.10	2.81
M-Estimation Geman McClure	70.85	69.49	72.21	2.72
M-Estimation SEF 0.5	70.81	69.45	72.19	2.74
M-Estimation SEF 0.0	70.82	69.44	72.19	2.75
M-Estimation SEF -0.5	70.83	69.46	72.20	2.74

TABLE 3.13 – Intervalles de confiance pour L_{Amax} sur la tranche de vitesse [55 ; 65] avec échantillonnage sur 50 données

Contrairement à l'étude avec 100 données, les amplitudes des intervalles de confiance sont assez conséquentes. De plus les intervalles de confiance, obtenus lors d'un échantillonnage, de plus petites amplitudes ne sont pas contenus dans les intervalles de confiance, obtenus lors d'un autre échantillonnage, de plus grandes amplitudes.

En comparant ces intervalles de confiance avec ceux obtenus au tableau 3.3, on remarque que les intervalles de confiance obtenus à partir de toutes les données ne sont pas contenus dans les intervalles de confiance obtenus à partir de 50 données, pour différents échantillonnage.

Récolter seulement 50 données sur le terrain ne semble pas raisonnable.

3.4 Calcul du biais

On va calculer le biais de la moyenne énergétique des L_{Amax} obtenue sur les véhicules isolés par rapport à la moyenne énergétique des véhicules maîtrisés. Pour cela on va calculer la différence entre la valeur de son espérance et la valeur de la variable qu'il est censé estimer. On obtient cette valeur de référence à partir des véhicules maîtrisés sur lesquels on calcul la moyenne énergétique des L_{Amax}, noté $\overline{L_{Amax_{VM}}}$.

$$Biais(\overline{L_{Amax}}) = E(\overline{L_{Amax}}) - \overline{L_{Amax_{VM}}} \tag{3.2}$$

On réalise le calcul du biais pour différents intervalles choisis pour le calcul de la référence à partir des véhicules maîtrisés et pour différents intervalles choisis pour les véhicules isolés. En voici deux exemples :

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Ecart type LAmax	Biais
Données brutes	71.83	1.68	2.76
Régression linéaire	71.43	0.66	2.36
M-Estimation Geman McClure	71.43	0.66	2.36
M-Estimation SEF $\alpha = 0.5$	71.41	0.66	2.34
M-Estimation SEF $\alpha = 0.0$	71.42	0.66	2.35
M-Estimation SEF α = -0.5	71.43	0.66	2.36

TABLE 3.14 – Biais avec comme référence VM [48,55] avec données [55-65]

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Ecart type LAmax	Biais
Données brutes	70.68	1.53	0.81
Régression linéaire	70.64	0.48	0.78
M-Estimation Geman McClure	70.65	0.48	0.78
M-Estimation SEF $\alpha = 0.5$	70.63	0.48	0.76
M-Estimation SEF $\alpha = 0.0$	70.64	0.48	0.77
M-Estimation SEF α = -0.5	70.64	0.48	0.78

TABLE 3.15 – Biais avec comme référence VM [53,55] avec données [55-60]

En comparant tous les résultats obtenus on voit que quel que soit l'intervalle de vitesse pris pour les véhicules maîtrisés qui sert de référence et quelle que soit la tranche de vitesse prise pour les véhicules isolés, on obtient quasiment les mêmes résultats pour toutes les méthodes. La méthode avec laquelle on obtient néanmoins le biais minimum est la méthode M-Estimation SEF avec $\alpha = 0.5$.

De plus, on remarque que plus l'intervalle de vitesse pris pour les véhicules maîtrisés, qui sert de référence, et l'intervalle de vitesse pris pour les véhicules isolés sont proches et plus le biais diminue.

3.5 Introduction de valeurs aberrantes

D'après la figure 3.5, on a remarqué que même en faisant de l'échantillonnage sur 100 données les différents M-estimateurs restent équivalents.

On va maintenant introduire des valeurs aberrantes dans le jeu de données et effectué de l'échantillonnage sur 100 données.

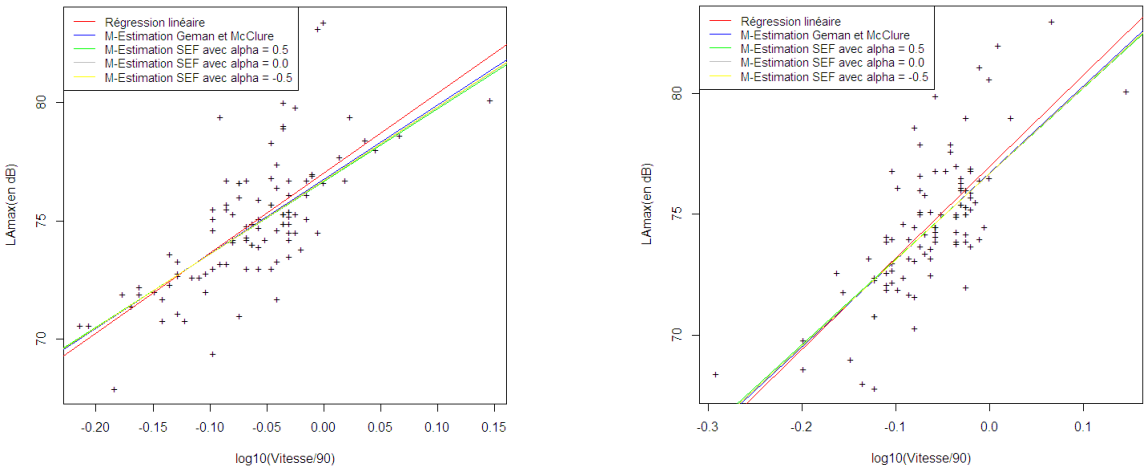


FIGURE 3.6 – Deux échantillons de 100 véhicules isolés avec des valeurs aberrantes

Sur la figure 3.6 on remarque qu'excepté la droite obtenue par régression linéaire, qui se détache mieux des autres par l'introduction de données aberrantes, toutes les autres droites sont confondues.

Malgré l'introduction de valeurs aberrantes dans le jeu de données, les différentes méthodes de M-estimations restent équivalentes.

3.6 Conclusions

Sur ce jeu de données, on peut dire que toutes les méthodes de M-estimations testées sont équivalentes. En effet ce jeu de données présente une bonne répartition des données et ne présente pas de données aberrantes. Cela s'explique par le fait qu'on a réalisé une campagne de mesures sur des véhicules légers exclusivement. Sur le terrain lorsque l'opérateur enregistre un véhicule il ne peut pas se tromper sur sa catégorie et de même lors du dépouillement.

On a donc introduit des données aberrantes dans le jeu de données. Cela a permit de mettre en évidence une différence entre la régression linéaire et les M-estimateurs. De plus cela justifie une nouvelle fois le fait que les méthodes de M-estimations sont équivalentes sur ce jeu de données.

En prenant en considération ce qui a été fait les années précédentes sur des autres jeux de données et ce qui vient d'être fait sur ce nouveau jeu de données, on peut dire que toutes les méthodes de M-estimations sont équivalentes.

Le fait d'avoir calculé des intervalles de confiance, en prenant différents échantillons de 100 données, a permit de montrer que les intervalles de confiance obtenus pour 100 données englobent celui obtenu à partir du jeu de données complet. Cela permet de conforter l'idée que récolter 100 données, lors de campagnes de mesures, permet tout de même une bonne approximation du niveau de bruit à basse vitesse.

En revanche l'échantillonnage sur 50 données n'est pas probant. Récolter seulement 50 données lors de campagnes de mesures ne semble pas raisonnable.

Chapitre 4

Classification automatique des véhicules

Lors du dépouillement d'une campagne de mesures de bruit de roulement, l'opérateur doit rentrer manuellement dans le logiciel la catégorie du véhicule en fonction de ce qui a été dit sur l'enregistrement audio. On examine ici la possibilité de déterminer automatiquement la catégorie des véhicules. L'idée est d'initialiser les deux classes véhicules légers et trains routiers en saisissant manuellement la catégorie pour quelques passages au début du dépouillement.

Dans cette partie j'ai exploité les données issues de quatre campagnes différentes : Haguenau2008, RN59, Rothau2008 et Erstein. Le premier jeu de données Haguenau2008 contient 169 données dont 114 véhicules légers et 55 trains routiers. Le second, RN59, contient 197 données dont 135 véhicules légers et 62 trains routiers. Le troisième, Rothau2008, contient 139 données dont 85 véhicules légers et 54 trains routiers. Pour finir, le jeu de données Ertein contient 223 données dont 138 véhicules légers et 85 trains routiers.

Ces jeux de données contiennent comme information sur chaque donnée la Vitesse, le L_{Amax} et le spectre (18 données, noté $L1$,...,$L18$). L'objectif est de déterminer la catégorie des véhicules à partir des variables Vitesse et L_{Amax} et de rajouter éventuellement des valeurs du spectre ou de partir d'autres combinaisons de variables.

4.1 Analyse discriminante linéaire

L'analyse discriminante linéaire (cf Annexe B) s'effectue sur des données quantitatives dont les individus sont affectés à des classes. Elle tente de déterminer la contribution des variables qui expliquent l'appartenance à des groupes. On compare plusieurs groupes sur plusieurs variables pour déterminer s'ils différent et pour comprendre la nature de ces différences. Cette méthode permet aussi d'affecter de nouveaux individus aux groupes. C'est à la fois une méthode descriptive et prédictive.

Dans le premier cas, l'objectif est de produire un système de représentation synthétique où l'on distinguerait au mieux les groupes, en fournissant les éléments d'interprétation permettant de comprendre ce qui les réunit ou les différencie. Dans le second cas, on cherche à produire un système de classement qui permet d'affecter un groupe à un individu selon ses caractéristiques.

L'analyse discriminante linéaire cumule des qualités intéressantes, elle est rapide sur des grandes bases, robuste et stable même appliquée sur des petites bases.

4.1.1 Vérification de l'hypothèse de normalité

L'utilisation de l'analyse discriminante linéaire nécessite la normalité des données. Pour vérifier cette hypothèse on utilise un test de multinormalité, qui est une modification du test de Shapiro-Wilk unidimensionnel, à l'aide du package "mvnormtest" sous *R*.

On commence par réaliser ce test sur l'ensemble des variables du jeux de données, on veut tester :

H_0 : les données suivent une loi normale multidimensionnelle

contre

H_1 : les données ne suivent pas une loi normale multidimensionnelle

Jeu de données	W	p-value
Haguenau2008	0.87	9.86e-11
RN59	0.77	5.515e-15
Rothau2008	0.87	1.102e-10
Erstein	0.88	1.298e-10

TABLE 4.1 – Test de normalité des différents jeux de données avec toutes les variables

Pour ces tests, les p-values sont toutes inférieures à 0.05, on décide au seuil α=5% de rejeter l'hypothèse nulle H_0. Par conséquent on peut dire qu'aucun des jeux de données, contenant toutes les variables, ne suit une loi normale multidimensionnelle.

On réalise maintenant le test seulement sur les variables $Vitesse$ et L_{Amax} des jeux de données, on veut tester :

H_0 : les données suivent une loi normale multidimensionnelle

contre

H_1 : les données ne suivent pas une loi normale multidimensionnelle

Jeu de données	W	p-value
Haguenau2008	0.98	0.05116
RN59	0.97	0.05207
Rothau2008	0.98	0.2691
Erstein	0.96	7.679e-06

TABLE 4.2 – Test de normalité des différents jeux de données avec les variables $Vitesse$ et L_{Amax}

Pour les trois premiers jeux de données, les p-values sont supérieures à 0.05 donc le test n'est pas significatif. L'hypothèse nulle H_0 ne peut pas être rejetée, c'est-à-dire qu'on décide

que l'hypothèse de normalité multidimensionnelle est satisfaite pour les trois jeux de données, contenant uniquement les variables $Vitesse$ et L_{Amax}, Haguenau2008, RN59 et Rothau2008.

En revanche pour le jeu de données Erstein, la p-value est inférieure à 0.05 donc le test est significatif. On décide de rejeter l'hypothèse nulle H_0 et de décider que l'hypothèse alternative H_1 est vraie , c'est-à-dire que l'hypothèse de normalité multidimensionnelle n'est pas satisfaite pour le jeux de données Erstein, contenant uniquement les variables $Vitesse$ et L_{Amax}.

4.1.2 Représentations graphiques des différents jeux de données

On commence par représenter graphiquement les véhicules légers et les trains routiers des différents jeux de données, à partir des variables $Vitesse$ et L_{Amax}.

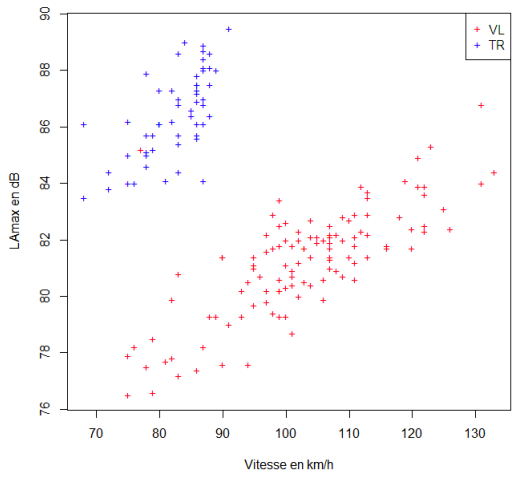


FIGURE 4.1 – Haguenau2008

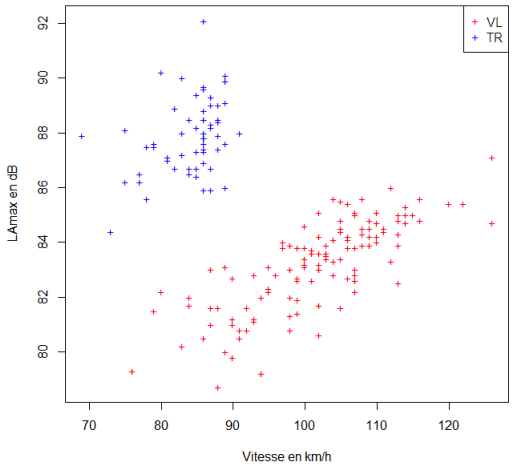


FIGURE 4.2 – RN59

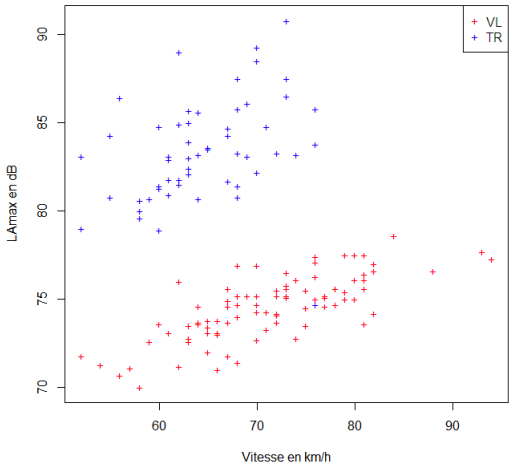


FIGURE 4.3 – Rothau2008

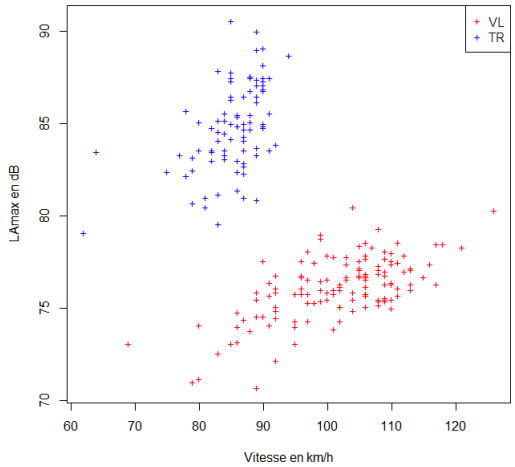


FIGURE 4.4 – Erstein

En ce qui concerne le jeu de données Haguenau2008 (sur la figure 4.1), on observe un point représentant un véhicule léger dans le nuage de points associé aux trains routiers. Ce point correspond à l'observation qui a pour $Vitesse$ 77 et pour L_{Amax} 85,2. En ce qui concerne le jeu de données Rothau2008 (sur la figure 4.3), on observe un point représentant un train routier dans le nuage de points associé aux véhicules légers. Ce point correspond à l'observation qui a pour $Vitesse$ 76 et pour L_{Amax} 74,5. On peut donc s'interroger sur la catégorie de ces véhicules qui semblent être mal classés par l'opérateur.

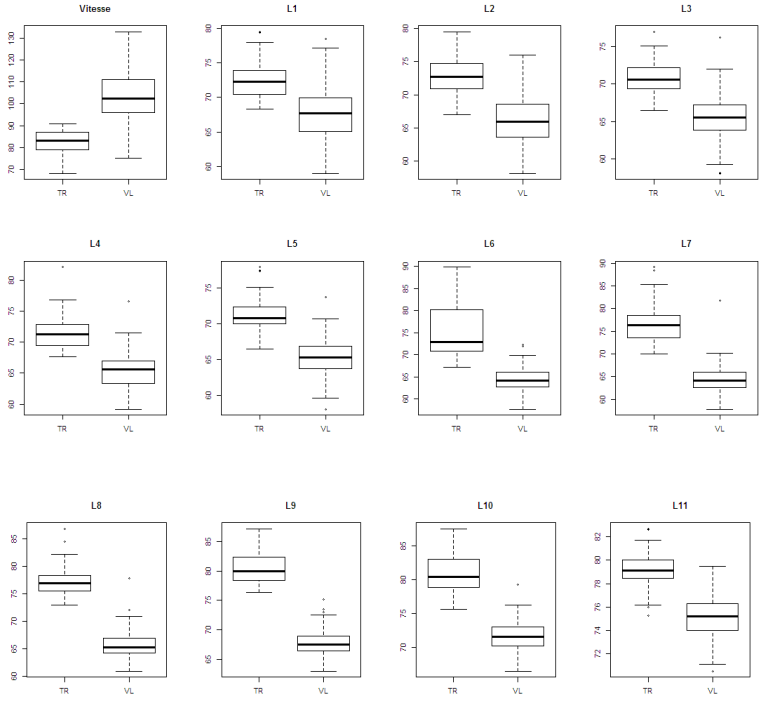
On remarque, sur les figures 4.1 à 4.4, que l'on différencie assez facilement les véhicules légers et les trains routiers.

4.2 Analyse discriminante descriptive

Il s'agit d'identifier les variables ou combinaisons de variables qui expliquent au mieux l'appartenance aux deux différents groupes.

4.2.1 Boxplot

Pour trouver les variables les plus discriminantes, une succession de boîtes à moustaches est très informative. Ces représentations concerne le jeu de données Haguenau2008. Les représentations ont aussi été faites pour les autres jeux de données, on en tire les mêmes conclusions.



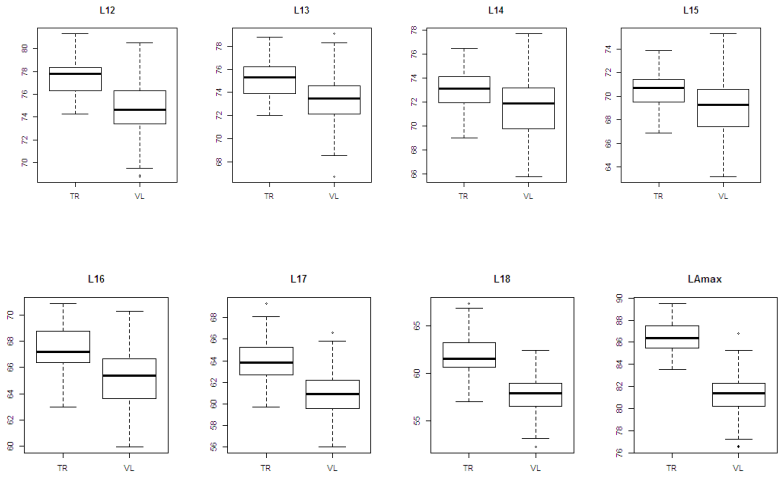
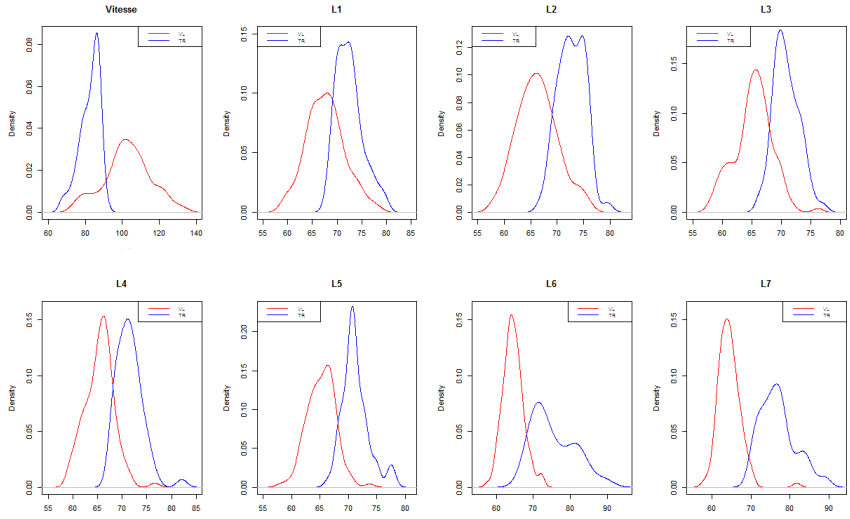


FIGURE 4.5 – Boxplot des différentes variables selon la catégorie des véhicules

On voit, sur la figure 4.5, que les variables $Vitesse, L7, L8, L9, L10$ et L_{Amax} différencient le mieux les catégories.

4.2.2 Pouvoir discriminant et estimation de densités

Partant de l'idée de représentation par espèce, on peut reproduire les densités de chacune des variables quantitatives pour les deux catégories. Ces densités correspondent au jeu de données Haguenau2008. Les densités ont aussi été représentées pour les autres jeux de données, on en tire les mêmes conclusions.



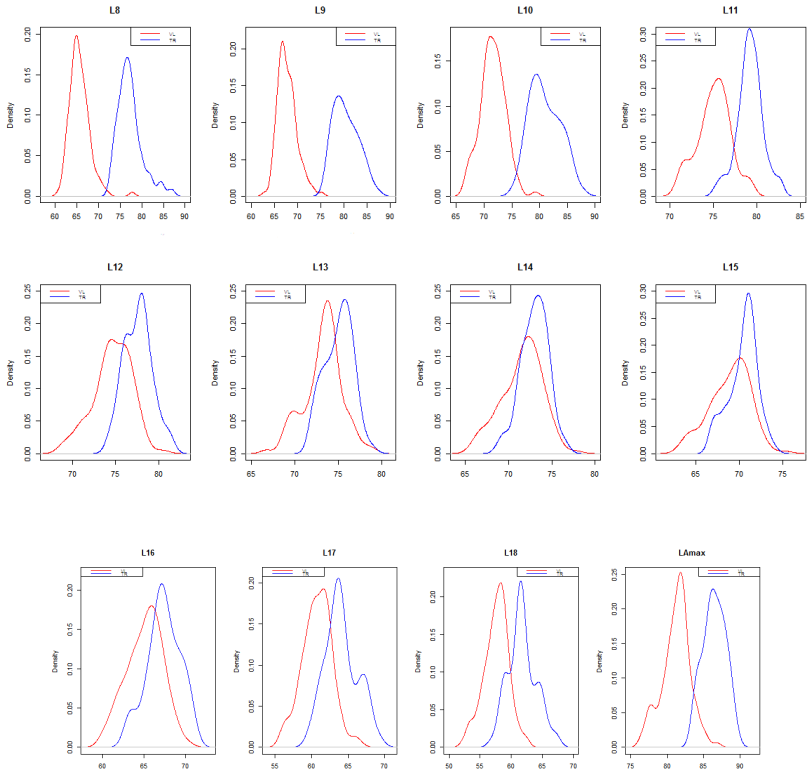


FIGURE 4.6 – Densités estimées de la distribution des différentes variables selon la catégorie des véhicules

On retrouve, sur la figure 4.6, les observations faites à partir des boxplots c'est-à-dire que les variables $Vitesse, L7, L8, L9, L10$ et L_{Amax} différencient le mieux les catégories. On peut également comparer ces graphiques à ceux de la figure 4.7 qui suit, qui donnent les distributions de chaque variable et fournissent ainsi une nouvelle façon d'appréhender le pouvoir discriminant d'une variable.

Ces graphiques fournissent plusieurs observations. Tout d'abord, un mélange de populations différentes ne se caractérise pas forcément par une densité multimodale (cf les distributions des $L1, L2, L12$ et $L16$). Par ailleurs, la comparaison des graphiques 4.6 et 4.7 semble montrer que le caractère discriminant d'une variable est lié à la relation entre la dispersion totale de cette variable et celle observée dans chacune des deux catégories. Ce qui nous amène au paragraphe suivant.

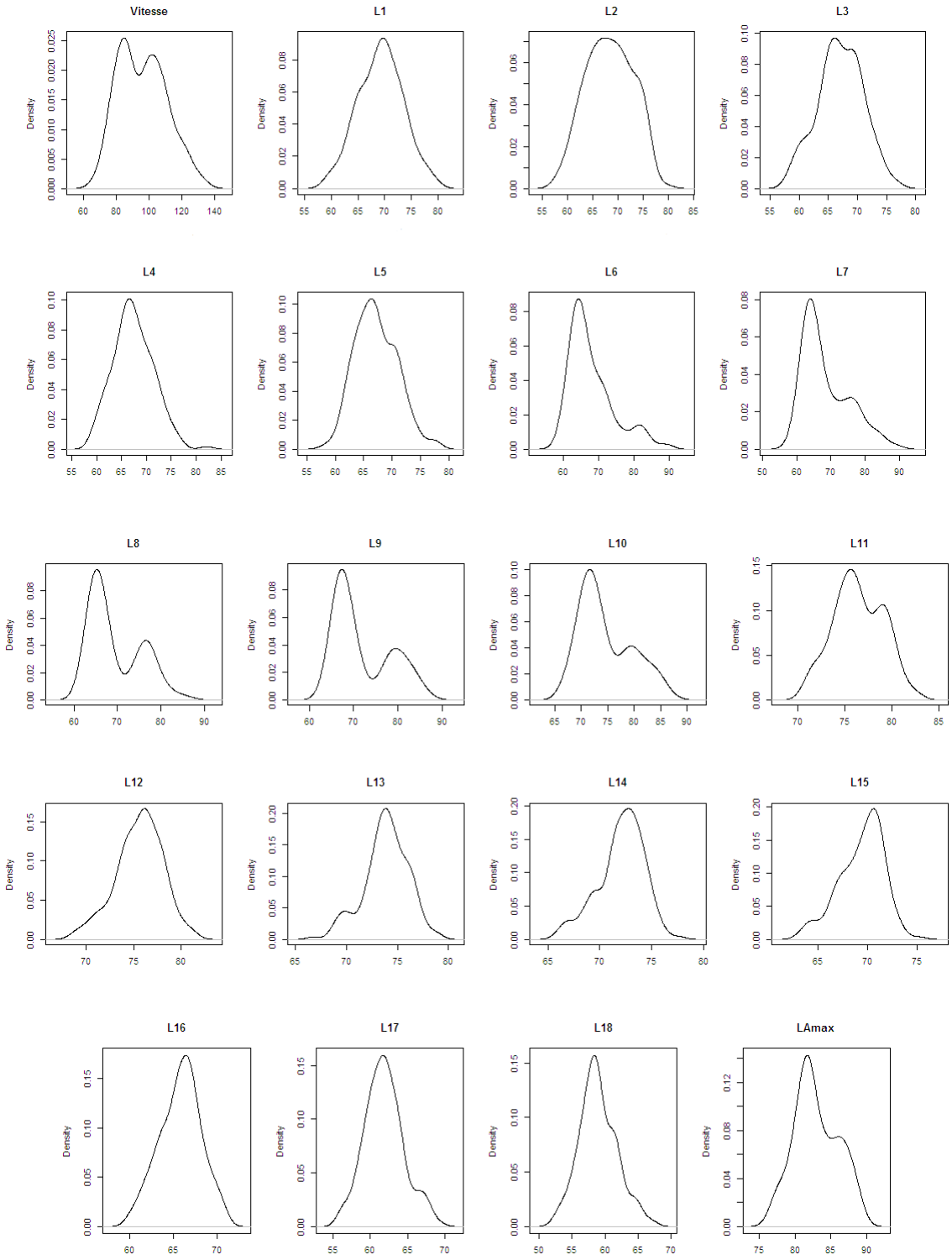


FIGURE 4.7 – Densités estimées de la distribution des différentes variables

4.2.3 Pouvoir discriminant et variance

Afin d'exalter au mieux les différences entre les deux classes, on recherche le fait que la variance entre les classes (inter-classes) soit maximale et que la variance à l'intérieur des classes (intra-classes) soit minimale pour que l'étendue dans les classes soit délimitée.

σ désigne l'écart-type de la variable observées, σ_{VL} désigne l'écart-type de la variable observée pour la catégorie véhicule léger (VL) et σ_{TR} l'écart-type de la variable observée pour la catégorie train routier (TR). La dernière colonne représente la moyenne des deux rapports σ_{VL}/σ et σ_{TR}/σ.

Ainsi plus la *Moyenne des rapports* est faible, plus la variable est utile pour discriminer les catégories.

Variable	σ	σ_{VL}	σ_{TR}	σ_{VL}/σ	σ_{TR}/σ	Moyenne des rapports
Vitesse	14.50	12.75	5.36	0.88	0.37	0.62
L1	4.20	3.90	2.69	0.93	0.64	0.78
L2	4.59	3.82	2.53	0.83	0.55	0.69
L3	3.90	3.20	2.18	0.82	0.56	0.69
L4	4.02	2.92	2.66	0.73	0.66	0.69
L5	3.72	2.48	2.36	0.67	0.64	0.65
L6	6.49	2.71	5.82	0.42	0.90	0.66
L7	6.87	2.96	4.62	0.43	0.67	0.55
L8	6.02	2.36	2.88	0.39	0.48	0.44
L9	6.39	2.08	2.68	0.33	0.42	0.37
L10	5.03	2.22	2.76	0.44	0.55	0.49
L11	2.66	1.94	1.47	0.73	0.55	0.64
L12	2.46	2.23	1.61	0.91	0.65	0.78
L13	2.19	2.21	1.61	1.01	0.74	0.87
L14	2.22	2.34	1.55	1.06	0.70	0.88
L15	2.25	2.34	1.64	1.04	0.73	0.89
L16	2.36	2.17	1.98	0.92	0.84	0.88
L17	2.61	2.04	2.21	0.78	0.85	0.82
L18	2.85	1.98	2.25	0.70	0.79	0.74
LAmax	3.04	1.93	1.52	0.64	0.50	0.56

TABLE 4.3 – Pouvoir discriminant de chaque variable sur le jeu de données Haguenau2008

Ainsi les variables $Vitesse, L7, L8, L9, L10$ et L_{Amax} discriminent le mieux les deux catégories. Ceci est cohérent avec les observations faites précédemment.

4.3 Analyse discriminante prédictive

D'après l'analyse discriminante descriptive, les variables $Vitesse$, $L7$, $L8$, $L9$, $L10$ et L_{Amax} discriminent le mieux les deux catégories. D'après le paragraphe 4.1.1, seuls les jeux de données Haguenau2008, RN59 et Rothau2008 contenant uniquement les variables $Vitesse$ et L_{Amax} suivent une loi normale multidimensionnelle. En réalisant d'autres tests de normalité avec les variables $Vitesse$, $L7$, $L8$, $L9$, $L10$ et L_{Amax}, aucune autre combinaison de variables ne semble suivre une loi normale.

J'ai tout de même réalisé l'analyse discriminante prédictive à partir de différentes variables et les meilleurs résultats s'obtiennent avec les variables $Vitesse$ et L_{Amax}.

Les résultats concernant l'analyse discriminante prédictive ont été obtenus à l'aide des packages "MASS", "klaR" et "cluster" sous R.

4.3.1 Analyse discriminante prédictive à partir des variables Vitesse et LAmax

Le but est de prédire la catégorie des véhicules à l'aide de deux informations, $Vitesse$ et L_{Amax}, que l'on a sur ces véhicules.

On voudrait notamment savoir combien il faut d'éléments dans l'échantillon de référence pour réaliser la classification des autres véhicules et quelle est l'erreur par rapport au classement effectué par l'opérateur.

L'objectif est d'initialiser les deux classes, véhicule léger et train routier, en saisissant manuellement la catégorie pour quelques passages au début du dépouillement afin de pouvoir classer automatiquement les passages suivants selon leurs catégories.

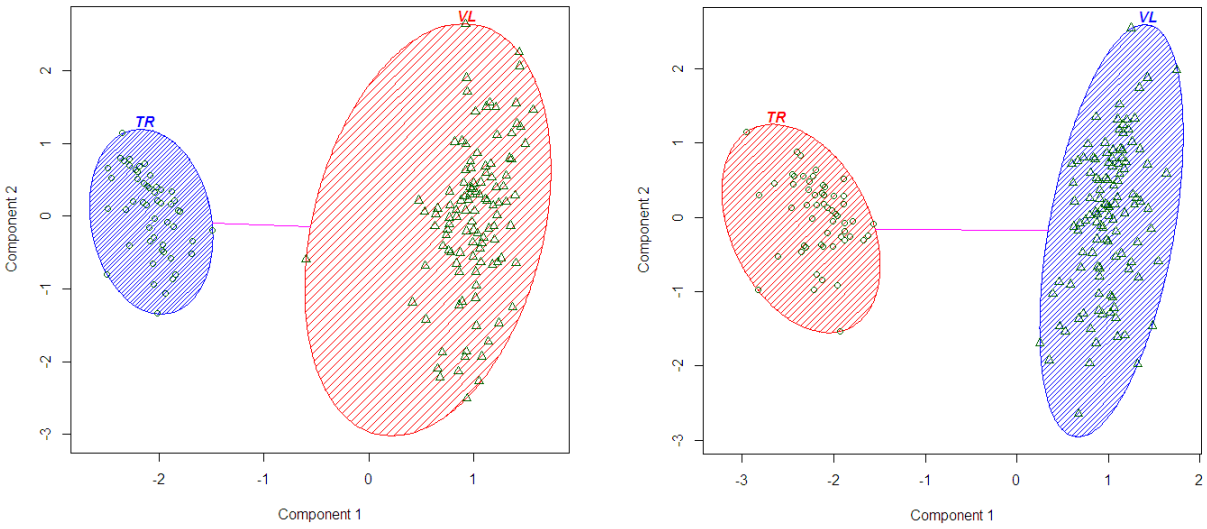


FIGURE 4.8 – Représentation des groupes pour les jeux de données Haguenau2008 et RN59

Sur la figure 4.8, on remarque que l'on distingue facilement les deux catégories.

On va réaliser la classification sur nos quatre jeux de données.

Je commence par appliquer cette méthode sur les jeux de données tel quel en choisissant comme échantillon de référence les premières données du fichier en prenant au minimum un véhicule de chaque catégorie.

Pour le jeu de données Haguenau2008, j'applique l'algorithme avec comme échantillon de référence les 4 premières données (3 VL et 1 TR). On obtient une seule erreur de classement par rapport à celui réalisé par l'opérateur. La donnée qui est mal classée est l'observation 122, l'analyse discriminante linéaire lui attribue la catégorie train routier alors que l'opérateur la catégorie véhicule léger. En prenant comme échantillon de départ les 121 premières données, l'observation 122 sera tout de même attribuée à la mauvaise catégorie. Cette donnée a comme Vitesse 77 et comme LAmax 85,2. On avait observé cette donnée sur la figure 4.1.

Pour le jeu de données RN59, j'applique l'algorithme avec comme échantillon de référence les 17 premières données (16 VL et 1TR). On obtient aucune erreur de classement par rapport à celui réalisé par l'opérateur.

Pour le jeu de données Rothau2008, j'applique l'algorithme avec comme échantillon de référence les 4 premières données (3 VL et 1 TR). On obtient une seule erreur de classement par rapport à celui réalisé par l'opérateur. La donnée qui est mal classée est l'observation 61, l'analyse discriminante linéaire lui attribue la catégorie véhicule léger alors que l'opérateur la catégorie train routier. En prenant comme échantillon de départ les 60 premières données, l'observation 61 sera tout de même attribuée à la mauvaise catégorie. Cette observation a comme Vitesse 76 et comme LAmax 74,7. On avait observé cette donnée sur la figure 4.3.

Pour le jeu de données Erstein, j'applique l'algorithme avec comme échantillon de base les 5 premières données (1 VL et 4 TR). On obtient aucune erreur de classement par rapport à celui réalisé par l'opérateur.

Ces premiers résultats sont très bon, l'analyse discriminante effectue le même classement que l'opérateur et corrige même ses erreurs de classement, à partir de très peu de données pour lesquelles on donne la catégorie.

4.3.2 Échantillonnage et erreur de classement

On va maintenant effectuer de l'échantillonnage sur les jeux de données afin d'appliquer l'analyse discriminante linéaire avec différents échantillons de référence de différentes tailles.

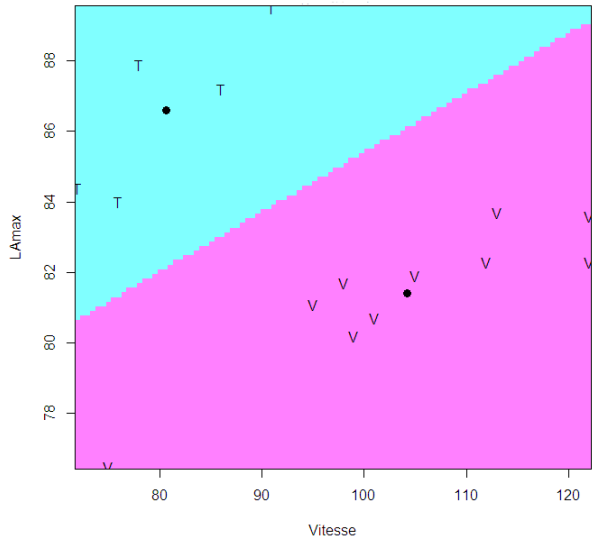


FIGURE 4.9 – Initialisation de la classification lors d'un échantillonnage à partir d'un échantillon de référence de 10 données sur le jeu de données Haguenau2008

J'ai écrit un script sous *R* pour calculer l'erreur de classement entre l'opérateur et l'analyse discriminante, à partir d'un grand nombre d'échantillonnage dans les jeux de données.

Jeu de données Haguenau2008

Pourcentage d'erreur par rapport au classement réalisée par l'opérateur	Pourcentage de chance d'avoir un classement contenant ce pourcentage d'erreur		
0.6%	88%	97%	99%
1%	7%	1.5%	1%
2%	2%	0.5%	0%
4%	1.3%	0.5%	0%
5%	0.7%	0.5%	0%
7%	1%	0%	0%
	10 données	15 données	20 données

TABLE 4.4 – Erreur dans le classement des véhicules pour le jeu de données Haguenau2008

Le 0.6% d'erreur obtenu par rapport au classement réalisé par l'opérateur est la donnée observée sur la figure 4.1 qui semble être une erreur de classement de la part de l'opérateur.

Jeu de données RN59

Pourcentage d'erreur par rapport au classement réalisée par l'opérateur	Pourcentage de chance d'avoir un classement contenant ce pourcentage d'erreur		
0%	80%	91%	93%
0.5%	7%	5%	6%
1%	9%	3%	1%
2%	2%	1%	0%
4%	1%	0%	0%
5%	1%	0%	0%
	10 données	15 données	20 données

TABLE 4.5 – Erreur dans le classement des véhicules pour le jeu de données RN59

Jeu de données Rothau2008

Pourcentage d'erreur par rapport au classement réalisée par l'opérateur	Pourcentage de chance d'avoir un classement contenant ce pourcentage d'erreur		
0.8%	86.5%	96.8%	98%
2%	8%	2%	2%
3%	1%	0.4%	0%
5%	1.5%	0.4%	0%
7%	2%	0.4%	0%
9%	1%	0%	0%
	10 données	15 données	20 données

TABLE 4.6 – Erreur dans le classement des véhicules pour le jeu de données Rothau2008

Le 0.8% d'erreur obtenu par rapport au classement réalisé par l'opérateur est la donnée observée sur la figure 4.3 qui semble être une erreur de classement de la part de l'opérateur.

Jeu de données Erstein

Pourcentage d'erreur par rapport au classement réalisée par l'opérateur	Pourcentage de chance d'avoir un classement contenant ce pourcentage d'erreur		
0%	65.4%	85%	90%
0.5%	23%	13%	9%
1%	10%	1.5%	1%
2%	1%	0.5%	0%
4%	0.6%	0%	0%
	10 données	15 données	20 données

TABLE 4.7 – Erreur dans le classement des véhicules pour le jeu de données Erstein

On remarque, à partir des tableaux 4.4 à 4.7, que plus on a de données dans l'échantillon de référence, plus on a de chance de réaliser le même classement des véhicules que l'opérateur. De plus, avec un plus grand nombre de données dans l'échantillon de départ lorsqu'il y a une erreur entre le classement réalisé par l'analyse discriminante prédictive et celui réalisé par l'opérateur, le pourcentage d'erreur entre les deux classements diminue. J'ai également effectué l'analyse discriminante sur ces jeux de données à partir de plus grands échantillons de référence, cependant on n'observe pas une réelle amélioration par rapport à ceux que l'on obtient avec 20 données dans l'échantillon de référence.

On observe également que les pourcentages de chance d'obtenir le même classement que celui réalisé par l'opérateur varie suivant le jeu de données. En effet on obtient de très bon résultat pour les jeux de données Haguenau2008 et Rothau2008 et d'un peu moins bon résultat pour les jeux de données RN59 et Erstein.

4.3.3 Classification évolutive

Une autre technique consiste à fixer un échantillon de référence au départ et de le faire évoluer au fur et à mesure. A partir de notre échantillon de référence on attribue à une nouvelle donnée une catégorie et cette nouvelle donnée s'intègre dans l'échantillon de référence. L'échantillon de référence évolue au fur et à mesure que l'on attribue une nouvelle donnée à une catégorie.

Jeu de données Haguenau2008

En appliquant l'algorithme avec un échantillon de référence de 10 données, on a 98% de chance de faire 0.6% d'erreur dans le classement (la donnée (77,85.2) classé en VL par l'opérateur et TR par la LDA, qui semble être un erreur de classement de la part de l'opérateur) par rapport à celui réalisé par l'opérateur. On a 2% de chance de ne faire aucune erreur (quand la donnée (77,85.2) se trouve dans l'échantillon de référence de départ).

Jeu de données RN59

En appliquant l'algorithme avec un échantillon de référence de 10 données, on a 98% de chance de ne pas faire d'erreur dans le classement des véhicules par rapport à celui réalisé par l'opérateur. On a 1.2% de risque de ne pas réaliser le même classement que l'opérateur avec 0.5% d'erreur entre les deux classements et on a 0.8% de risque de ne pas réaliser le même classement que l'opérateur avec 1% d'erreur entre les deux classements.

Jeu de données Rothau2008

En appliquant l'algorithme avec un échantillon de référence de 10 données, on a 99% de chance de faire 0.8% d'erreur dans le classement (la donnée (76,74.7) classé en TR par l'opérateur et VL par la LDA, qui semble être un erreur de classement de la part de l'opérateur) par rapport à celui réalisé par l'opérateur. On a 1% de risque de ne pas réaliser le même classement que l'opérateur avec 2% d'erreur entre les deux classements.

Jeu de données Erstein

En appliquant l'algorithme avec un échantillon de référence de 10 données, on a 97% de chance de ne pas faire d'erreur dans le classement des véhicules par rapport à celui réalisé par l'opérateur. On a 3% de risque de ne pas réaliser le même classement que l'opérateur avec 0.5% de différence entre les deux classements.

L'analyse discriminante prédictive avec classification évolutive permet d'obtenir de très bon résultat. En effet sur ces quatre jeux de données on a au plus un risque de 3% de ne pas réaliser le même classement des véhicules que l'opérateur avec au plus 2% d'erreur entre les deux classements.

En comparant ces résultats avec ceux obtenus dans les tableaux 4.4 à 4.7, on observe que pour les jeux de données RN59 et Erstein les pourcentages de chance d'obtenir le même classement que celui réalisé par l'opérateur sont nettement améliorés.

La classification évolutive permet d'obtenir de meilleurs résultats avec moins de données dans l'échantillon de référence.

4.4 Conclusions

L'analyse discriminante linéaire donne de bon résultat dans l'attribution de catégories au véhicules. Avec comme variables explicatives $Vitesse$ et L_{Amax}, en réalisant une classification évolutive, on obtient quasiment le même classement des véhicules que l'opérateur. Sur ces quatre jeux de données on a au plus un risque de 3% de ne pas réaliser le même classement des véhicules que l'opérateur avec au plus 2% d'erreur entre les deux classements.

De plus cette méthode permet de pallier le problème de mauvais classement effectué par l'opérateur qui peut survenir lors de la réalisation de la campagne sur le terrain ou lors du dépouillement des données.

Chapitre 5

Synthèse

Dans le but d'étudier les méthodes de M-estimations sur un jeu de données plus important que ce que l'on obtient généralement lors d'une campagne de mesure, et afin de calculer des intervalles de confiance plus précis, il a fallu dimensionner la taille adéquate de l'échantillon et réaliser les mesures sur le terrain.

L'analyse des M-estimateurs sur ce jeu de données a permis de mettre en évidence le fait que les différentes méthodes de M-estimations, Geman et McClure et « Smooth Exponential Family », sont équivalentes. En effet, même en réalisant de l'échantillonnage afin de travailler sur différents échantillons du jeu de données plus restreint, ces méthodes restent équivalentes.

Du fait qu'on ait réalisé cette campagne de mesure uniquement sur des véhicules légers, le jeu de données présente une assez bonne répartition des données avec très peu de données aberrantes.

Le calcul d' intervalles de confiance a permit de conforter l'idée que récolter 100 données, lors de campagnes de mesures, permet tout de même une bonne approximation du niveau de bruit à basse vitesse.

En revanche récolter seulement 50 données lors de campagnes de mesures ne semble pas raisonnable.

En étudiant les possibilités de classement des véhicules selon leurs catégories à partir de l'analyse discriminante, on a pu mettre en avant le fait qu'à partir des variables $Vitesse$ et L_{Amax} des véhicules et des catégories des 10 premiers véhicules en utilisant une classification évolutive, on a 97% de chance de réaliser le même classement que l'opérateur.

Dans la perspective d'automatiser les campagnes de mesures, l'analyse discriminante linéaire permettra d'effectuer l'attribution des catégories aux véhicules.

L'analyse discriminante utilisée pour classer les véhicules selon leur catégorie permet un gain de temps pour l'opérateur. De plus elle permet également de pallier le problème des données aberrantes qui étaient éliminées à la main par l'opérateur. Après avoir appliqué cette méthode d'analyse discriminante prédictive pour classer les véhicules selon leurs catégories on peut appliquer une des méthodes de M-estimation sur les données. Du fait de la non présence de données aberrantes, les droites d'estimation ne seront pas influencées par une ou plusieurs données aberrantes.

Annexes

Annexe A

M-Estimateurs

Cette annexe a été écrite grâce au livre [7] et [4] et à l'article [8]

Les M-estimateurs constituent une large classe de statistiques obtenues par la minimisation d'une fonction dépendant des données et des paramètres du modèle. Le processus du calcul d'un M-estimateur est appelé M-estimation. De nombreuses méthodes d'estimation statistiques peuvent être considérées comme des M-estimateurs. Dépendant de la fonction à minimiser lors de la M-estimation, les M-estimateurs peuvent permettre d'obtenir des estimateurs plus robustes que les méthodes plus classiques.

Les M-estimateurs ont été introduits en 1964 par Peter Huber sous la forme d'une généralisation de l'estimation par maximum de vraisemblance à la minimisation d'une fonction ρ sur l'ensemble des données. Ainsi, le (ou les) M-estimateur associé aux données et à la fonction ρ est estimé par :

$$\widehat{\theta} = argmin_\theta \sum_{i=1}^{n} \rho(x_i, \theta)$$

Le M de M-estimateur provient donc de Maximum de vraisemblance (Maximum likelihood-type en anglais) et les estimateurs par maximum de vraisemblance sont un cas particulier des M-estimateurs.

A Estimateurs du maximum de vraisemblance

Si g est la densité d'une certaine loi de probabilité de paramètre θ inconnu et si nous avons n observations $x_1,\cdots,x_n$, l'estimateur du maximum de vraisemblance de θ est donné par :

$$\hat{\theta} = \arg\max_{\theta\in\mathbb{R}} \prod_{i=1}^{n} g(x_i;\theta) \tag{A.1}$$

ou en passant au logarithme et en prenant l'opposé :

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}} -\sum_{i=1}^{n} log(g(x_i;\theta)) \tag{A.2}$$

Dans le cas gaussien, c'est à dire lorsque les variables d'erreurs ϵ_i sont distribuées indépendamment selon une loi normale $N(0, \sigma^2)$ alors les variables aléatoires Y_i suivent également une loi normale, de paramètres $(M.f)_i$ et σ^2. On peut alors donner la loi du vecteur $Y : Y \rightsquigarrow N_n(M.f, \sigma^2.I_n)$, c'est une loi normale à n dimensions.

La densité de celle-ci est :

$$\exp(-\frac{1}{2\sigma^2}||Y - M.f||^2) \tag{A.3}$$

On en déduit que l'estimateur du maximum de vraisemblance $\hat{f}^{MV}$ de f est donné par :

$$\hat{f}^{MV} = \arg\max_{f \in \mathbb{R}^m} \exp(-\frac{1}{2\sigma^2}||Y - M.f||^2) \tag{A.4}$$

ou encore :

$$\hat{f}^{MV} = \arg\min_{f \in \mathbb{R}^m} ||Y - M.f||^2 \tag{A.5}$$

Soient $r_i(f) = y_i - (M.f)_i$, $i = 1, \cdots, n$ les résidus au modèle. L'estimateur des moindres carrées $\hat{f}$ de f vérifie l'équation :

$$\hat{f}^{MV} = arg\min_{f \in \mathbb{R}^m} ||r(f)||^2 = arg\min_{f \in \mathbb{R}^m} \sum_{i=1}^{n} r_i(f)^2 \tag{A.6}$$

B Modèle de départ

Soit le modèle linéaire $Y = M.f + \epsilon$, où Y est le vecteur des réponses, M la matrice des données, f le vecteur des paramètres inconnus et ϵ le bruit associé au modèle. Nous avons n réalisations indépendantes $y_i = (M.f)_i + \epsilon_i$ de ce modèle. On note p la dimension de f.
Ce modèle s'écrit sous forme matricielle :

$$\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & m_{11} & \cdots & m_{pn} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & m_{1n} & \cdots & m_{pn} \end{pmatrix} \cdot \begin{pmatrix} f_0 \\ \vdots \\ f_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

L'une des hypothèses utilisées pour réaliser des tests d'influence ou des intervalles de confiance sur les estimateurs des paramètres inconnus est la normalité des variables ϵ_i et donc des variables réponses Y_i. Soient μ et σ^2 l'espérance et la variance de cette loi normale. Cela signifie que la probabilité pour une réponse de se trouver hors de l'intervalle [-3σ,3σ] est très faible. Or dans de nombreux jeu de données, il existe des points appelés points aberrants qui se trouvent éloignés des autres points ou ne semblent pas correspondre à ce qui était attendu. Cela remet en cause l'hypothèse de normalité. Cela nous amène à considérer les M-estimateurs.

Si les variables d'erreurs ϵ_i sont distribuées indépendamment selon une loi normale $N(0, \sigma^2)$ alors les variables aléatoires Y_i suivent également une loi normale, de paramètres $(M.f)_i$ et σ^2. On peut alors donner la loi du vecteur $Y : Y \rightsquigarrow N_n(M.f, \sigma^2.I_n)$, c'est une loi normale à n dimensions.

C Construction du M-estimateur

On remarque que dans les équations de la section A.1, les résidus sont élevés au carré pour faire le calcul de l'estimateur. Si pour une observation i le résidu associé est important, celui-ci peut être une valeur aberrante. L'idée des M-estimateurs est de réduire l'incidence de telles observations. Au lieu d'utiliser la fonction $x \longmapsto x^2$ on va prendre une autre fonction, noté ρ et chercher à minimiser en f la quantité

$$\sum_{i=1}^{n} \rho(r_i(f)) \tag{A.7}$$

C.1 M-estimateur Geman et McClure

Dans la littérature, il existe un certain nombre de fonctions ou de familles de fonctions ρ. Nous n'en présenterons que deux ici.

Le M-estimateur de Geman et McClure est donné par la fonction :

$$\rho_{GM}(r) = \frac{r^2}{1 + r^2} \tag{A.8}$$

et est représenté sur le graphique :

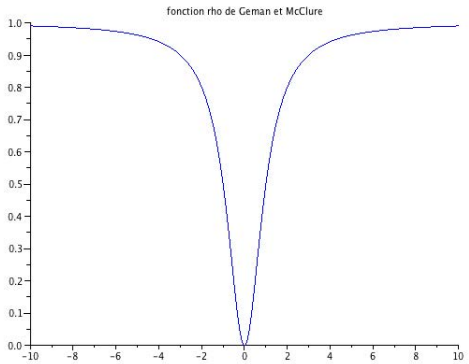


FIGURE A.1 – Geman et McClure

C.2 M-estimateur « Smooth Exponential Family »

Le M-estimateur « Smooth Exponential Family », paramétré par une valeur α, est donné par la fonction :

$$\rho_\alpha(r) = \frac{1}{\alpha}((1 + r^2)^\alpha - 1) \tag{A.9}$$

Les M-estimateurs servent à pallier le problème de la loi normale avec laquelle nous pouvons difficilement traiter les données aberrants. Ce qu'on souhaite en utilisant les M-estimateurs est la robustesse de l'estimation, c'est-à-dire de trouver un estimateur qui ne perde pas ses propriétés et ses qualités en présence de donnés aberrants.

D Estimation de l'échelle

Les M-estimateurs ne sont pas invariants par rapport au paramètre d'échelle ρ.Afin de pouvoir faire les estimations des paramètres de la droite de régression, il s'agit donc d'abord d'estimer cette échelle ρ.
Mathématiquement parlant, pour Y un vecteur aléatoire gaussien de dimension n dont les composantes Y_i sont indépendantes, et de loi normale $\mathcal{N}(\theta_i, \sigma^2)$, en utilisant l'algorithme du maximum de vraisemblance, on obtient :

$$l(y_1, \cdots, y_n; \theta_i, \sigma) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta_i)^2 \tag{A.10}$$

Et en prenant un minimum en σ de la quantité :

$$e_\sigma = n\ln(\sigma) + \frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i^2 \tag{A.11}$$

En calculant les M-estimateurs, nous avons introduits une fonction ρ qui appliquée à la place de la fonction $x \longmapsto x^2$.En faisant ici cette transformation, on obtient :

$$e_{M,\sigma} = n\ln\sigma + \frac{1}{2}\sum_{i=1}^{n}\rho\left(\frac{r_i}{\sigma}\right) \tag{A.12}$$

E Méthode du bootstrap

La méthode du bootstrap a été introduit à la fin des années 1970 avec B. Efron. C'est une méthode de rééchantillonnage qui permet de faire de l'estimation en créant de nouveaux jeux de données à partir du jeu de données de départ. Cette méthode est utilisée dans des cas où les hypothèses habituelles ne sont pas vérifiées ou lorsque la distribution des paramètres est inconnue. En effet l'un de ses principaux avantages est que peu d'hypothèses sont nécessaires pour la mettre en oeuvre.

On a un n-échantillon indépendant x = $(x_1, ..., x_n)$ issu des variables aléatoires $(X_1, ..., X_n)$. Ces variables aléatoires sont distribuées comme une variable X de fonction de répartition F inconnue. La distribution de X est donc inconnue. Nous avons des réalisations de cette variable aléatoire et nous pouvons alors construire une distribution empirique de X.

Deux principales méthodes sont utilisées, la méthode paramétrique et la méthode non paramétrique.

La méthode paramétrique peut être utilisée si on connaît ou si on a une idée de la famille de distributions à laquelle appartient X. Dans ce cas on estime la distribution de X avec le premier échantillon obtenu et on ajuste le modèle. On réalise ensuite des estimations dans le but d'obtenir d'autres échantillons par rapport à cette distribution estimée et non par rapport à l'échantillon de départ en lui-même. On simule N n-échantillons suivant la distribution estimée et pour chacun de ces échantillons on peut recalculer l'estimation du paramètre voulu.

La méthode non paramétrique quant à elle peut être utilisée lorsqu'on ne connaît pas du tout la distribution de X. On simule N n-échantillon à partir de l'échantillon de départ. On réalise N fois n tirages avec remise dans la population de départ.

Pour chacun des N échantillons simulés, on obtient une valeur $\hat{\theta}_i$. La distribution des $\hat{\theta}_i$, avec i=1,...,N, est une approximation de la distribution théorique de $\hat{\theta}$. On moyenne empirique des ces $\hat{\theta}_i$ sera alors une nouvelle estimation de $\hat{\theta}$.

Annexe B

Analyse discriminante linéaire

Cette annexe a été écrite grâce aux livres [6] et [9].

On désigne sous le nom d'analyse discriminante une famille de techniques destinées à classer (affecter à des classes préexistantes) des individus caractérisés par un certain nombre de variables numériques ou nominales.

L'origine de cette méthode remonte aux travaux de Fisher (1936) ou, de façon moins directe, à ceux de Mahalanobis (1936). Elle est une des techniques d'analyse multidimensionnelle les plus utilisées en pratique.

L'analyse linéaire discriminante ou analyse factorielle discriminante est une méthode à la fois descriptive et prédictive, qui donne lieu, comme des méthodes factorielles (par exemple l'ACP), à des calculs d'axes principaux. Elle peut être considérée comme une extension de la régression multiple dans le cas où la variable à expliquer est nominale et constitue la variable de partition.

A Formulation du problème et notations

On dispose de n individus ou observations, indexés par $i, 1 \leq i \leq n$, décrits par un ensemble de p variables $(x_1, \cdots, x_p)$ et réparties en q classes définies a priori par la variable y nominale à q modalités.

L'analyse discriminante se propose dans un premier temps de séparer au mieux les q classes à l'aide des p variables explicatives. Dans un deuxième temps, elle cherche à résoudre le problème de l'affectation d'individus nouveaux, caractérisés par les p variables, à certaines classes déjà identifiées sur l'échantillon des n individus (appelé *échantillon d'apprentissage*).

On distingue par conséquent deux démarches successives, d'ordre descriptif puis décisionnel :
- chercher des fonctions linéaires discriminantes sur l'échantillon d'apprentissage de taille n qui sont des combinaisons linéaires des variables explicatives $(x_1, \cdots, x_p)$ dont les valeurs séparent au mieux les q classes
- connaître la classe d'affectation de n' nouveaux individus décrits par les variables explicatives $(x_1, \cdots, x_p)$. Il s'agit ici d'un problème de *classement* dans des classes préexistantes.

Soit le tableau des données $\mathbf{X}$ à n lignes (individus ou observations) et p colonnes (variables), de terme général x_{ij}. Les n individus sont partitionnés en q classes. Chaque classe k caractérise un sous-nuages I_k de n_k individus i avec :

$$\sum_{k=1}^{q} n_k = n \tag{B.1}$$

Par $\overline{x}_{kj}$ on désigne la moyenne de la variable x_j dans la classe k. C'est la $j^{ème}$ coordonnée du centre de gravité G_k du sous-nuage I_k.

$$\overline{x}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} x_{ij} = G_{kj} \tag{B.2}$$

La moyenne de la variable x_j sur l'ensemble des individus qui correspond à la $j^{ème}$ coordonnées du centre de gravité G du nuage des individus vaut :

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} = \sum_{k=1}^{q} \frac{n_k}{n} \overline{x}_{kj} = G_j \tag{B.3}$$

B Fonctions linéaires discriminantes

L'analyse factorielle discriminante consiste à rechercher les combinaisons linéaires de p variables explicatives $(x_1, \cdots, x_p)$ qui permettent de séparer au mieux les q classes.

La première combinaison linéaire sera celle dont la variance entre les classes (inter-classes) est maximale, afin d'exalter les différences entre les classes, et dont la variance à l'intérieur des classes (intra-classes) minimale pour que l'étendue dans les classes soit délimitée. Puis, parmi les combinaisons linéaires non corrélées à la première, on recherchera celle qui discrimine le mieux les classes. Ces combinaisons linéaires seront les *fonctions linéaires discriminantes.*

Désignons par *a(i)* la valeur, pour l'individu i, d'une combinaison linéaire $\mathbf{a}$ des p variables préalablement centrées :

$$a(i) = \sum_{j=1}^{p} a_j (x_{ij} - \overline{x}_j) \tag{B.4}$$

La variance *var(**a**)* de la nouvelle variable *a(i)* vaut, puisque *a(i)* est centrée :

$$var(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} a^2(i) = \frac{1}{n} \sum_{i=1}^{n} [\sum_{j=1}^{p} a_j (x_{ij} - \overline{x}_j)]^2 \tag{B.5}$$

$$var(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{j'=1}^{p} a_j a_{j'} (x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}_{j'}) \tag{B.6}$$

En posant :

$$t_{jj'} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}_{j'}) = cov(x_j, x_{j'}) \tag{B.7}$$

la variance de la combinaison des variables $\mathbf{a}$ peut s'écrire :

$$var(\mathbf{a}) = \sum_{j=1}^{p} \sum_{j'=1}^{p} a_j a_{j'} cov(x_j, x_{j'}) = \mathbf{a'Ta} \tag{B.8}$$

où $\mathbf{a}$ désigne le vecteur dont les p composantes sont $a_1, \cdots, a_p$ et $\mathbf{T}$ désigne la matrice des covariances des p variables, de terme général $t_{jj'}$.

Nous allons montrer que la variance de $\mathbf{a}$ se décompose en variance intra-classes et en variance inter-classes, ce qui correspond à une décomposition analogue de la matrice des covariances $\mathbf{T}$.

B.1 Décomposition de la matrice de covariance

La covariance totale entre deux variables x_j et xj' s'écrit :

$$cov(x_j, x_{j'}) = \frac{1}{n} \sum_{k=1}^{q} [\sum_{i \in I_k} (x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}_{j'})] = t_{jj'} \tag{B.9}$$

Comme en *analyse de la variance*, nous allons décomposer $cov(x_j, x_{j'})$ en somme de covariances *intra-classes* (à l'intérieur des classes) et covariances *inter-classes* (entre les classes).

Pour cela on va partir de l'identité, pour i, j, k :

$$(x_{ij} - \overline{x}_j) = (x_{ij} - \overline{x}_{kj}) + (\overline{x}_{kj} - \overline{x}_j) \tag{B.10}$$

La somme entre crochets dans la formule de la covariance se décompose alors en quatre termes, dont deux sont nuls.

En effet par définition de $\overline{x}_{kj}$:

$$\sum_{i \in I_k} (x_{ij} - \overline{x}_{kj})(x_{kj'} - \overline{x}_{j'}) = (x_{kj'} - \overline{x}_{j'}) \sum_{i \in I_k} (x_{ij} - \overline{x}_{kj}) = 0 \tag{B.11}$$

De la même manière

$$\sum_{i \in I_k} (x_{kj} - \overline{x}_j)(x_{ij'} - \overline{x}_{kj'}) = (x_{kj} - \overline{x}_j) \sum_{i \in I_k} (x_{ij'} - \overline{x}_{kj'}) = 0 \tag{B.12}$$

Il reste la formule dite *formule de décomposition de Huyghens* :

$$t_{jj'} = d_{jj'} + e_{jj'} \tag{B.13}$$

avec

$$d_{jj'} = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \in I_k} (x_{ij} - \overline{x}_{kj})(x_{ij'} - \overline{x}_{kj'}) \tag{B.14}$$

$$e_{jj'} = \sum_{k+1}^{q} \frac{n_k}{n} (\overline{x}_{kj} - \overline{x}_j)(\overline{x}_{kj'} - x_{j'}) \tag{B.15}$$

Ces relations se notent sous forme matricielle :

$$\mathbf{T} = \mathbf{D} + \mathbf{E} \tag{B.16}$$

Ainsi la variance d'une combinaison linéaire **a** des variables se décompose d'après la relation B.16 en variance interne et variance externe :

$$\mathbf{a'Ta} = \mathbf{a'Da} + \mathbf{a'Ea} \tag{B.17}$$

On cherche, parmi toutes les combinaisons linéaires des variables, celles qui ont une variance intra-classes minimale et une variance inter-classes maximale. En projection sur l'axe discriminant **a**, chaque sous-nuage doit être, dans la mesure du possible, à la fois bien regroupé et bien séparé des autres sous-nuages.

Il s'agit donc de chercher **a** tel que le quotient $\frac{\mathbf{a'Ea}}{\mathbf{a'Da}}$ soit maximal (ou $\frac{\mathbf{a'Da}}{\mathbf{a'Ea}}$ minimal).
D'après la relation B.17 il est équivalent de minimiser $\frac{\mathbf{a'Ta}}{\mathbf{a'Ea}}$ ou de rendre maximun $f(\mathbf{a})$ tel que :

$$f(\mathbf{a}) = \frac{\mathbf{a'Ea}}{\mathbf{a'Ta}} \tag{B.18}$$

B.2 Calcul des fonctions linéaires discriminantes

La fonction $f(\mathbf{a})$ à maximiser est le rapport de la variance inter-classes à la variance totale. Cette fonction étant homogène de degré 0 en **a** (invariante si **a** est changé en ξ **a**, ξ étant un scalaire quelconque), il est équivalent de rechercher le maximum de la forme quadratique **a'Ea** sous la *contrainte* quadratique **a'Ta** = 1.

Ceci conduit à la relation :

$$\mathbf{Ea} = \lambda\mathbf{Ta} \tag{B.19}$$

Lorsque la matrice des covariances **T** est inversible, on obtient :

$$\mathbf{T^{-1}Ea} = \lambda\mathbf{a} \tag{B.20}$$

a est le vecteur propre de $\mathbf{T^{-1}E}$ relatif à la plus grande valeur propre λ.

En prémultipliant les deux membres de B.19 par le vecteur **a'** on constate que **a'Ea**, le maximum cherché, n'est autre que λ.

La plus grande valeur propre λ, quotient de la variance *externe* de la fonction discriminante par la variance *totale*, est inférieur à 1 d'après la relation B.16. On l'appelle quelquefois *pouvoir discriminant* de la fonction **a**.

B.3 Cas de deux classes : équivalence avec la régression multiple

Lorsque la variable **y** ne prend que deux valeurs, chacune caractérisant une classe, des simplifications apparaissent. L'analyse discriminante est alors un cas particulier de la régression multiple.

On repérera les deux classes par les indices *1* et *2*. La matrice des covariances **E** entre classes a pour terme général :

$$e_{jj'} = \frac{n_1}{n}(\overline{x}_{1j} - \overline{x}_j)(\overline{x}_{1j'} - \overline{x}_{j'}) + \frac{n_2}{n}(\overline{x}_{2j} - \overline{x}_j)(\overline{x}_{2j'} - \overline{x}_{j'}) \tag{B.21}$$

avec :

$$\overline{x}_j = \frac{n_1}{n}\overline{x}_{1j} + \frac{n_2}{n}\overline{x}_{2j} \tag{B.22}$$

En remplaçant $\overline{x}_j$ par sa valeur et en tenant compte du fait que $n_1 + n_2 =$ n, on trouve :

$$e_{jj'} = \frac{n_1 n_2}{n^2}(\overline{x}_{1j} - \overline{x}_{2j})(\overline{x}_{1j'} - \overline{x}_{2j'}) \tag{B.23}$$

La matrice **E** d'ordre (*p*,*p*) et de rang *1*, peut être considérée comme le produit d'une matrice colonne **c** par sa transposée :

$$\mathbf{E} = \mathbf{cc'} \tag{B.24}$$

avec :

$$c_j = \frac{\sqrt{n_1 n_2}}{n}(\overline{x}_{1j} - \overline{x}_{2j}) \tag{B.25}$$

La relation B.19 s'écrit alors :

$$\mathbf{T^{-1}cc'a} = \lambda\mathbf{a} \tag{B.26}$$

Prémultiplions les deux membres par **c'** :

$$[\mathbf{c'T^{-1}c}]\mathbf{c'a} = \lambda\mathbf{c'a} \tag{B.27}$$

La quantité entre crochets est un scalaire, égal par conséquent à λ qui est ici une valeur propre unique car **E** est de rang *1*.

Cette valeur propre vaut donc : $\lambda = \mathbf{c'T^{-1}c}$

λ est appelée *distance généralisée* entre les deux classes ou encore "*Distance de Mahalanobis*". Le vecteur propre correspondant : $\mathbf{a} = \mathbf{T^{-1}c}$ est l'unique fonction discriminante.

Considérons un vecteur **w** à *n* composantes, défini par :

$$w_i = \begin{cases} \sqrt{n_1/n_2} & \text{si l'individu } i \text{ est membre de la classe } 1 \\ -\sqrt{n_1/n_2} & \text{si l'individu } i \text{ est membre de la classe } 2 \end{cases}$$

La régression multiple expliquant **w** par les colonnes de **X** conduit au vecteur de coefficients noté ici **b** :

$\mathbf{b} = \mathbf{(X'X)^{-1}X'w}$, avec : $\frac{1}{n}\ \mathbf{X'X} = \mathbf{T}$

On vérifie que : $\frac{1}{n}\ \mathbf{X'w} = \mathbf{c}$ d'où $\mathbf{b} = \mathbf{T^{-1}c}$

Le vecteur des *coefficients de régression* **b** coïncide par conséquent avec le vecteur des composantes de la *fonction discriminante* **a** calculé précédemment.

C Principes des règles d'affectation (ou de classement)

Une fois trouvées les fonctions discriminantes qui séparent au mieux les individus répartis en *q* classes, on veut trouver la classe d'affectation d'un nouvel individu, pour lequel on connaît les valeurs des variables $(x_1, \cdots, x_p)$. Une règle simple et géométrique d'affectation est de choisir la classe dont le centre de gravité est le plus proche du point-individu. La métrique généralement utilisée dans les applications est celle de *Mahalanobis globale*.

Cette approche purement géométrique ne prend cependant pas en compte les probabilités *a priori* des différentes classes, qui peuvent être très inégales dans certaines applications. Le modèle bayésien d'affectation permet d'enrichir ce point de vue.

C.1 Le modèle bayésien d'affectation

Au moment de l'apprentissage, nous savons que l'individu i appartient au groupe I_k (appartenance codée par la valeur : $y_i = k$) et nous calculons une estimation de la probabilité $\mathrm{P}(x_i|I_k)$, c'est-à-dire la probabilité de x_i sachant que I_k est réalisé.

Au moment de l'affectation d'un individu nouveau noté $\mathbf{x}$, on peut calculer les différents $\mathrm{P}(\mathbf{x}|I_k)$ pour $k = 1,2,\ldots,q$. Il paraît raisonnable d'affecter $\mathbf{x}$ à la classe I_k pour laquelle $\mathrm{P}(\mathbf{x}|I_k)$ est maximale. Cependant ce ne sont pas les probabilités $\mathrm{P}(\mathbf{x}|I_k)$ qu'il faudrait connaître mais les probabilités $\mathrm{P}(I_k|\mathbf{x})$, c'est-à-dire la probabilité du groupe I_k sachant que $\mathbf{x}$ est réalisé.

Le théorème de Bayes permet de procéder à cette *inversion des probabilités*. Il exprime $\mathrm{P}(I_k|\mathbf{x})$ en fonction de $\mathrm{P}(\mathbf{x}|I_k)$, $\mathrm{P}(I_k)$ et $\mathrm{P}(\mathbf{x})$:

$$P(I_k|\mathbf{x}) = \frac{P(\mathbf{x}|I_k)P(I_k)}{P(\mathbf{x})} \tag{B.28}$$

$\mathrm{P}(I_k)$ est la probabilité *a priori* du groupe k. $\mathrm{P}(\mathbf{x})$ s'exprime en fonction de $\mathrm{P}(\mathbf{x}|I_k)$ et de $\mathrm{P}(I_k)$; d'où la formulation classique du théorème de Bayes :

$$P(I_k|\mathbf{x}) = \frac{P(\mathbf{x}|I_k)P(I_k)}{\sum_{k=1}^{q} P(\mathbf{x}|I_k)P(I_k)} \tag{B.29}$$

Le dénominateur est le même pour toutes les classes. La classe d'affectation de $\mathbf{x}$ sera celle pour laquelle le produit $\mathrm{P}(\mathbf{x}|I_k) \times \mathrm{P}(I_k)$ est maximal. Si les probabilités *a priori* $\mathrm{P}(I_k)$ des classes sont égales pour toutes les valeurs de k, les classements selon $\mathrm{P}(I_k|\mathbf{x})$ et $\mathrm{P}(\mathbf{x}|I_k)$ sont identiques.

Pour tester l'efficacité des règles d'affectation, on mesure les erreurs de classement par des méthodes de rééchantillonnage, notamment la validation croisée ou le bootstrap. Comme dans le cas du modèle linéaire, le choix des variables explicatives est une opération délicate. L'étude de la stabilité des fonctions discriminantes est difficile. Les règles d'affectation ainsi que l'estimation des taux d'erreur de classement dépendent souvent de la taille de l'échantillon d'apprentissage.

Bibliographie

[1] Guillaume Dutilleux. *Projet dBEuler : Outils pour la mesure de bruit de roulement au passage*, 2006.

[2] Marie-Paule Ehrhart. Application de méhodes de statistique robuste a l'analyse de mesures de bruit de roulement. Master's thesis, UFR Maths-info Strasbourg, 2011.

[3] Frédéric Bertrand et Myriam Maumy. *Plans d'expériences : Modéles d'analyse de la variance*, 2013.

[4] Virginie Delsart et Nicolas Vannecloo. *Estimation, tests, échantillonnage*. Septentrion, 2011.

[5] Makarim Ghazza. Statistiques robustes appliquées aux mesures de bruit de roulement. Master's thesis, UFR Maths-info Strasbourg, 2012.

[6] Ludovic Lebart, Alain Morineau, and Marie Piron. *Statistique exploratoire multidimensionnelle*. Dunod, 2006.

[7] Ricardo A. Maronna. *Robust Statistics - Theory and methods*. John Wiley and Sons, 2006.

[8] Cosma Shalizi. The bootstrap. *American Scientist*, 98, 2010.

[9] Michel Tenenhaus. *Statistique : Méthodes pour décrire, expliquer et prévoir*. Dunod, 2007.