Statistiques robustes appliquées au bruit de roulement

Makarim Ghazza

UNIVERSITÉ DE STRASBOURG

Statistiques robustes appliquées aux mesures de bruit de roulement

Rapport de stage

Laboratoire Régional des Ponts et Chaussées de Strasbourg

Groupes «Acoustique» et «Méthodes Physiques»
Sous la direction de Guillaume Dutilleux et Pierre Charbonnier

Makarim GHAZZA
Master Statistique 1ère année
2011-2012

Résumé

Dans le cadre de la réflexion sur l'amélioration de l'analyse des mesures de bruit de roulement au passage, des techniques d'estimation alternatives à la régression linéaire ont été étudiées en 2011 lors du stage de Marie-Paule Ehrhart. Toutefois, il n'a pas été possible de hiérarchiser les différentes techniques étudiée dans le but d'en retenir une qui puisse être intégrée à une procédure de dépouillement standardisée. Le stage de M.P. Ehrhart repose aussi la question de la pertinence de l'analyse du bruit des poids lourds, catégorie pour laquelle le nuage de points (Vitesse, L_{Amax}) est très dispersé.
La première partie de mon travail consiste à tester les différentes méthodes pour départager les droites d'estimations, en travaillant sur une petite tranche de vitesse.
La deuxième partie concerne les poids lourds . On se propose de remplacer le L_{Amax} par un indicateur moyenné sur la durée du passage du véhicule, comme le L_{AE} ou le L_{Aeq}. Il faudra évaluer l'impact du changement d'indicateur sur la dispersion des données.
La troisième partie concerne la classification non supervisée en testant l'algorithme EM dans un premier temps, et en procédant à une nouvelle approche utilisant les spectres de passages.
Les méthodes statistiques utilisées durant le stage sont a priori la régression classique ou les M-estimateurs, l'ACP, l'IRLS, l'algorithme EM, l'ISOMAP[3], K-means "clustering", et les tests de normalité.
Les traitements des mesures seront réalisés principalement avec **Scilab**, **R** et **Matlab**.

Mots-clés

Mesure de bruit de roulement, valeurs aberrantes, régression, M-estimation, ACP robuste, algorithme EM, Isomap, Isodata, K-means..

Remerciements

Je tiens à remercier dans un premier temps M. Georges Kuntz, directeur du Laboratoire Régional de Strasbourg, de m'avoir accueillie au sein du Laboratoire et de m'avoir permis d'effectuer mon stage dans de bonnes conditions.

Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance à Guillaume Dutilleux et Pierre Charbonnier, mes maîtres de stage, pour l'expérience enrichissante et pleine d'intérêt qu'ils m'ont fait vivre durant cette période de stage au sein du laboratoire et pour leurs conseils tout au long du stage.

Je remercie enfin les membres de l'équipe «Acoustique» et «Méthodes Physiques» qui m'ont permis de découvrir tous les différents projets menés par le laboratoire et de m'avoir bien accueilli au sein de leur groupe, sans oublier de remercier Mr Nicolas Poulin, ingénieur de Recherche à l'IRMA, qui m'a aidée au début de mon stage.

Table des matières

Chapitre 1

Présentation et objectif du stage

1.1 Organisme d'accueil

Mon stage s'est déroulé au groupe acoustique du Laboratoire Régional des Ponts et Chaussées (LRPC) de Strasbourg, qui fait partie du CETE de l'Est et se trouve au 11, rue Jean Mentelin à Koenigshoffen. Il a été effectué sous la direction de Mr. Guillaume Dutilleux et Mr Pierre Charbonnier. Le stage a eu lieu du 4 juin au 14 juillet et du 1^{er} au 31 août 2012

1.1.1 CETE de l'EST

Les CETE, pour « Centre d'Etudes Techniques de l'Equipement », sont au nombre de huit répartis dans toute la France dont dépendent des Laboratoires Régionaux. Ces organismes sont des bureaux d'études d'ingénierie publique et font partie du Ministère de l'Ecologie, du Développement Durable, et de l'Energie. Le CETE de l'Est répartit ses compétences sur trois sites : direction et département d'études à Metz, et deux laboratoires régionaux, à Nancy et Strasbourg. Les domaines d'interventions traditionnels des CETE sont les infrastructures de transport et de la construction, des missions concernant l'aménagement du territoire. Le CETE réalise des prestations pour le compte de l'État et également pour les collectivités territoriales, ou pour d'autres organismes publics ou privés.
Les CETE développent aussi des activités de recherche, notamment dans le cadre des équipes de Recherche associées à l'IFSTTAR.

1.1.2 LRPC de Strasbourg

Le Laboratoire Régional des Ponts et Chaussés de Strasbourg, dirigé par M. Georges Kuntz, compte 77 employés, emploie ponctuellement des vacataires et accueille régulièrement des stagiaires. Il regroupe des compétences diverses en cinq départements :

– Géotechnique – Terrassement – Chaussées
– Ouvrages d'art
– Construction
– Acoustique
– Méthodes Physiques
Un organigramme du laboratoire est présenté à la figure 1.1 :

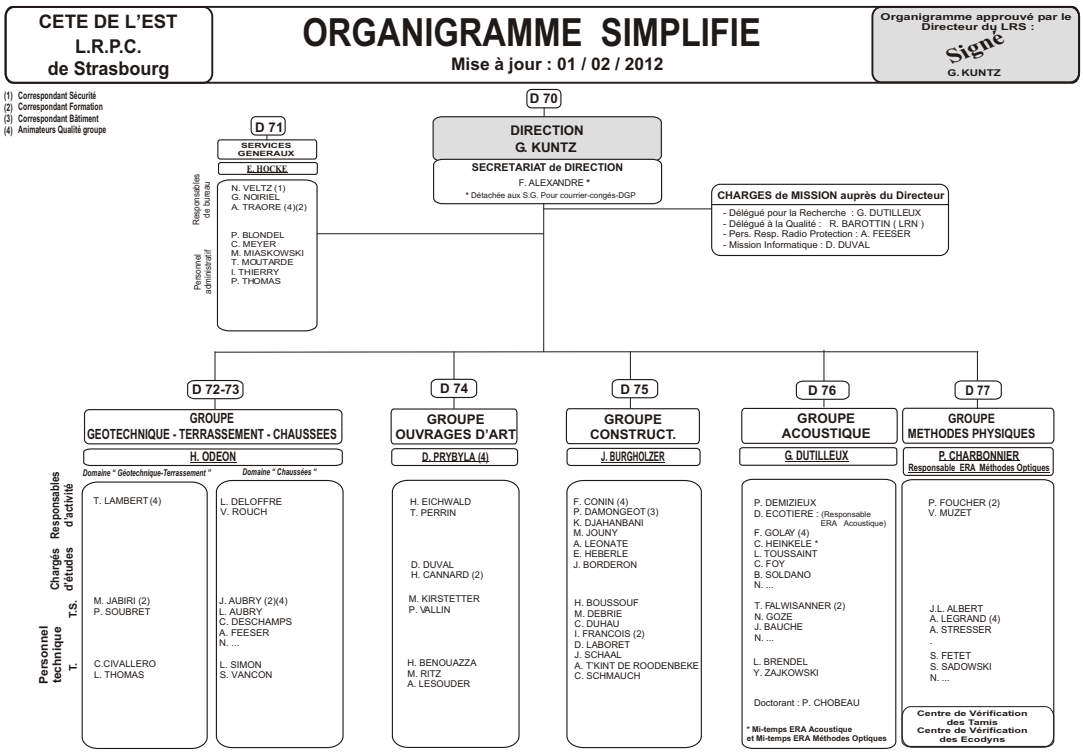


FIGURE 1.1 – Organigramme - Laboratoire Régional de Strasbourg

Le Groupes Acoustiques

Dirigé par M. Guillaume Dutilleux, le groupe acoustique contribue à lutter concrètement contre les nuisances sonores, par des mesures sur le terrain, des études, des expertises. La recherche fait aussi partie du travail effectué par le groupe, dans le cadre de l'Equipe Recherche Associée à l'Ifsttar.
Comme exemples de mission :la cartographie du bruit des grandes villes et infrastructures, des études concernant l'acoustique des bâtiments, ou encore le bruit industriel et de voisinage. Les campagnes de mesure de bruit de roulement, qui font l'objet de mon stage, servent à classifier les différents revêtements routiers selon leurs performances acoustiques et le type des véhicules qui les empruntent.
Le travail de ce groupe est très divers mais tout tourne autour de l'acoustique, et contribuent à améliorer le cadre de vie des riverains.

Le Groupe Méthodes Physiques

Comme le groupe Acoustique, le groupe Méthodes Physiques est constitué d'une dizaine de personnes. Il est dirigé par M. Pierre Charbonnier. La recherche occupe 60% du travail de ce groupe, dans le cadre de l'Equipe de Recherche Associée « Imagerie et Méthodes optiques » de l'Iffstar. Les 40% restants sont des applications servant à mettre en oeuvre et à valoriser les produits de la recherche.
Les applications principales se font dans le domaine de la sécurité routière, ou encore dans l'inspection d'ouvrages d'art.

1.2 Contexte du Stage

1.2.1 Contexte administratif

Déroulement d'une campagne de mesure du bruit de roulement selon la norme AFNOR S31-119[5]

Une campagne peut se découper en plusieurs parties(voir FIG 1.2). Tout d'abord, un opérateur, le technicien chargé des mesures, se place à une certaine distance d'une chaussée dont on veut étudier le revêtement et enregistre à l'aide d'un microphone la presssion acoustique engendré $p(t)$ par les véhicules sur leur passage. A côté de cela il enregistre pour chaque véhicule sa vitesse ainsi que sa catégorie, par exemple VL(pour Véhicules Légers), PL(pour Poids Lourds), ou TR(pour Trains Routiers)(Voir Annexe A).
Ensuite les enregistrements sont dépouillés, c'est-à-dire que l'on analyse l'enregistrement audio pour en retirer les informations pertinentes. Pour cela il faut également pouvoir délimiter les passages au sein des enregistrements.
En effet un enregistrement unique peut correspondre aux passage de plusieurs véhicules qui se suivent. Une fois le passage délimité on peut extraire les informations qui nous intéressent tels que le niveau de bruit pondéré A noté L_{Amax},l'intensité sonore qui soit plus proche de l'intensité perçue.
La finalité d'une telle campagne est d'obtenir un niveau sonore type pour un revêtement donné

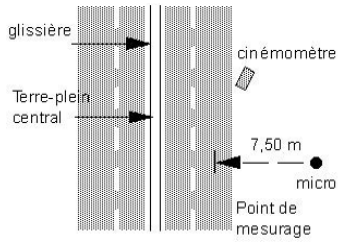


FIGURE 1.2 – Vue en plan

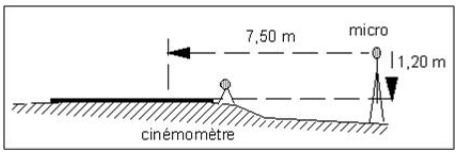


FIGURE 1.3 – Vue en travers

à une vitesse de référence conventionnel et une loi de comportement en fonction de la vitesse.

Pour y arriver, les informations récupérées permettent de tracer une droite de régression du niveau sonore maximum noté L_{Amax} en fonction de la vitesse modifiée par le logarithme de base 10.

$$L_{Amax} = a \log_{10}(\frac{vitesse}{v_{réf}}) + b$$

avec a et b les paramètres inconnus de la droite de régression, et $v_{réf}$ la vitesse de référence.

Ensuite le niveau sonore "type" correspond sur la droite au L_{Amax} à une vitesse de référence donnée selon la catégorie de véhicule.

Le Groupe Acoustique du Laboratoire Régional de Strasbourg s'occupe de classifier les revêtements selon le niveau de bruit de roulement après avoir importé et dépouillé les données avec

le logiciel **dBEuler**[11] (voir Annexe A). Ce logiciel, développé par le LRPC sous l'environnement **Scialb**, réalise une régression linéaire au sens des moindres carrées, mais cette méthode est très sensible aux valeurs aberrantes. Pour pallier ce problème, plusieurs méthodes d'estimations ont été testées en 2011 comme la régression avec un M-estimateur, l'ACP (Analyse en Composantes Principales) ou encore l'ACP robuste qui restent inexactes.
Durant mon stage j'ai utilisé les versions suivantes **Scilab** v5.3.3, **R** v2.15.1 et **Matlab** v16.

1.3 Objectif du stage

Une procédure réalisé par Marie-Paule Ehrhart[6], stagiaire au LPRC en 2011, a permis d'analyser des données issues des campagnes de mesure de bruit de roulement. Plusieurs méthodes d'estimation des paramètres de la droite cherchée ont été utilisées à savoir la régression linéaire, la régression robuste avec un **M-estimateur**(Annexe A), l'**ACP**(Annexe A), l'**ACP** robuste(Annexe A) et la régression experte, cependant on arrivait pas à départager ces méthodes, l'idée dans un premier temps(chapitre 1)est de prendre une petite tranche de vitesse et de calculer un intervalle de confiance pour les données brutes ainsi que pour les divers estimations, dans ce but plusieurs méthodes de statistiques robustes ont été testées.
Un deuxième objectif de stage est d'améliorer la qualité de dépouillement de poids lourds en procédant à un changement d'indices de bruit, comme déjà mentionné dans le contexte de stage lors du dépouillement on obtenait le niveau sonore maximum en fonction de la vitesse, dans le chapitre 2 on va testé l'impact du changement de cet indice par le niveau d'exposition au bruit "SEL", ou par le niveau de pression acoustique "LA_{eq}".
Ensuite, comme cela a été initié par M.P Ehrhart, on peut faire un lien entre les points aberrants des deux catégories VL et PL. Le chapitre 3 présente l'analyse faite avec l'algorithme EM, qui est ici une autre manière de traiter les points aberrants. Il permet aussi de traiter ensemble les données VL et les données PL. Cet algorithme fournit une nouvelle classification des données en deux catégories, dans le cadre de mon stage je vais tester l'algorithme EM pour les données issus des derniers dépouillements.
Le chapitre 4 présente, une autre alternative proposé par mes maitres de stage, qui consiste à regarder du coté des spectres en procédant à une réduction de dimensionnalité et à un algorithme du K-means, en vue d'une classification non supervisée.

Chapitre 2

Différents moyens pour départager les méthodes d'estimations robustes

2.1 Intervalles de confiance pour une tranche de vitesse

2.1.1 Introduction

Dans cette partie, on s'est concentré sur les données déjà exploitées par Marie-Paule Ehrhart(Rothau2009, Haguenau2009) Grâce à ces différentes méthodes, un algorithme général avec plusieurs estimations de droite a été mis en place.Le problème était d'en choisir une, et cela grâce à l'estimation de référence qui est la régression experte. Une solution possible était le calcul du niveau sonore équivalent Leq. Malheureusement, cette alternative ne s'avère pas suffisamment discriminante.
Une autre approche m'a été proposée par mon maitre de stage Mr Guillaume Dutilleux, est celle d'essayer de départager les différentes estimations de droites sous une petite tranche de vitesse, dans la suite de mon travail l'intervalle de vitesse pris est [55 à 65 $km.h^{-1}$] , et les données étudiées sont celles de Rothau2009.
Pour le calcul de l'intervalle de confiance de L_{Amax} **"niveau maximum sonore"**, on utilise la moyenne énergétique définie par :

$$\overline{L_{Amax}} = 10 \log_{10} \frac{\sum_{i=1}^{n} 10^{\frac{L_{Amaxi}}{10}}}{n}$$

Après le calcul de l'intervalle de confiance sous R, on obtient :
$L_{Amax} = 79.34dB$
et $\sigma = 5.17dB$ (σ :écart-type de $LAmax$)
Un intervalle de confiance pour la tranche de vitesse [55 à 65 $km.h^{-1}$] est donc :[77.07 ; 81.6]
Après calcul du L_{Amax} moyen et les intervalles de confiance pour toutes les méthodes d'estimation disponibles on obtient le tableau suivant :

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne Sup
Données mesurées	79.34	77.07	81.6
Régression linéaire	75.16	75.05	75.27
M-Estimation Geman McClure	73	72.75	73.23
ACP	74.01	73.78	74.25
ACP Robuste	72.8	72.55	73.06
Régression Experte	72.85	72.61	73.1

TABLE 2.1 – Intervalle de confiance Rothau2009

On remarque dans la Table 2.1 que la taille d'échantillon pour les données brutes est d'amplitude $4.53db$ ce qui est assez important avec une précision de $2.34dB$

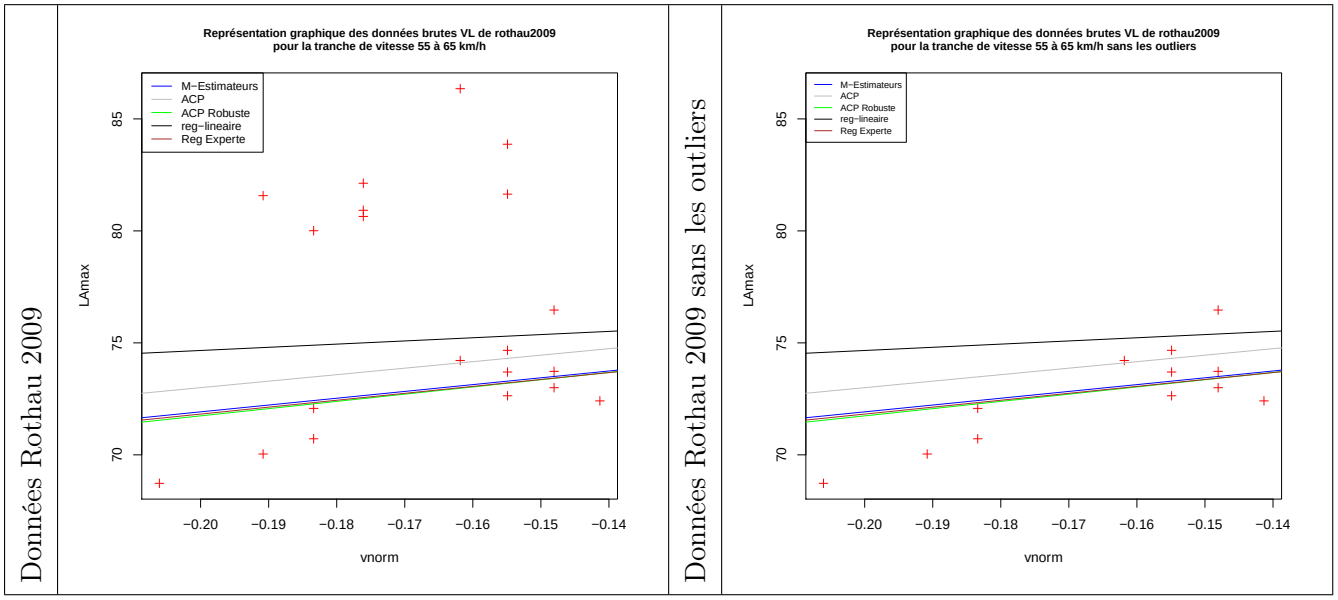
Pour réduire la taille de cette échantillon, il existe des formules et des abaques, connues sous le nom de " Table de Gauss " qui permettent de calculer la taille de l'échantillon et le taux de précision. Pour être sûr à 80% que la réponse est fiable, la formule est la suivante :

$$1.28.\sqrt{1 - N/P}.\sqrt{A.(100 - A)/N} = 1$$

A : Pourcentage de réponses à mesurer
P : Taille de la population
N : Taille de l'échantillon
On voudrait, garder la tranche de vitesse [55 à 65 $km.h^{-1}$] et jouer sur la population nécessaire pour une meilleure précision, vue qu'on avait 20 observation sur 122 pour cette tranche de vitesse, donc pour améliorer l'intervalle de confiance et garder la tranche de vitesse, alors il faudra augmenter la taille de la population, qui devrait passer de 122 à $(65.122/20) = 397$, donc il faudra une population de 397 observation(soit une augmentation de 3.25 l'échantillon de départ), pour avoir un intervalle de confiance d'amplitude 1dB.

2.1.2 Représentation graphique des données Rothau 2009 pour [55 à 65 $km.h^{-1}$]

La présence d'outliers peut nuire au choix de la méthode d'estimation, pour en avoir la certitude j'ai procédé avec le logiciel **R** à une suppression manuelle des valeurs aberrantes.
Ci-dessous les graphes correspondants aux résultats.



En recalculant l'intervalle de confiance sans valeurs aberrantes, avec la même procédure évoquée précédemment, on obtient la Table 2.1.

Méthode d'estimation	$\overline{L_{Amax}}$(en dB)	Borne inf	Borne Sup
Données Brutes	73.15	71.95	74.36
Régression linéaire	75.17	75	75.34
M-Estimation Geman McClure	73.03	72.67	73.4
ACP	74.06	73.72	74.40
ACP Robuste	72.93	72.54	73.30
Régression Experte	72.95	72.6	73.32

TABLE 2.2 – Intervalle de confiance Rothau2009 sans outliers

On voit bien qu'enlevant même les valeurs aberrantes, cela n'améliore pas l'amplitude de l'intervalle de confiance.Dans la suite j'ai testé les différentes méthodes d'estimations, et j'ai recalculé l'intervalle de confiance correspondant.

2.2 Le bootstrap[10]

Pour avoir de meilleurs résultats, plusieurs méthodes ont été testées, en commençant par la méthode du bootstrap paramétrique qui consiste à réaliser de l' inférence statistique quand la distribution des données n'est pas connue ou que les propriétés statistiques d'intérêt sont difficiles à dériver analytiquement.
La construction de l'intervalle de confiance par la méthode du bootstrap consiste à :
-Générer B échantillons bootstrap, $X^*(b)$.
-Calculer T(X*(b)) pour chaque échantillon, b=1,2...B,ou T() est la statistique .
-Estimer la distribution échantillonnage bootstrap de T(c) :
G*(x)=1/B $\sum_{b=1}^{B}$ I(T(X*(b))≤x)
-Rechercher les percentiles $\alpha/2$ et 1-$\alpha/2$ de G*(x) : v*($\alpha/2$) et v*(1- $\alpha/2$)
-L'intervalle de confiance bootstrap de type « percentile » et de niveau 1- α est défini par :
[v*($\alpha/2$), v*(1- $\alpha/2$)]
Le calcul de l'intervalle de confiance avec cette méthode a utilisé la fonction **boot**, sous **R** nécessitant le package **boot**, nous constaterons que les intervalles de confiance de L_{Amax} obtenus par les différentes méthodes d'estimations se chevauchent et donc impossible de les départager par ce biais. Vous trouverez les résultats obtenus avec le **bootstrap** :

Méthode d'estimation	$\overline{LA_{max}}$(en dB)	Borne inf	Borne Sup
Données Brutes	77	77.07	81.6
Régression linéaire	75.12	75.01	75.23
M-Estimation Geman McClure	72.94	72.70	73.17
ACP	73.97	73.75	74.19
ACP Robuste	72.82	72.58	73.07
Régression Experte	72.85	72.61	73.09

TABLE 2.3 – Bootstrap Rothau2009

2.3 Méthode des résidus

-Après discussion avec Mr Nicolas Poulin, responsable de stage et ingénieur de recherche statistique à l'UFR de maths-info de Strasbourg, il m'a conforté dans l'idée que ce n'est pas possible

de départager les méthodes en utilisant l'intervalle de confiance, et d'ailleurs ça s'est vérifié, et que la seule façon c'était de comparer les résidus des modèles, ce que j'ai pu faire en créant une fonction sous R, en voici la sortie et les résultats obtenus sont :

Méthode d'estimation	Sommes des carrés résiduelles	Somme des résidus	S des valeurs absolus des résidus
Régression linéaire	540.24	27.07	88.35
M-Estimation Geman McClure	756.76	71.34	90.70
ACP	630.53	50.65	87.36
ACP Robuste	773.97	73.71	91.27
Régression Experte	768.86	73.02	91.16

TABLE 2.4 – Calcul de résidus pour Rothau2009

Ce résultat n'est pas pertinent, dans la mesure ou la comparaison n'est pas possible, car la distance utilisée diffère d'une méthode à l'autre, de plus l'utilisation qui est faite de résidus diffère également d'une méthode à l'autre.

2.3.1 application du MM.estimateur et du S.estimateur

Pour l'application de ces deux méthodes d'estimations, un programme sous R a été nécessaire.**un MM-estimateur** est utilisé pour déterminer un modèle robuste.Cet estimateur est calculé avec un algorithme itératif, initialisé par une estimation robuste de la matrice de covariance. Cette approche consiste à minimiser une mesure robuste des distances orthogonales des observations au sous-espace de l'ACP (espace résiduel)
Le **S.estimateur** peut être définit comme suit :
Soit ρ une fonction qui satisfait les conditions suivantes :
C_1 :ρ est symétrique, possède une dérivée continue ψ et $\rho(0) = 0$
C_2 :il existe une constante $0 \leq C_0$, tel que ρ est strictement croissante sur $[0\ ,\ C_0]$ et constante sur $[C_0,\infty]$.soit $a_0 = \sup \rho$
C_3 :ρ possède une dérivée seconde ψ' avec $\psi(t)'$ et $u(t) = \psi(t)/t$ continues et bornées
Si $t_i = [(X_i - \mu_n' V_n^{-1}(X_i - \mu_n)]^{1/2}$, et$i = 1...n\ , les S.estimateurs(\mu_n, V_n)$ sont définis comme les solutions du problème d'optimisation
$min|(V_n)|$ définit à $\sum_{i=1}^{n} \rho(t_i) = b_0$
En appliquant le M-M.estimateur et le S.estimateur sur la tranche de vitesse [55 à 65 $km.h^{-1}$] aux données Rohau2009 , le résultat obtenu est présenté dans la Table 2.3 :

Méthode d'estimation	$\overline{L_{Amax}}$	Borne inf	Borne Sup
MM.estimateur	72.72	72.46	72.97
S.estimateur	73.04	72.81	73.28

TABLE 2.5 – MM.estimateur et S.estimateur sur Rothau2009

On voit bien que ni le S.estimateur ou le MM.estimateur ne donnent de meilleurs résultats par rapport à ceux déjà obtenu avec les autres méthodes d'estimations, d'ailleurs ils se rapprochent toutes les deux du M.estimateur avec une meilleure précision certes, mais sans plus.

2.4 Calcul d'intervalles de confiance pour différentes données

L'idée est de calculer d'intervalles de confiance pour différentes donnés, et d'en déduire qu'il n'y a pas moyen de départager les méthodes d'estimations :

-En calculant l'intervalle de confiance pour toutes les valeurs de Rotheau2009VL, et pour les différentes méthodes d'estimation, on obtient :

Méthode d'estimation	$\overline{L_{Amax}}$	Borne inf	Borne Sup
Données Brutes	79.56	77.07	81.60
M-Estimation Geman McClure	79.56	79.31	79.81
ACP	80.28	80.04	80.52
ACP Robuste	79.87	79.6	80.14
Régression Experte	79.6	79.34	79.85
Régression linéaire	60.21	59.95	60.47

TABLE 2.6 – Intervalle de confiance Rothau2009VL

Le calcul de l'intervalle de confiance pour les données de Haguenau2009VL est :

Méthode d'estimation	$\overline{L_{Amax}}$	Borne inf	Borne Sup
Données Brutes	82.86	82.54	83.19
M-Estimation Geman McClure	82.56	82.35	82.77
ACP	82.69	82.45	82.94
ACP Robuste	82.49	82.23	82.74
Régression Experte	82.38	82.17	82.59
Régression linéaire	82.86	82.54	83.19

TABLE 2.7 – Haguenau2009VL

2.4.1 Données Rothau 2009 cat. Poids lourds

Pour calculer l'intervalle de confiance pour les poids lourds pour les données de Rothau2009 PL, ce qui donne :

Méthode d'estimation	$\overline{LA_{max}}$	Borne inf	Borne Sup
Données Brutes	83.59	82.84	84.34
M-Estimation Geman McClure	74.15	73.72	74.58
ACP	75.11	74.07	75.52
ACP Robuste	74.124	73.67	74.58
Régression Experte	74.09	73.65	74.52
Régression linéaire	75.65	75.44	75.84

TABLE 2.8 – Données Rothau2009PL

2.4.2 Données Haguenau 2009 cat. Poids lourds

J'ai regarder aussi pour les données de Haguenau2009PL, et j'ai obtenu la Table 2.7 :

Méthode d'estimation	$\overline{L_{Amax}}$	Borne inf	Borne Sup
Données Brutes	87.72	87.13	88.31
M-Estimation Geman McClure	87.73	87.40	88.06
ACP	88.18	87.32	89.04
ACP Robuste	88.03	87.50	88.55
Régression Experte	87.68	87.39	87.97
Régression linéaire	87.36	87.10	87.61

TABLE 2.9 – Données Haguenau2009PL

Pour les données de Haguenau2009PL, on remarque que les résultats sont meilleures, vue qu'on retrouve un intervalle de confiance plus précis et les méthodes d'estimations se rapprochent mieux des données brutes.

2.5 Conclusion

Pour résumer ce chapitre, on pourra dire que l'idée de départager les droites d'estimations robustes sur une tranche de vitesse(la plus petite)n'est pas probante, comme j'ai mentionné auparavant la taille de l'échantillon doit augmenter pour avoir un intervalle de confiance d'amplitude moins que 1dB, pour l'exemple que j'ai étudié(Rothau2009), il fallait avoir plus que 3 fois l'échantillon de départ, ce qui est difficile à obtenir durant une compagne de mesure au passage.Finalement Le calcul d'intervalles de confiance pour différentes méthodes d'estimation, ne nous a pas plus avancé.
Dans le chapitre suivant, on se focalisera sur les données de poids lourds, on va évaluer l'impact du changement d'indicateur sur la dispersion des données.

Chapitre 3

Évaluation d'indicateurs alternatives au L_{Amax} pour les PL[5]

Cette partie a nécessité le code de Scilab pour le calcul du niveau de pression acoustique équivalent LA_{eq}, ainsi que pour le niveau d'exposition sonore SEL, et un programme R pour le calcul des sommes de résidus, ainsi qu'un test de normalité.
Lors du dépouillement l'opérateur a procédé à l'analyse statistique qui est une régression linéaire du niveau de bruit L_{Amax} par rapport à la vitesse, seulement que pour les données de poids lourds le nuage de points(Vitesse, L_{Amax})est très dispersé, on se propose dans ce chapitre de remplacer le L_{Amax}, par un indicateur moyenné sur la durée de passage du véhicule, comme le SEL "niveau d'exposition au bruit", ou le LAeq "niveau de pression acoustique équivalent"

3.0.1 Introduction

Un bruit est un phénomène physique caractérisé notamment par son niveau de pression acoustique et par sa composition fréquentielle. Ces paramètres constituent les composantes objectives du bruit, Pour tenir compte de la sensibilité de l'oreille humaine, ces paramètres physiques sont pondérés un filtre fréquentiel.
Dans cette partie on se propose de remplacer le L_{Amax} par un indicateur moyenné sur la durée du passage du véhicule, comme le SEL ou le L_{Aeq}(définis ci-dessous), pour évaluer l'impact du changement d'indicateur sur la dispersion des données.

3.0.2 Les indices de bruit :

- L_{Amax} **(ou « niveau instantané maximum »)**

Le L_{Amax} est le niveau maximum de bruit mesuré (avec une pondération fréquentielle A) durant une période de temps donnée. Il correspond à un niveau sonore qui n'est jamais .

- L_{Aeq} **:Equivalent Continuous Sound Level (niveau de pression acoustique équivalent)**

Le L_{Aeq} est la valeur d'un niveau de pression acoustique constant sur toute la durée de l'événement T, qui possède la même énergie acoustique que le bruit variable du passage de la source. Il représente le niveau énergétique moyen (puissance) mise en jeu pendant la durée (T) du passage Il est exprimé en décibel dB(A).On définit le $L_{Aeq}(T)$le niveau énergétique sonore moyen sur une période T par la relation :

$$L_{Aeq} = 10\log_{10}[\tfrac{1}{T}\int_{t_1}^{t_2}(\tfrac{P_A(t)}{P_0})^2 dt]$$

- En dB(A) avec $T_i = |t_1 - t_2|$.
- $L_p(t)$:l'évolution du niveau sonore en un point en fonction du temps

Pour la détermination de $t_{1,i}$ et $t_{2,i}$ la figure (3.1) representatif de la fonction $L_p(t)$, donne un exemple de calculs de ces durées

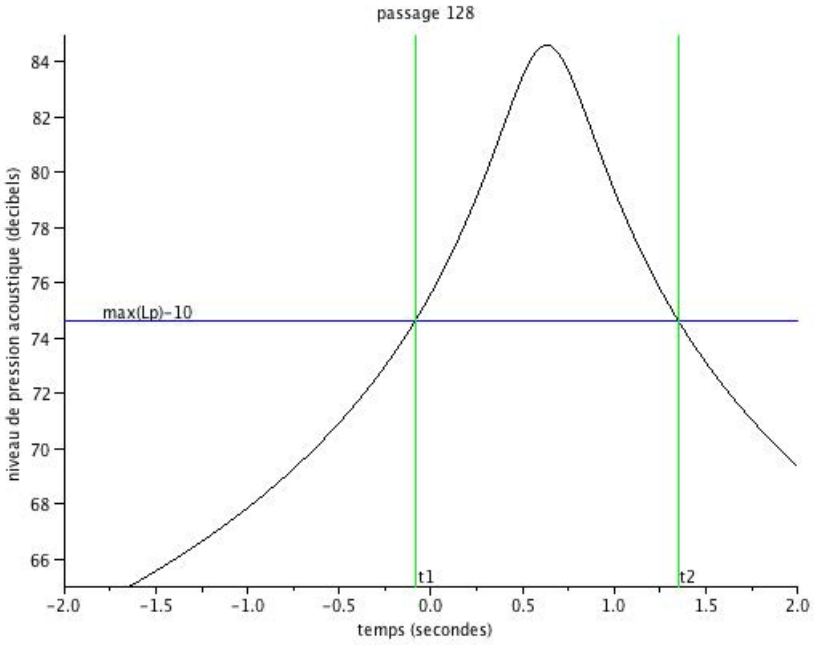


Figure 3.1 – Détermination des abscisses $t_{1,128}$ et $t_{2,128}$

• **SEL Sound Exposure Level (niveau d'exposition au bruit)**
L'indice SEL prend également en compte la hausse et la baisse du son. Cet indice indique le niveau d'une poussée de bruit hypothétique d'une seconde dans laquelle est comprimée toute l'énergie sonore du véritable événement de bruit.
L'indice SEL est donc fondé sur l'énergie et son unité de mesure est le décibel A dB(A). La pondération A prend en compte la sensibilité variable de l'oreille humaine à la fréquence sonore.et pour comparer entre eux les évènements sonores issus d'une même source. Le SEL se calcule suivant la formule :

$$SEL = 10\log_{10}[\frac{1}{T_0}\int(\frac{P_A(t)}{P_0})^2 dt]$$

- avec t = durée de l'événement exprimée en secondes .
- $P_A(t)$:pression acoustique instantanée pondérée A.
- P_0 :Pression acoustique de référence.

3.0.3 Résultats obtenus

Le calcul du SEL et L_{Aeq} sur les données de Hagunau2011 et Erstein2011, en utilisant le code scialab pour le dépouillement, j'ai présenté les nouveaux données avec les nouveaux indicateurs afin de calculer grâce à R les résidus et le test de multinormalité(m.shapiro).Les figures 3.2 et 3.3 donnent les résultats obtenus pour les deux sites de dépouillement.

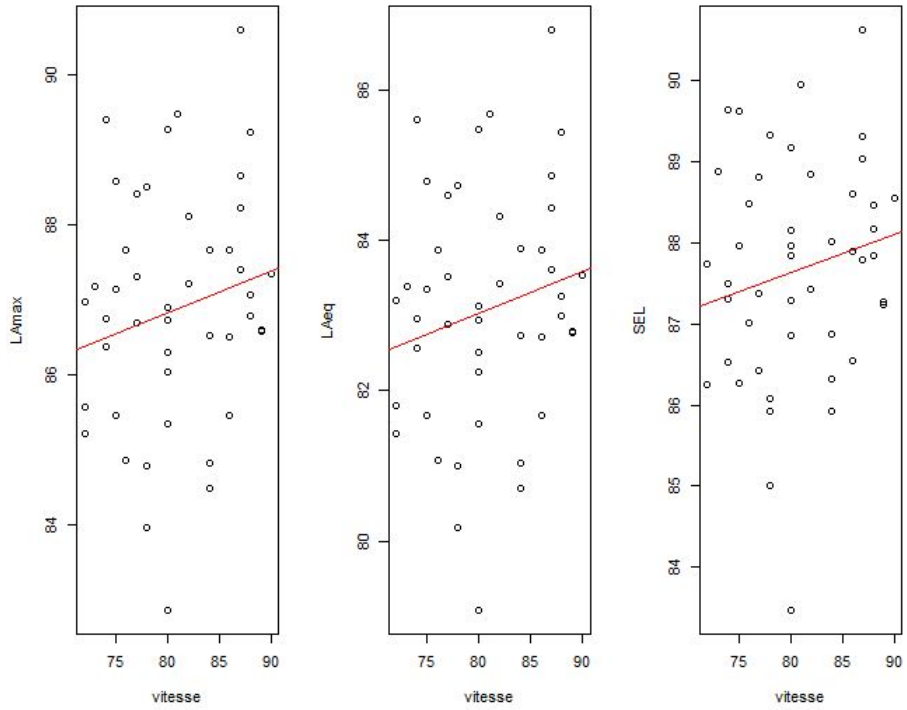


Figure 3.2 – Données Haguenau2011

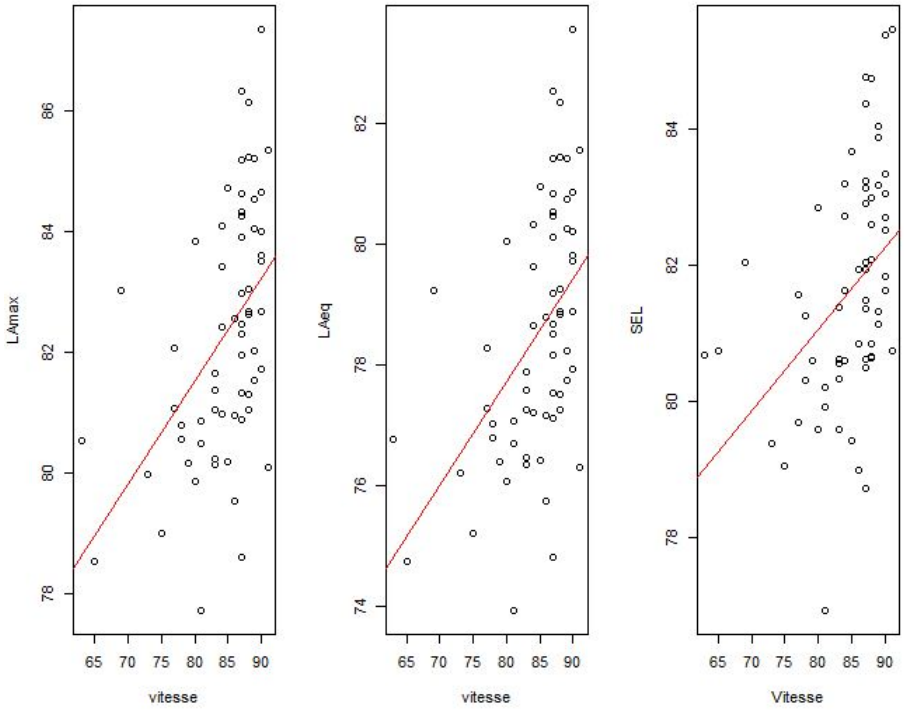


Figure 3.3 – Données Erstein2011

Après calcul des sommes de résidus pour L_{Amax}, L_{Aeq}, et SEL, les tableaux ci-dessous présente les résultats pour Haguenau2011 et Erstein2011 :

Indice	Somme des résidus Haguenau2011
L_{Amax}	111.38
LAeq	111.02
SEL	84.36

Indice	Somme des résidus Erstein2011
L_{Amax}	207.44
LAeq	207.36
SEL	152.17

On remarque qu'il n'y a pas de grands changements sur la somme des résidus entre les indices L_{Amax} et L_{Aeq}, par contre concernant le niveau d'exposition au bruit SEL, la somme des résidus est plus petite. Dans le cadre de mon stage, on voulait savoir si on obtenait la normalités des données de poids lourds, en changeant d'indice, un test de multinormalité s'est imposé pour les données de poids lourds des deux sites d'expériences.le test de multinormalité est réalisé par la commande m.shapiro.

Ci-dessous les résultats de multinormalité :
On remarque que pour les données de Haguneau2011, on obtient la normalités des données,

Indice	Weight	p-value
L_{Amax}	0.54	7.183e-12
L_{Aeq}	0.99	0.99
SEL	0.97	0.56

Table 3.1 – Test de multinormalité Haguenau2011

Indice	Weight	p-value
L_{Amax}	0.82	3.95e-07
LAeq	0.82	3.95e-07
SEL	0.81	1.45e-07

Table 3.2 – Test de multinormalité Erstein2011

mais cela n'est pas une règle générale, puisque sur les données d'Erstein2011, la p-value est plus petite que 0.05.

3.0.4 Conclusion

Les résultats obtenus en utilisant les indices L_{Aeq} ou SEL ne sont pas pertinents et le changement de l'indice L_{Amax} par L_{Aeq} ou par SEL s'avère sans utilité, en attendant de trouver une autre perspective, les travaux faits auparavant avec L_{Amax} restent exploitables .

Chapitre 4

Classification non supervisée et algorithme EM[4]

Lors du dépouillement de la campagne de mesures de bruit de roulement, l'étude est faite séparément pour les VL (véhicules légers) et les PL (poids lourds), sans oublier des données n'appartenant ni à VL ni à PL, ces dernières sont jugées aberrantes. Or si on superpose, pour une même campagne, les données VL et PL, on observe que les points aberrants de la catégorie VL se trouvent dans le nuage de points des PL, et inversement.Pour pallier au problème des points aberrants, qui peuvent provenir d'un mauvais enregistrement, ou d'un autre problème lors du dépouillement, on cherche à faire une classification non supervisée à l'aide d'un algorithme EM. Le code Scilab utilisé est celui écrit par Pierre Charbonnier, j'ai toutefois apporté quelques modifications.
Dans cette partie j'ai utilisé les données de la compagne RN592012, et les données non exploitables de 2011(Erstein2011,Matzenheim2011) pour test.
J'ai notamment rédigé des programmes R pour le test de multinormalité, l'algorithme EM :pour comparer les résultats obtenus avec le code Scilab.

4.1 Algorithme EM

L'algorithme EM(Expectation Maximization) est un outil mathématique, utilisée pour la classification non supérvisée (une définition plus élaborée est donnée en Annexe A), ce qu'on doit retenir dans le cadre de mon stage c'est qu'on a deux classes (VL et PL)qui ne sont pas nécessairement issues d'une loi normale à deux dimensions. Ces étiquettes sont ensuite oubliées pour n'avoir plus qu'une seule classe. On commence alors l'algorithme EM qui permettra après plusieurs itérations de faire une classification, le but de cet algorithme est d'assister l'opérateur en cas d'erreurs lors du dépouillement. Le travail c'est année consiste à tester cet algorithme pour les données récentes, et pouvoir conclure son efficacité.

Données RN59 Les données RN59 sont le derniers données dépouillées en juillet 2012, j'ai procédé à une vérification de l'algorithme EM pour ces données, mais j'ai obtenu une classification parfaite, et l'algorithme EM n'avait pas d'intérêt, et après discussion avec mes maitres de stage, j'ai modifié volontairement les données d'entrés, en changement les étiquettes de 4 passages, dans le but de vérifier l'efficacité de l'algorithme EM.Donc dans toute la suite de cette partie les données RN59 sont ceux qui sont modifiés.

4.1.1 Vérification de l'hypothèse de normalité

Pour appliquer une classification non supervisée à l'aide de l'algorithme EM, il faut d'abord procéder à une vérification de normalité de données , en utilisant un test de multinormalité une

modification du test de shapiro-wilk[2] unidimensionnel, qui nécessite le package "mvnormtest", on trouve :

Données RN59 2012	Weight	p-value
Catégorie VL	0.75	1.71e-08
Catégorie PL	0.76	1.04e-13

Dans les figures 4.1 et 4.2 présentent les données VL et PL de la compagne RN59 2012.

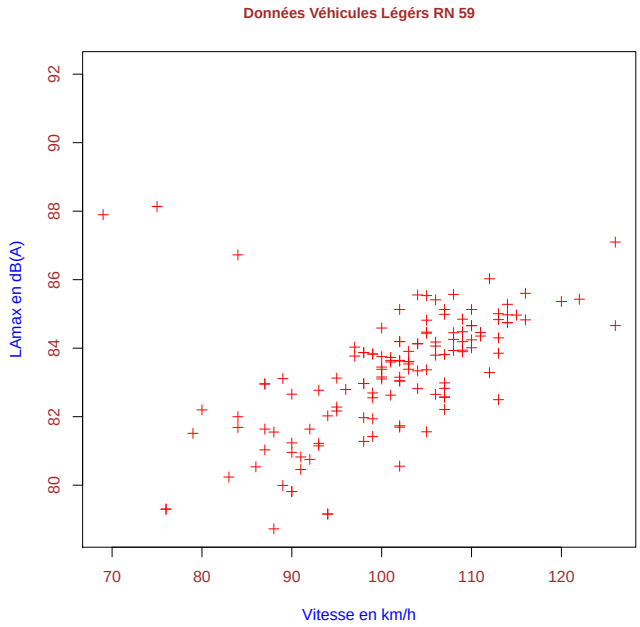


FIGURE 4.1 – Données Véhicules Légers RN59.

On arrive à percevoir dans la figure 4.1 qu'il y a 3 points mal classifiées en haut à gauche.

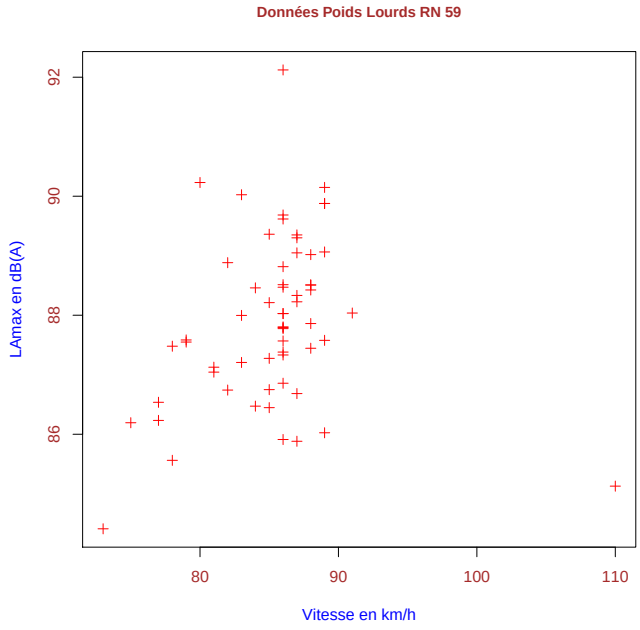


FIGURE 4.2 – Données Poids Lourds RN59.

Dans la figure 4.2, en bas à droite le point est en réalité un VL.

4.2 Sorties d'Algorithme EM

Dans les Tables ci-dessous, on a les données de RN59 avant classification(Table 4.1,Table 4.2) et en utilisant le code **Scilab** pour le calcul de la pente et de l'ordonnée à l'origine des diverses méthodes d'estimations utilisées, on obtient les résultats suivants pour les VL et PL, avant et après classification :

4.2.1 Avant classification

Pour les données VL on a :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	18.97	82.36
M-estimation Geman McClure	31.48	81.7
ACP	29.09	81.88
ACP ROBUSTE	33.97	81.55

TABLE 4.1 – Paramètres droites RN59 cat. VL avant classification EM

Pour les données PL on a :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	11.4	87.59
M-estimation Geman McClure	33.21	87.03
ACP	85.35	85.62
ACP ROBUSTE	360.39	76.09

TABLE 4.2 – Paramètres droites RN59 cat. PL avant classification EM

4.2.2 Après classification

Pour les données VL, le résultat est présentée à la Table 4.3 :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	30.17	81.66
M-estimation Geman McClure	31.83	81.68
ACP	36.22	81.35
ACP ROBUSTE	34.02	81.56

TABLE 4.3 – Paramètres droites RN59 cat. VL aprés classification EM

Pour les données PL, on a le tableau suivant :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	21.24	87.45
M-estimation Geman McClure	31.376	87.05
ACP	69	86.38
ACP ROBUSTE	263.05	79.4

TABLE 4.4 – Paramètres droites RN59 cat. PL aprés classification EM

Remarques

On remarque que dans les résultats obtenus, surtout sur les pentes d'ACP et d'ACP robuste sont supérieurs à celles obtenus avant classification, pour mes maitres de stage, on ne peut pas retenir ces méthodes, cet écart peut être expliquer par la construction de l'ACP qui nécessite une normalisation des axes.

4.2.3 Graphes avant et après classification

Après avoir utilisé le code de Scilab, on obtient des graphes avant(figure 4.3) et après classification(figure 4.4) : On remarque bien qu'il y a 3 observations de VL qui passent en PL, et une

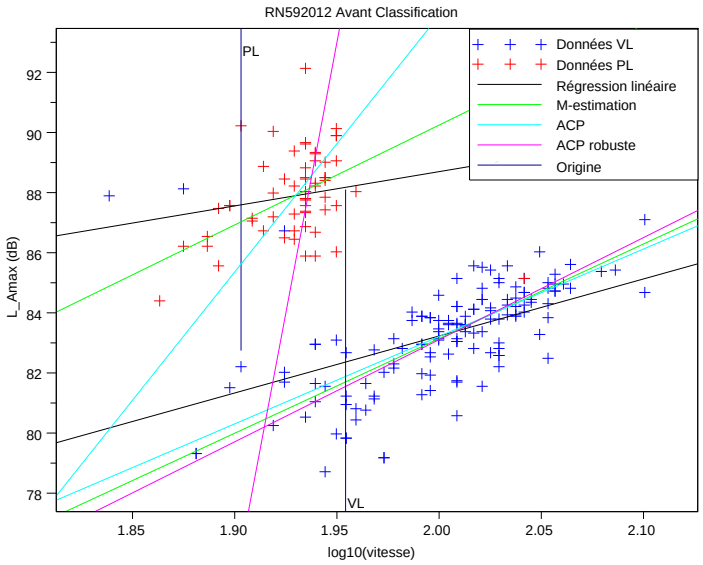


FIGURE 4.3 – Données RN59 2012 avant classification

observation PL qui passe en VL, d'où l'intérêt de l'algorithme EM.

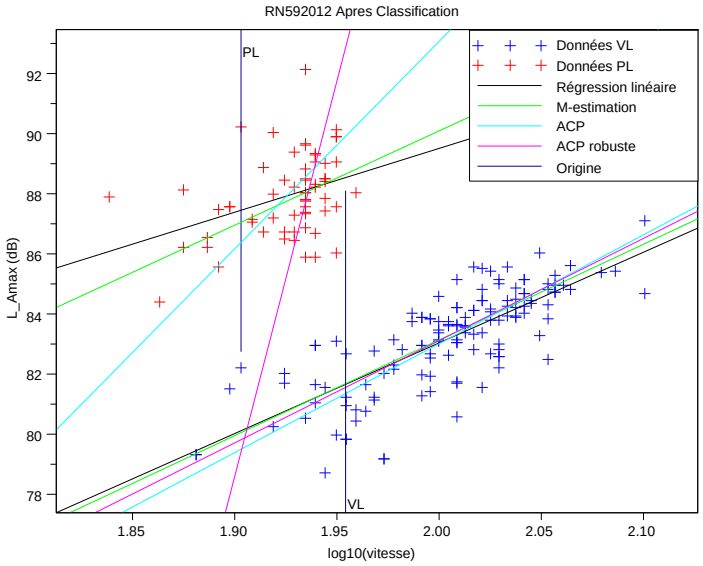


FIGURE 4.4 – Données RN59 2012 après classification

4.2.4 Ellipses d'isoprobabilité

Avec le code de Scilab déjà utilisé on obtient les ellipses d'isoprobabilités avant (figure 4.5) et après classification (figure 4.6) :

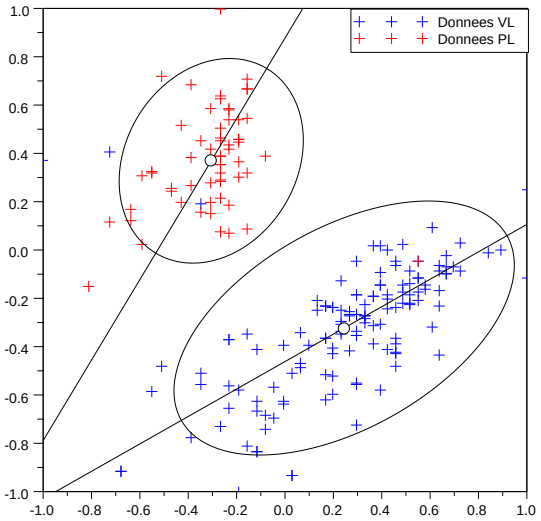


FIGURE 4.5 – Ellipse d'isoprobabilité des données RN59 2012 avant classification

On voit bien les 3 observations de Vl qui passe en PL, et l'observation PL qui passe en VL, ceci traduit la modification que j'ai effectué sur les passages.

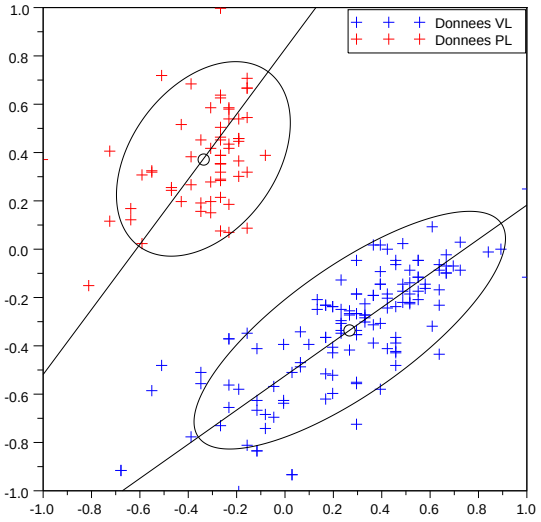


FIGURE 4.6 – Ellipse d'isoprobabilité des données RN59 2012 après classification

Remarques

L'algorithme EM a bien joué son rôle de classification .

4.2.5 Algorithme EM sur les données de Erstein2011,Matzenheim2011

Avant de procéder l'algorithme, j'ai vérifié la normalité des données issues des zones de dépouillement effectués sur Erstein et Matzenheim , en 2011.
Un premier test de mshapiro sur les données ultérieurs que se soit pour les véhicules légers ou les poids lourds nous donne le résultat suivant :

Données	weight	p-value
Erstein2011VL	0.9831	0.1849
Matzenheim2011VL	0.9666	0.00499
Erstein2011PL	0.8464	$2.003.10^{-}6$
Matzenheim2011PL	0.9209	0.00048

Pour les données de véhicules légers, on remarque qu'ils vérifient le critère de normalité, presque juste pour les données de Matzenheim2011, par contre les poids lourds ne vérifient pas le critère de normalité.

```
-------------------------CLASSIFICATION-------------------------
----------------------------------------------------------------
nombre de points initialemetnt dans la categorie VL : 111
nombre de points VL passés en PL : 1

nombre de points initialement dans la categorie PL : 63
nombre de points PL passés en VL : 0

nombre total de points: 174
nombre total de points ayant changé de catégorie : 1

les points passes de VL a PL sont les points du vecteur vl d'indice:
43


aucun point n'est passe de PL a VL
```

FIGURE 4.7 – classification

Dans la figure 4.7 y a un point VL qui est passé en PL, forcément due à une faute d'enregistrement.

4.2.6 Paramètres de droites avant classification

Pour les données VL, on a :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	25.817904	72.621131
M-estimation Geman McClure	27.197353	72.446114
ACP	34.769941	72.345071
ACP ROBUSTE	32.587567	72.234169

Pour les données de poids lourds, on a :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	30.274164	81.599767
M-estimation Geman McClure	35.278848	81.15036
ACP	66.707061	80.786706
ACP ROBUSTE	222.20439	75.04595

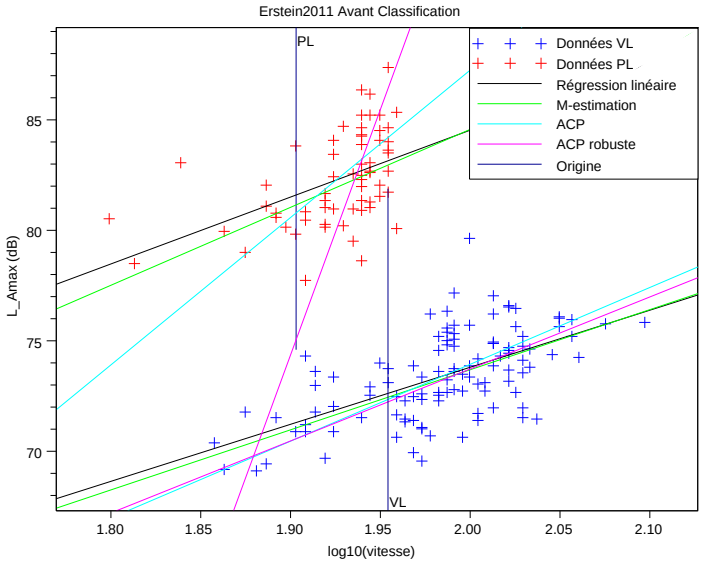


FIGURE 4.8 – Données Erstein 2011 avant classification

4.2.7 Paramètres de droite après classification

Pour les données VL, on a :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	25.446682	72.579233
M-estimation Geman McClure	27.245445	72.436733
ACP	33.01	72.34
ACP ROBUSTE	32.59	72.22

Pour les PL :

Méthode d'estimation	Pente	ordonnée à l'origine
Régression linéaire	25.33	81.63
M-estimation Geman McClure	34.62	81.15
ACP	63.9	80.74
ACP ROBUSTE	218.13	75.21

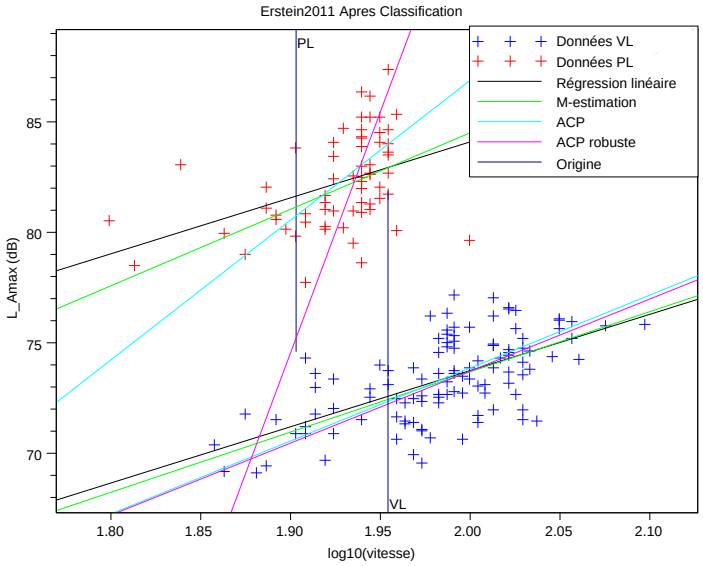


FIGURE 4.9 – Données Erstein 2011 après classification

4.2.8 Ellipses d'isoprobabilité

On trouvera ci-joint les ellipses d'isoprobabiltés avant et après classification pour les données d'Erstein2011.

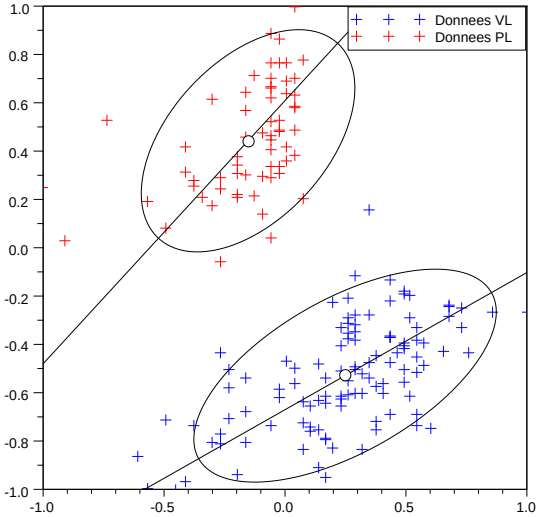


FIGURE 4.10 – Ellipse d'isoprobabilité des données Erstein 2011 avant classification

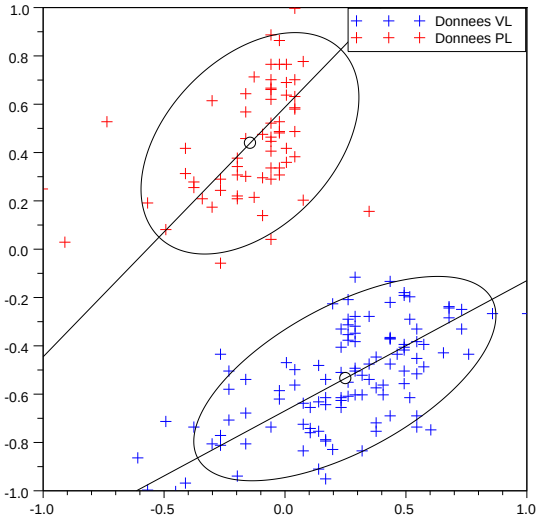


FIGURE 4.11 – Ellipse d'isoprobabilité des données Erstein 2011 après classification

J'ai effectué les mêmes opérations sur les compagnes effectués en 2011 de Haguenau, et de Matzenheim, et ce qu'on peut remarquer c'est que la qualité d'enregistrement et de dépouillement est meilleure que celle effectuée sur des données de 2009, ou on avait plus que 5 points qui changent de catégorie.L'idée est de regarder du coté des spectres, s'il y a moyen d'obtenir de meilleurs résultats en procédant à une réduction de dimensionnalité **Isomap** et un K-means.

Chapitre 5

Réduction de dimensionnalité Isomap[3]

Toujours dans l'esprit de classification non supervisée, et dans l'amélioration de cette dernière, l'alternative qui est traitée dans ce chapitre vise à pallier le problème des erreurs d'enregistrement, en travaillant sur les spectres, qui représentent l'intensité en fonction de la fréquence.
Les données qui m'ont été mis à disposition(RN59,Haguenau2011)comportait le vecteur vitesse de dimension 1, le L_{Amax}de dimension 1, et le spectre est de dimension 18, l'ISOMAP trouve son utilité pour réduire la dimensionnalité des données de spectres(dim 1 ou 2), pour pouvoir les traiter par la suite

5.1 Principe

On entend par réduction de dimensionnalité, la recherche de dimensions qui permettent d'expliquer dans la plus grande partie possible la variance observée entre les données.Cette technique est utile pour comprendre les liens entre les données, afin d'établir un modèle robuste dans le but de faire de la prédiction de classe sur de nouvelles données .
L'algorithme Isomap, qui est un algorithme de réduction de dimensionnalité non linéaire sera décrit puis appliqué aux données.

5.1.1 Algorithmes de réduction de dimentionnalité

L'algorithme **ACP**, pour Principal Component Analysis, est un algorithme de réduction de dimensionnalité qui cherche une enchâssement des données préservant au mieux la variance des points dans l'espace dimensionnel d'origine.L'algorithme ne permet d'approximer que les fonctions linéaires, l'algorithme MDS, pour Multidimensionnel Scaling trouve un enchâssement semblable au **ACP**, il y a même équivalence quand la distance est euclidienne.L'avantage de **MDS** sur **ACP** est qu'il peut utiliser n'importe quelle distance métrique.

L'algorithme **ISOMAP**, pour Complete Isometric feature Mapping, est un algorithme basé sur **MDS**mais, contrairement à celui-ci, ISOMAP permet l'approximation de fonctions non linéaires, ce qui en fait un outil plus puissant.et donc permettant de mieux approximer la dimensionnalité sous-jacente (manifold). Le concept clé différenciant **ISOMAP** des 2 autres algorithmes est qu'il utilise la distance géodésique, comme le montre la figure 7.3 :

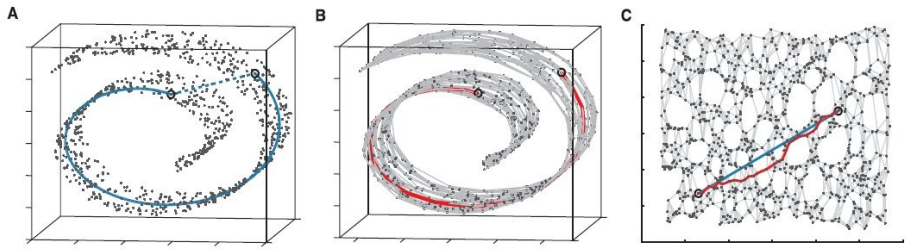


FIGURE 5.1 – Distribution de type "swiss roll", où la distance Euclidienne dans l'espace original est indiqué par le trait pointillé en A, alors que la distance géodésique est le trait plein, passant par k intermédiaires entre le point de départ, i, et le point final j.

La distance géodésique est en fait une approximation de la distance Euclidienne dans le sous espace dimensionnel des données. Celle-ci est calculée comme étant la somme des distances entre les points sur le chemin reliant i à j.

5.1.2 Le Multi-Dimensional Scaling (MDS)

-Parfois on a pas les coordonnées des exemples, mais seulement les distances(ou autre mesure de similarité) entre chaque paire d'exemples **MDS** classique trouve une présentation des exemples qui correspond exactement à la **ACP**, mais en partant de ces distances D_{ij}.

-L'algorithme est le suivant :
-Moyennes par rangées : $\mu_i = 1/n \sum_j D_{ij}$.
-Double centrage (distance vers produit scalaire) : $P_{ij} = -1/2(D_{ij} - \mu_i - \mu_j + 1/n \sum_i \mu_i)$
-Calcul des vecteurs propres v_j et valeurs propres λ_j principales de la matrice P (avec λ_j^2 plus grand).
-La i-ème coordonnée réduite de l'exemple j est $\sqrt{\lambda_i}$

5.1.3 Algorithme ISOMAP

[1] -Cet algorithme se base sur les relations linéaires locales entre voisins pour capturer la structure de la variété, mais il a aussi une composante "globale", en essayant de préserver les distances **le long de la variété**.Pour cela on essaie d'approximer la **distance géodésique** sur la variété par la distance minimale dans un graphe dont les nœuds sont les exemple-set les arcs seulement entre voisins sont associées aux distances locales.
-L'algorithme est le suivant :

- Calculer les m plus proches voisins de chaque exemple, avec les distances $d(x_i, x_j)$ correspondantes, pour peupler le graphe.

- Calculer la longueur $D(x_i, x_j)$du chemin le plus court dans le graphe entre chaque paire d'exemples :
  $D(x_i, x_j) = min_p \sum_k d(p_k, p_{k+1})$ ou p est un chemin $(p_1, p_2, ...p_l)$ entre x_i et x_j dans le graphe $(p_1 = x_i, p_l = x_j)$.
- Appliquer l'algorithme **MDS** sur la matrice des distances géodésiques , $D_{ij} = D(x_i, x_j)$, ce qui donne les coordonnées réduites pour les exemples d'apprentissage.

5.1.4 Application de l'algorithme ISOMAP[3] sur les données RN59

Sur les données de RN59[12], j'ai rédigé un script sur Scilab pour tirer les valeurs des spectres (L3fmax), et en utilisons la fonction **Isomap** sous le logiciel **R** nécessitant le package vegan, la

sortie donne la figure 7.3 :

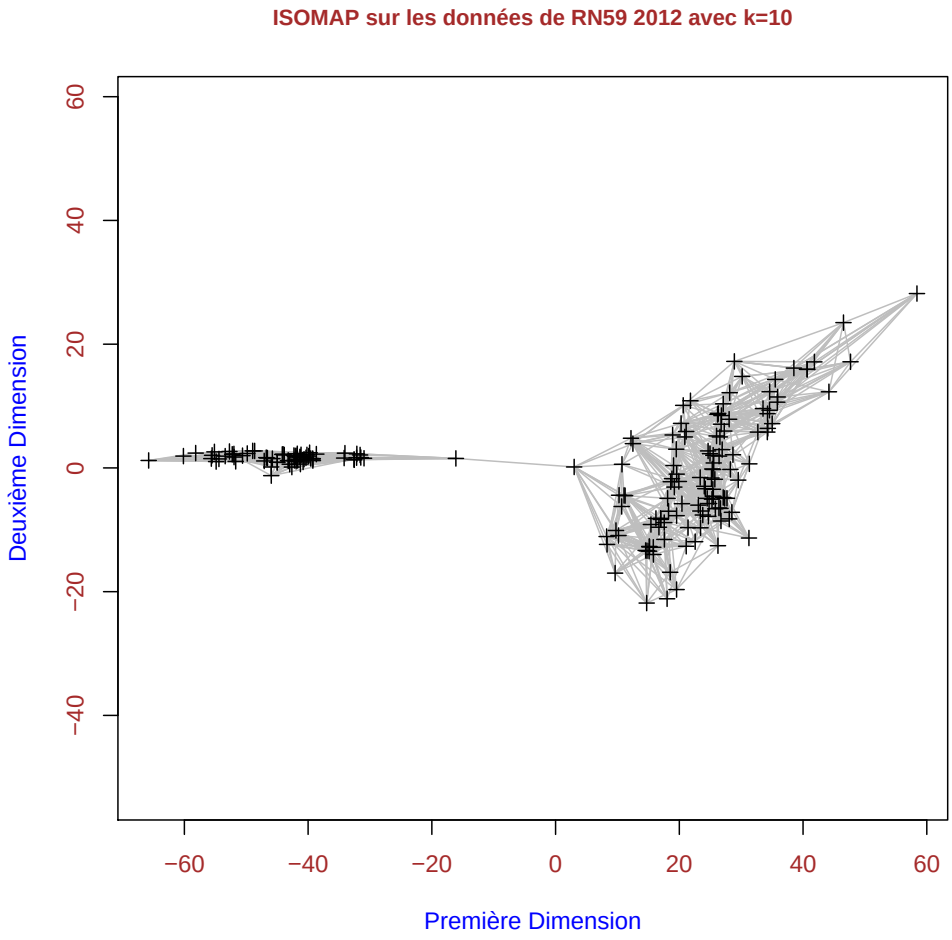


FIGURE 5.2 – Représentation ISOMAP avec k=10 pour les spectres en 1/3 d'octave de RN59 2012

- k=nombre de plus proches voisins pour un point

5.1.5 Remarques

Pour cette présentation d'Isomap, on voit bien que les points sont regroupées en deux noyaux, On peut bien distinguer que les données sont classés en deux catégories .Reste à savoir si les deux catégorie correspondent aux classes VL et PL.

5.1.6 THE MANI GUI

Une autre alternative est celle introduite via le logiciel Matlab à l'aide d'un programme "THE MANI GUI"c'est un programme qui incorpore des algorithmes de réduction de dimensionnalité à savoir Isomap, ACP, MDS..afin de choisir la meilleure méthode de réduction de dimensionnalitéles, je l'ai testé toutefois sur les données de spectres de "RN59", voici les résultats obtenus :

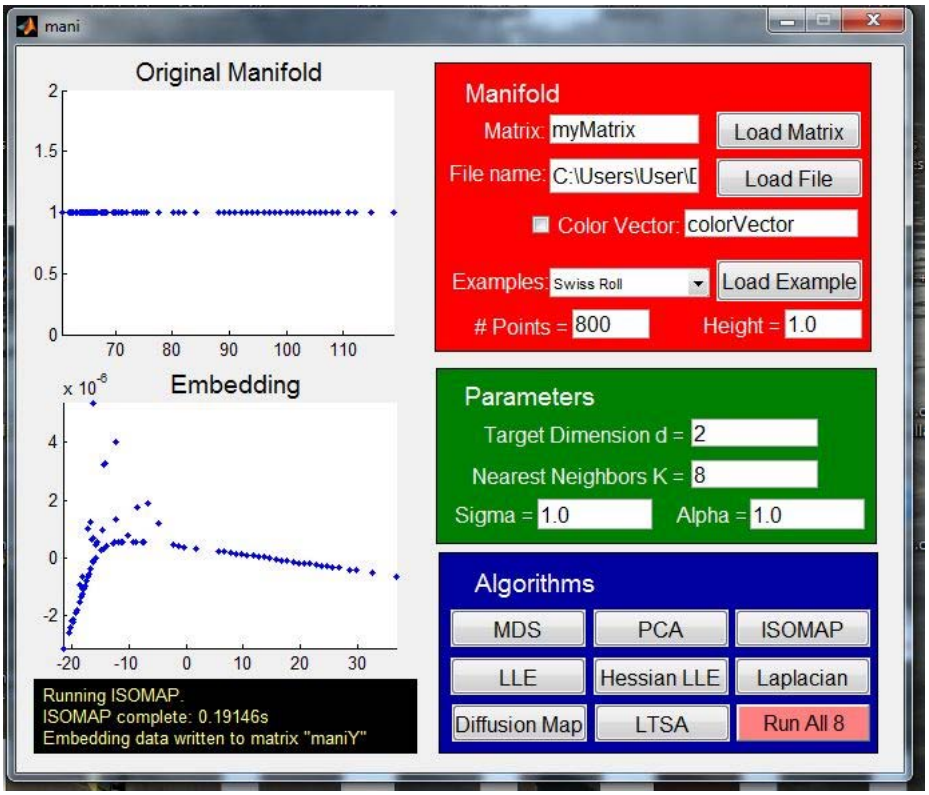


FIGURE 5.3 – Sortie programme THE MANI GUI pour les données " RN59"

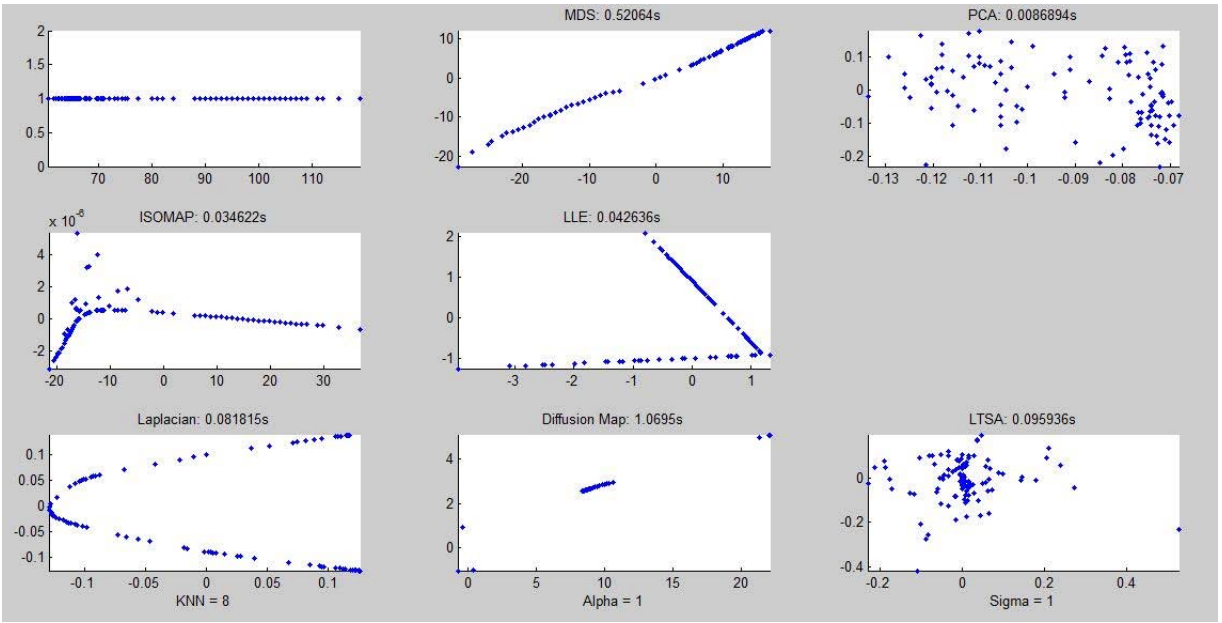


FIGURE 5.4 – Résultats de différents algorithmes de réductions pour " RN59"

Remarques

Cette méthode est certes intéressante, mais Certaines applications prennent un certain temps à s'exécuter, en particulier MDS et Isomap. Si l'ensemble de données est grande ou de grande dimension, la méthode peut prendre plusieurs minutes pour s'exécuter. Lorsque la méthode est terminée, la matrice de données est écrit dans l'espace de travail et le plongement est tracée dans le graphique(en haut).

Avec ce programme, on arrive également à distinguer deux ensembles distincts pour l'Isomap, maintenant pour choisir une méthode de réduction plus qu'une autre, y a d'autres critère qui entre en jeux à savoir la rapidité d'exécution, l'uniformité du graphe, la non convexité, la géométrie du collecteur...
Après avoir réalisé un Isomap sur les données, le chapitre 6 présentera l'idée d'un k-means sur les données issus de l'Isomap.

5.2 Regroupement et similitude"Clustering"

Le présent paragraphe vise la construction d'un classificateur intrinsèque, sans classes connues a priori, par apprentissage à partir d'échantillons donnés. Le nombre de classes possibles est en général connu, on veut construire le classificateur K : $M \rightarrow (1..c)$ partitionnant l'espace M des mesures. On fait l'hypothèse que plus deux échantillons sont proches de M, plus leur probabilité d'appartenir à la même classe est grande.
Il s'agit donc de déterminer le partitionnement en classes, ainsi que les critères d'appartenance d'un échantillon à la classe la plus probable, en partant uniquement d'une probabilité d'appartenance à une même classe (sans toutefois préjuger de laquelle), c'est-à-dire à partir d'une mesure de ressemblance entre échantillons.
On désigne aussi par regroupement ou clustering les méthodes d'inférence d'une classification. Les classes ou catégories obtenues sont appelées groupes ou agrégats. Nous verrons deux familles d'algorithmes de regroupement procédant l'une par optimisation itérative d'une classification donnée l'une et l'autre par agglomération hiérarchique d'échantillons partiellement agrégés à chaque niveau

5.2.1 Formulation mathématique du problème

Mathématiquement, le problème de l'apprentissage non supervisé peut être formulé comme l'approximation d'une distribution multimodale. On peut donc faire l'hypothèse que le nombre c de modes est connu, ainsi que la loi uni-modale de chaque mode. Toutefois les méthodes connues des moindres carrés ou d'estimation Bayésienne ne fournissent pas de résultats pratiques : la solution analytique n'est connue que dans des cas triviaux tandis qu'en règle générale les calculs effectifs croissent exponentiellement avec la taille de l'échantillon d'apprentissage On peut résumer le problème de la manière générale suivante : supposons donné un n-échantillon $(x_1, ..., x_n)$ dans un espace de mesures M quelconque, ainsi qu'une "dissemblance" (dissimilation measure) réelle ρ elle que :
$\rho(x, y) \leq 0$ pour x, y dans M. On cherche une classification K :$M \rightarrow$(0...c) regroupant les échantillons en c classes (c peut être considéré comme donné, son choix sera discuté ultérieurement) telle que la dissemblance $\rho(x, y)$ soit plus petite pour $K(x) = K(y)$ que pour $K(x) \neq K(y)$. Ce critère peut être précisé de diverses manières ; dans les méthodes hiérarchiques on exige qu'au niveau donné par le seuil θ, toute paire x, y satisfait la relation : $\rho(x, y) \leq \theta$ si et seulement si $K(x) = K(y)$. Dans les méthodes d'optimisation itérative on exige que la moyenne des $\rho(x, y)$ intra-classe (c'est-à-dire avec($K(x) = K(y)$) soit minimale relativement à la moyenne inter-classe (avec $K(x) \neq K(y)$).

Chapitre 6

Classification non supervisée avec Isomap et K-means[8]

6.1 K-means"Clustering"

Dans l'exploration de données , k-means est une méthode de l'analyse par grappes , qui vise à la partition n observations en k grappes dans lequel chaque observation appartient à l'amas avec la plus proche moyenne . Il en résulte un cloisonnement de l'espace des données dans les cellules de Voronoï(Annexe A).

6.1.1 Description

Étant donné un ensemble d'observations $(x_1, x_2, ..., x_n)$, où chaque observation est un vecteur d-dimensionnel réel, k-means vise à partitionner les n observations en k ensembles $(k \leq n)$ $S = S_1, S_2, ..., S_k$ de manière à minimiser la somme intragrappe des carrés (WCSS) :

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

6.1.2 Algorithme K-means

L'algorithme des k-moyennes ou K-means est en statistiques, et en apprentissage automatique (plus précisément en apprentissage non supervisé), un algorithme de partitionnement de données, c'est-à-dire une méthode dont le but est de diviser des observations en K partitions (clusters) dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche. Les nuées dynamiques sont une généralisation de ce principe, pour laquelle chaque cluster est représenté par un noyau pouvant être plus complexe qu'une moyenne. L'algorithme classique de K-means est le même que l'algorithme de quantification de Lloyd-Max.
Algorithme standard

- Choisir k moyennes $m_1^{(1)}, \ldots, m_k^{(1)}$ initiales (au hasard par exemple)
- Répéter jusqu'à convergence :
  -assigner chaque observation à la moyenne la plus proche (i.e effectuer une partition de Voronoï selon les moyennes).

  $$S_i^{(t)} = \{x_j : \|x_j - m_i^{(t)}\| \leq \|x_j - m_{i^*}^{(t)}\| \ \forall i^* \text{ dans } [1, k]\}$$

  -mettre à jour la moyenne de chaque cluster

  $$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i} x_j$$

  La convergence est atteinte quand il n'y a plus de changement

6.1.3 Avantages et Inconvénients

Un inconvénient de cet algorithme est que les clusters dépendent de l'initialisation et de la distance choisie. Le fait de devoir choisir déterminer a priori le paramètre k peut être perçu comme un inconvénient ou un avantage.

6.2 Application dans le stage

Comme c'est expliqué précédemment l'algorithme K-means vise à partitionner les données en grappes, l'idée serait d'appliquer cet algorithme sur les données produits après réduction de dimmensionnalité, reste à spécifier les classes en regroupant les observations appartenant à l'ensemble qui a la plus proche distance du clusters, autrement dit pour savoir qu'une donnée appartient à une classe plus qu'une autre il faudra calculer sa distance avec les deux clusters et la regrouper avec l'ensemble de plus petite distance.

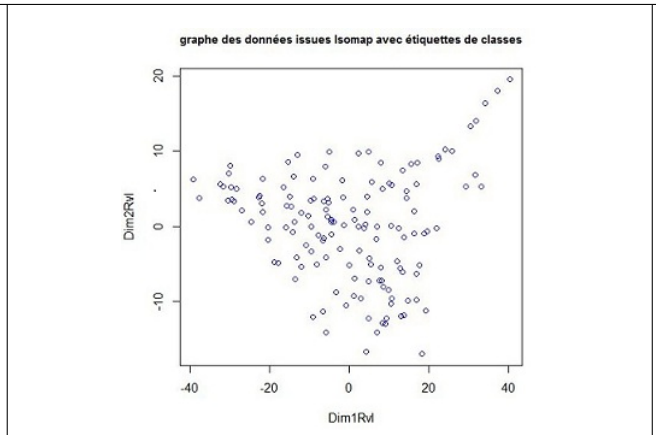
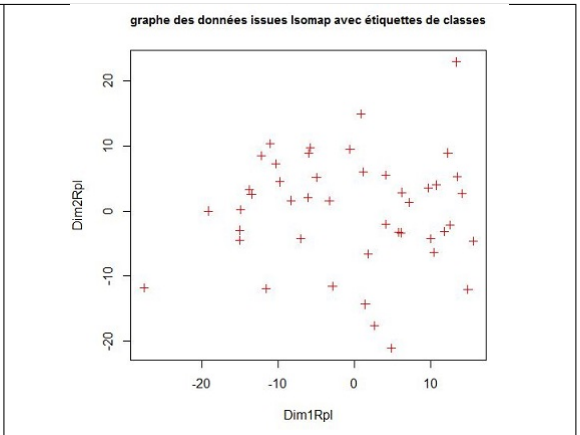
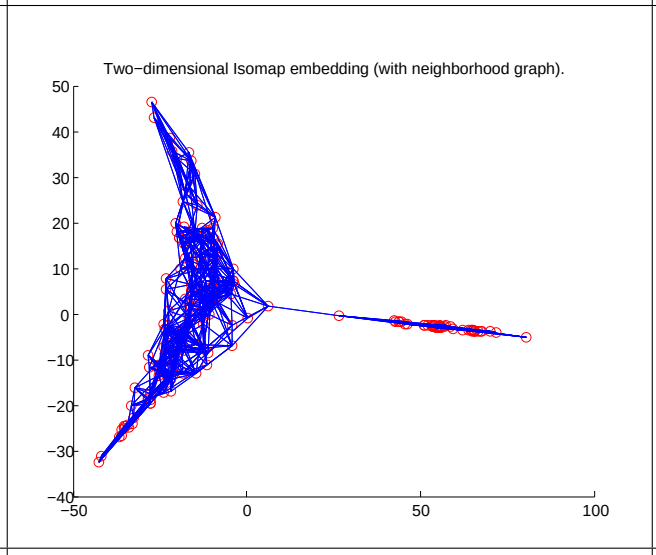
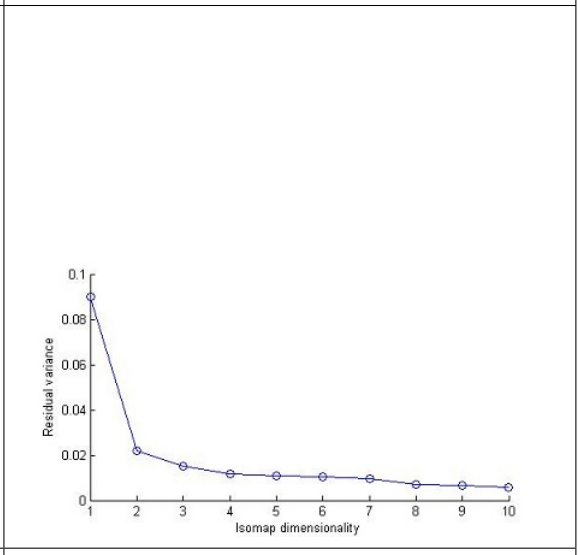
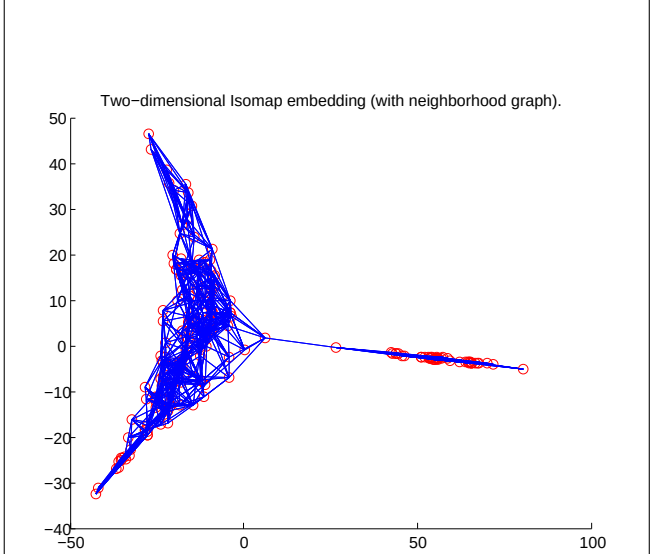
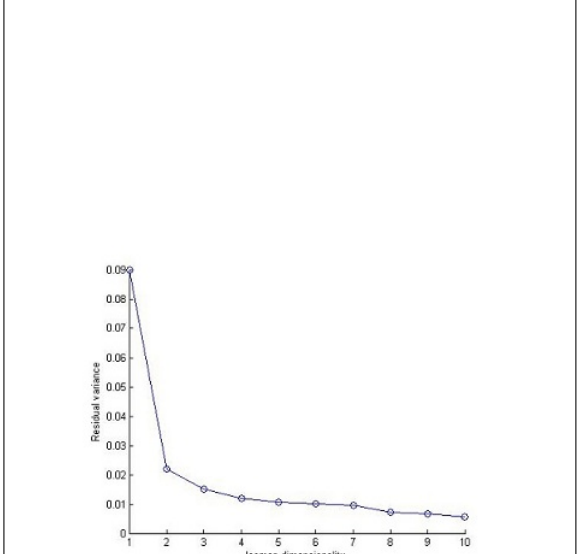
Fichiers textes

Les données étaient disponibles au format binaire. Après chargement dans Scilab et traitement des vitesses, j'ai obtenu des fichiers texte décrivant ces données. En fait, ce sont des tableaux dont les colonnes sont respectivement la vitesse, le paramètre adimensionnel de vitesse (variable X) , le niveau de bruit L_{Amax}, et le spectre $L3F_{max}$, c'est ce dernier paramètre sur lequel on va travailler, pour voir s'il y a moyen de réaliser une classification à partir des spectres, pour cela une réduction de dimensionnalité s'impose, le but serait de faire un K-means sur les données issus de l'Isomap.

6.3 Planches

A partir de ces fichiers, des planches ont été faites afin de comparer les différentes résultats obtenus, en gardant le vecteur vitesse, ou pas, en travaillant sur les données normalisés ou pas, en faisant un Isomap en premier lieu et un K-means en deuxième lieu, les résultats obtenus sont issus des données de Haguenau2011.et les fonctions Isomap et K-means ont été réalisés sous R et Matlab, le code Matlab de K-means était réalisé par Doulaye Dembele. J'ai rapporté toutefois, quelques modifications.

6.3.1 Planches ISOMAP

ISOMAP sur les données vl		ISOMAP sur les données pl	
ISOMAP sur les données brutes avec k=8		Dimension Isomap	
Isomap sur les données normalisées avec k=8		Dimension Isomap	

Remarques

D'après les résultats obtenus dans la planche ci-dessus, on remarque bien qu'on arrive à distinguer deux groupes et en effectuant un Isomap sur les données de spectres brutes ou normalisés, ce donne le même résultat, donc on peut se passer de la normalisation, Ceci peut être expliqué

par le fait que l'algorithme MDS, on normalise déjà les données.Aussi, on peut remarquer qu'avec une réduction en dim 2 on ne perd pas plus d'information qu'en dimension 10 ou moins.

Irréversibilité des résultas : En essayant de regrouper les données VL et PL issus de l'Isomap, pour retrouver le résultat obtenu sur les données brutes, cela s'avère irréalisable, et c'est dû principalement aux vecteurs propres de VL et PL qui ne sont pas pareils après réduction de dimensionnalité, et ainsi une projection n'est pas possible, voici toutefois le résultat obtenu à la figure 8.1 :

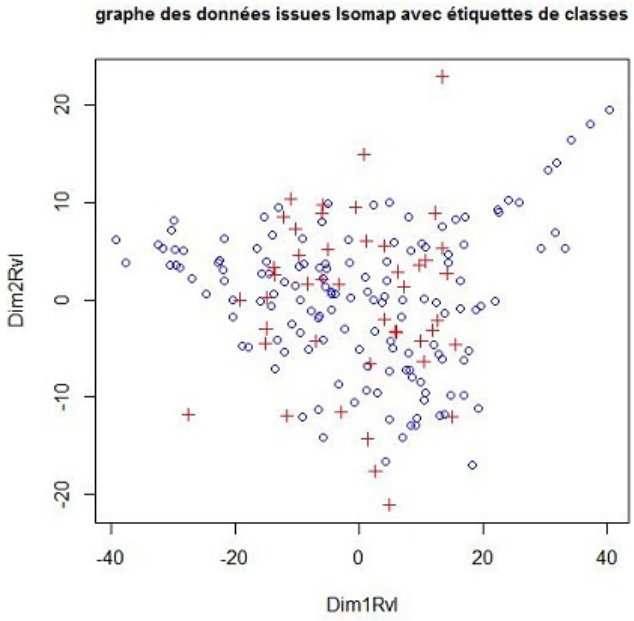


FIGURE 6.1 – Regroupement Isomap vl et pl

6.4 K-means

En effectuant un k-means sur les donnés issus de l'Isomap de Haguenau2011 (figure6.2), on arrive à distinguer deux classes (VL et PL), ainsi que les clusters ,rappelant que ce résultat a été obtenu après 25 itérations.

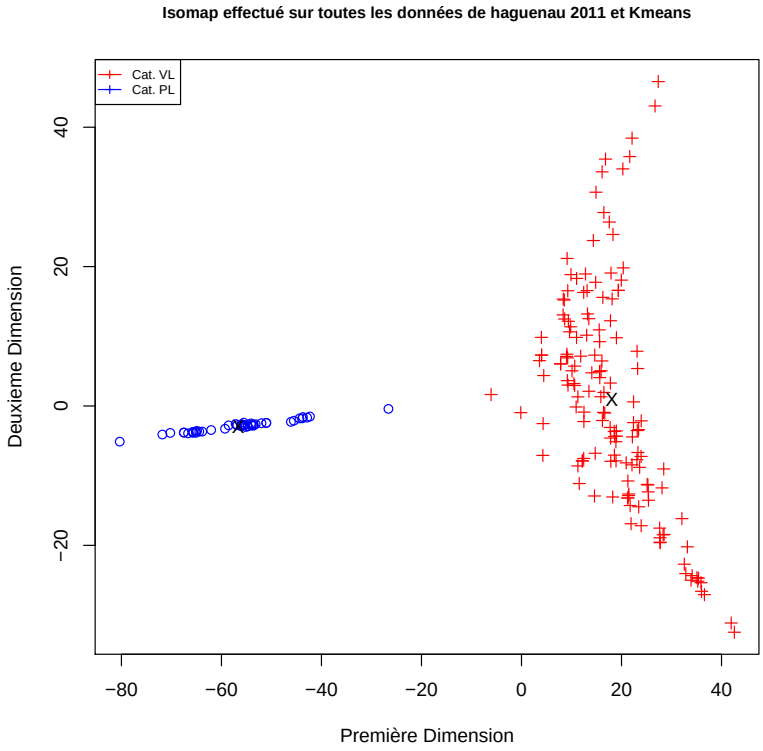
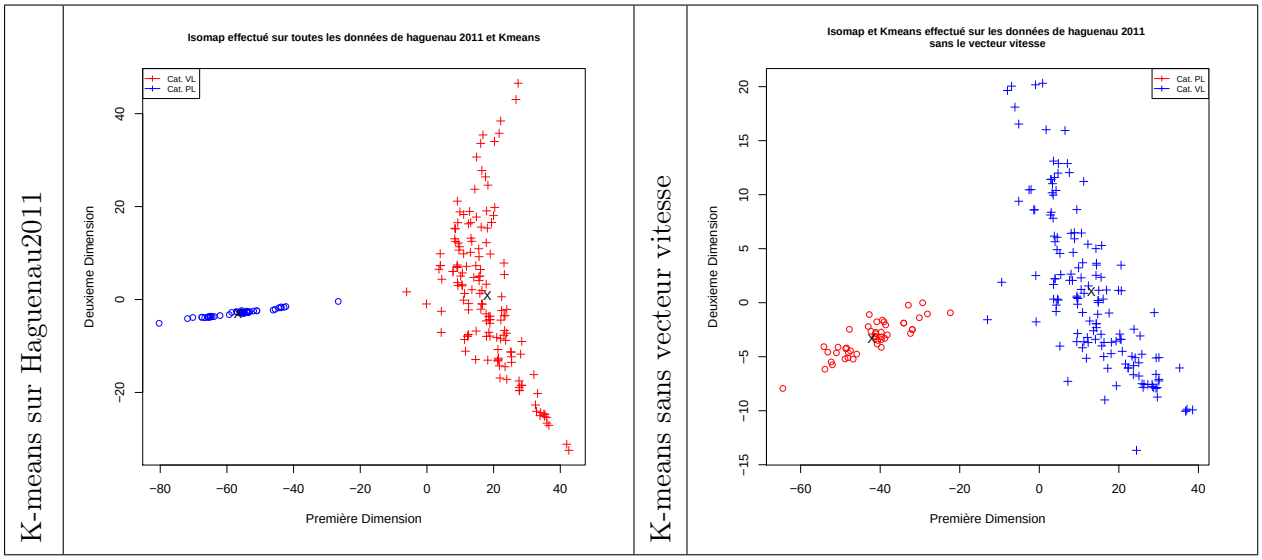


FIGURE 6.2 – Sortie k-means sur les données Haguenau2011 issues de l'isomap

Peut-on effectuer une classification indépendamment du vecteur vitesse ?



Remarques Ce qu'on peut remarquer c'est que le vecteur vitesse n'est pas indispensable, une différence d'une observation(voir AnnexeB), ce résultat est donc intéressant dans la mesure ou il permet une classification indépendamment de la vitesse, cette information qui demande un grand travail à l'opérateur.

Différentes essais

– En rajoutant le vecteur vitesse sur les données issus de l'Isomap(on fait donc une projection sur une seul vecteur propres) et en rajoutant le vecteur vitesse, on obtient la figure6.3 , ce qui intéressant c'est qu'on retrouve les mêmes observations qu'avec un k-means effectué sur toutes les données.

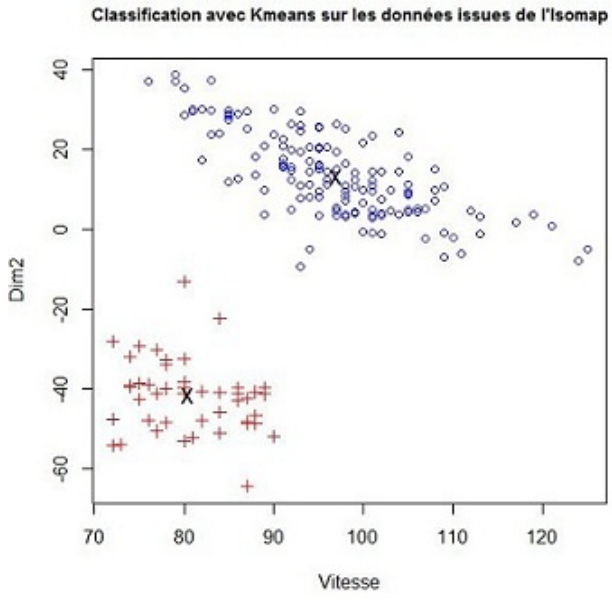


FIGURE 6.3 – k-means sur les données en rajoutant le vecteur vitesse

– En exécutant k-means plusieurs fois jusqu'à 25 itérations, on obtient les deux classes, avec inversion des étiquette. Le résultat est présenté dans la figure 6.3 :

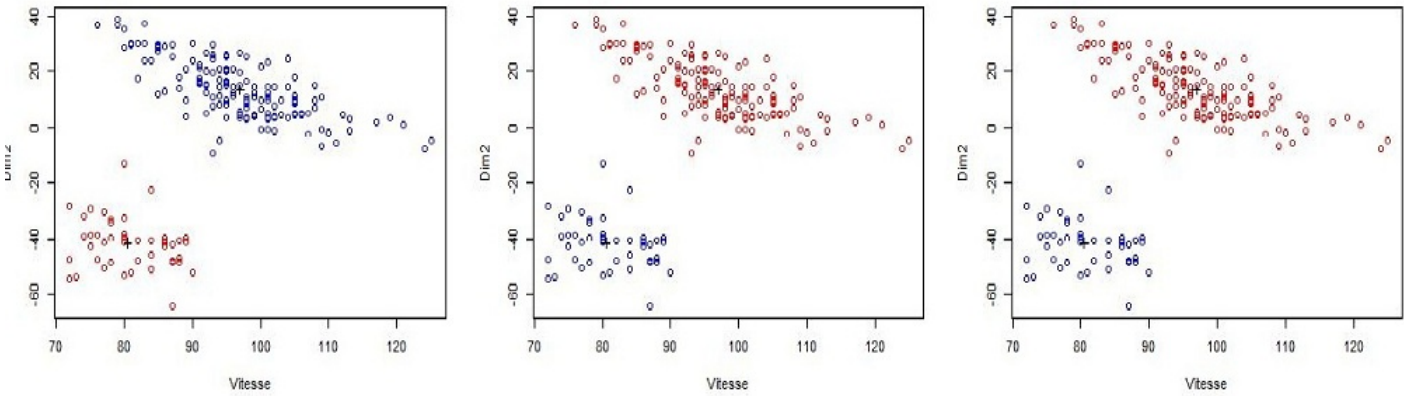


FIGURE 6.4 – Sortie K-means après plusieurs itérations sans vitesse

Comparaisons des résultats obtenus avec le k-means et le résultat du dépouillement

En utilisant les résultats de classification réalisés par l'opérateur dans dBeuler[11], et en comparant avec les résultats obtenus, avec l'Isomap et le k-means, on trouve une différence d'un point qui change de catégorie "l'observation 165", vous trouverez en annexe B les données obtenus par

l'opérateur et ceux après la classification. En appliquant un Isomap et K-means sur les données sans l'observation 165, on obtient la table 6.1 :

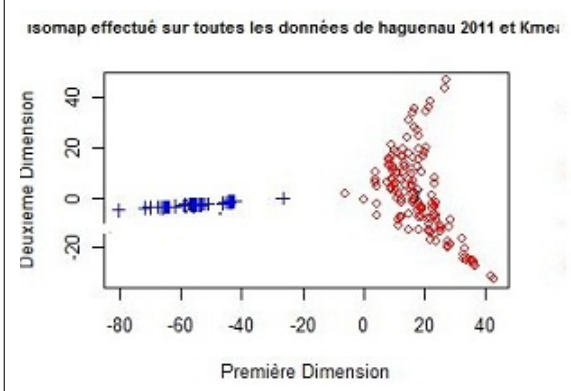
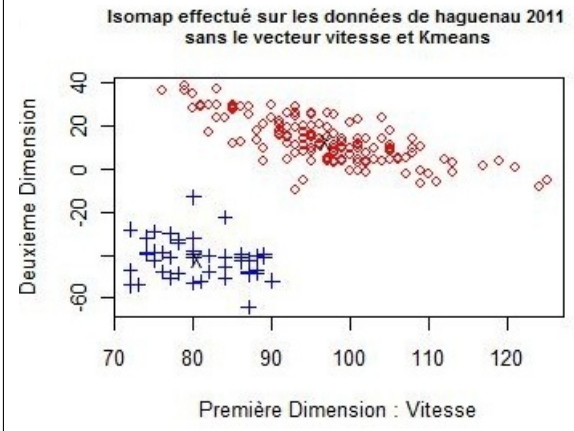
Isomap sur les données sans l'observation 165		K-means sur les données sans l'obs 165	

TABLE 6.1 – Isomap et k-means sans l'observation 165

6.4.1 Conclusion et Perspectives

La classification par le k-means, donne de bons résultats, on arrive à distinguer deux groupes , qui donne à une erreur prés pour les cas des données de Haguenau2011 les mêmes résultats obtenus par l'opérateur.Reste maintenant aux acousticiens de savoir si cette méthode de classification est pertinente.Le fait de pouvoir faire la classification indépendamment du facteur de la vitesse est un plus, la question maintenant est de savoir si cela est suffisant.

Chapitre 7

Conclusions Générale

Différentes méthodes ont été testé pour départager les droites d'estimations, malheureusement les résultats ne sont pas probants, malgré que le travail effectué est sur une petite tranche de vitesse, Cependant l'opérateur pourra utiliser n'importe quelle méthode d'estimations vues.

Le dépouillement des poids lourds présentait aussi le problème de dispersion de nuage des points, après avoir travaillé sur d'autres indicateurs SEL et le L_{Aeq} à la place du L_{Amax}, les résultats ont montré qu'un changement d'indicateur est sans utilité. Ceci conforte l'idée que les résultats trouvés en utilisant L_{Amax} sont exploitables

La classification non supervisée, qui a été appliquée avec l'algorithme EM, a porté ses fruits. Cela pourra permettre d'assister l'opérateur dans le dépouillement des campagnes de mesure. Après intégration de cette méthode dans dBEuler, l'opérateur pourrait s'appuyer sur la classification proposée par l'algorithme pour détecter d'éventuelles erreurs ou éliminer certains passages. Une

autres méthode de classification non supervisé " K-means " a été testé après une réduction de dimensionnalité "Isomap" sur les données de spectres et a donnée de bon résultats. Le fait aussi de pouvoir classifier indépendamment du facteur de la vitesse est un avantage pour l'opérateur ,contrairement à la classification avec l'algorithme EM

Maintenant il faudra savoir choisir quelle méthode de classification non supervisée entre l'algorithme EM et l'algorithme du K-means après réduction de dimensionnalité.Dans le cas ou le choix est celui du K-means ,il faudra penser à l'intégrer également dans dBEuler.

Annexe

7.1 Annexe A

7.1.1 Outils Mathématiques

M.estimateur[6]

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 fonctionρ sur l'ensemble des données. Ainsi, le (ou les) M-estimateur associé aux données et à la fonctionρ 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.

7.1.2 Construction d'un M.estimateur

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.
lorsque les variables erreurs ϵ_i sont distribuées indépendamment selon une loi normale $N(0, \sigma^2)$, 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 :

$$\exp -\frac{1}{2\sigma^2}||Y - M.f||^2 \tag{7.1}$$

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

$$\hat{f}^{MV} = \underset{f \in \mathbb{R}^m}{\arg\max} ||Y - M.f||^2 \tag{7.2}$$

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

$$\hat{f}^{MV} = \underset{f \in \mathbb{R}^m}{\arg\max} ||r(f)||^2 \tag{7.3}$$

L'idée des M-estimateurs alors, est de réduire l'incidence de telles observations. Au lieu d'utiliser la fonction x^2, on utilise la fonction ρ, et chercher à minimiser en f la quantité :

$$\sum_{i=1}^{n} \rho(r_i(f))$$

M.estimateur Geman et McClure

Le M.estimateur Geman et McClure est donnée par la fonction définit par :

$$\rho_{GM}(r) = \frac{r^2}{1+r^2} \tag{7.4}$$

Présentation de la fonction Geman et McClure (Figuree 7.1)

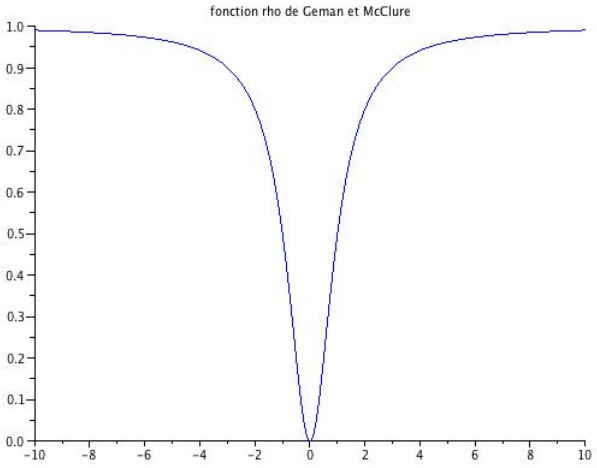


FIGURE 7.1 – Geman et McClure

Dans les calculs du M estimateur effectué dans le chapitre 2, la fonction utilisée est celle de Geman et McClure.

7.1.3 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 :

$$l(y_1, \cdots, y_n; \theta_i, \sigma) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta_i)^2 \tag{7.5}$$

Et en prenant un minimum en σ de la quantité :

$$e_\sigma = n\ln(\sigma) + \frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i^2 \tag{7.6}$$

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 :

$$e_{M,\sigma} = n\ln\sigma + \frac{1}{2}\sum_{i=1}^{n}\rho\left(\frac{r_i}{\sigma}\right) \tag{7.7}$$

7.1.4 Algorithme EM[9]

L'algorithme espérance-maximisation (en anglais Expectation-maximisation algorithm, souvent abrégé EM), proposé par Dempster et al. (1977)1, est une classe d'algorithmes qui permettent de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables

Principe de fonctionnement En considérant un échantillon $X = (x_1, ..., x_n)$ d'individus suivant une loi $f(x_i, \theta)$ paramétrée par θ, on cherche à déterminer le paramètre maximisant la log-vraisemblance donnée par

$$L(x;\theta) = \sum_{i=1}^{n} log f(x_i, \theta)$$

Cet algorithme est particulièrement utile lorsque la maximisation de L est très complexe mais que, sous réserve de connaître certaines données judicieusement choisies, on peut très simplement déterminer θ.
Dans ce cas, on s'appuie sur des données complétées par un vecteur inconnu. En notant $f(z_i \backslash x_i, \theta)$ la probabilité de z_i sachantx_i et le paramètre θ, on peut définir la log-vraisemblance complétée comme la quantité

$$L((x,z),\theta) = \sum_{i=1}^{n} (log(f(z_i \backslash x_i, \theta) + log f(x_i;\theta))$$

et donc :
$L(x;\theta) = L((x,z);\theta) - \sum_{i=1}^{n} log(f(z_i \backslash x_i, \theta)$
L'algorithme EM est une procédure itérative basée sur l'espérance des données complétées conditionnellement au paramètre courant. En notant ce paramètre $\theta^{(c)}$, on peut écrire

$$E[L((x,z);\theta] = E[L((x,z);\theta) \backslash \theta^{(c)})] - E[\sum_{i=1}^{n} log(f(z_i \backslash x_i, \theta^{(c)})]$$

ou encore :
$L(x,\theta) = Q(\theta, \theta^{(c)}) - H(\theta, \theta^{(c)})$
avec $Q(\theta, \theta^{(c)}) = E[L((x,z);\theta) \backslash \theta^{(c)})]$,et $H(\theta, \theta^{(c)}) = E[\sum_{i=1}^{n} log(f(z_i \backslash x_i, \theta^{(c)})]$
On montre que la suite définie par
$\theta^{(c+1)} = \underset{\theta}{\operatorname{argmin}}(Q(\theta, \theta^{(c)}))$
fait tendre $L(x;\theta^{(c+1)})$ vers un maximum local
on peut définir l'algorithme EM de la manière suivante :

- Initialisation au hasard de $\theta^{(0)}$
- $c = 0$
- Tant que l'algorithme n'a pas convergé, faire
- Evaluation de l'espérance (étape E) :$Q(\theta, \theta^{(c)}) = E[L((x,z);\theta) \backslash \theta^{(c)})]$
- Maximisation (étape M) :$Q(\theta, \theta^{(c)}) = E[L((x,z);\theta) \backslash \theta^{(c)})]$
- Méthodes $\theta^{(c+1)} = \underset{\theta}{\operatorname{argmax}}(Q(\theta, \theta^{(c)})$
- c=c+1
- Fin

En pratique, pour s'affranchir du caractère local du maximum atteint, on fait tourner l'algorithme EM un grand nombre de fois à partir de valeurs initiales différentes de manière à avoir de plus grandes chances d'atteindre le maximum global de vraisemblance.

7.1.5 Analyse en Composantes Principales (ACP)[7]

Introduction

L'ACP est une Méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes.

Lorsqu'on étudie simultanément un nombre important de variables quantitatives(ne serait-ce que 4 !), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales. C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique adéquat, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus selon l'ensemble des variables initiales (ainsi remplacées par les facteurs).

principe de l'ACP

Les données sont les mesures effectuées sur n unités $u_1, u_2, ..., u_i, ...u_n$. Les p variables quantitatives qui représentent ces mesures sont $v_1, v_2, ..., v_j, ...v_p$. Le tableau des données brutes 'a partir duquel on va faire l'analyse est noté X et a la forme suivante :

$$\text{X}=\begin{pmatrix} x_{11} & x_{12} & .. & x_{1j} & .. & x_{1p} \\ x_{21} & x_{22} & .. & x_{2j} & .. & x_{2p} \\ . & . & .. & . & .. & . \\ x_{i1} & xi2 & .. & x_{ij} & .. & x_{ip} \\ . & . & .. & . & .. & . \\ x_{n1} & xn2 & .. & x_{nj} & .. & x_{np} \end{pmatrix}$$

On peut représenter chaque unité par le vecteur de ses mesures sur les p variables :

${}^tU_i = (x_{i1}, x_{i2}, .., x_{ip})$ ce qui donne $U_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ .. \\ x_{ip} \end{pmatrix}$

Alors U_i est un vecteur de $\mathbb{R}^p$ Pour avoir une image de l'ensemble des unités, on se place dans un espace affine en choisissant comme origine un vecteur particulier de $\mathbb{R}^p$, par exemple le vecteur dont toutes les coordonnées sont nulles. Alors, chaque unité sera représentée par un point dans cet espace. L'ensemble des points qui représentent les unités est appelé traditionnellement "nuage des individus".
En faisant de même dans $\mathbb{R}^n$, chaque variable pourra être représenter par un point de l'espace affine correspondant. L'ensemble des points qui représentent les variables est appelé "nuage des variables".
On constate, que ces espaces étant de dimension supérieure en général à 2 et même 3, on ne peut visualiser ces représentations. L'idée générale des méthodes factorielles est de trouver un système d'axes et de plans tels que les projections de ces nuages de points sur ces axes et ces plans permettent de reconstituer les positions des points les uns par rapport aux autres, c'est-'a-dire avoir des images les moins déformées possible.

Présentation d'un point

Supposons qu'on ait des vecteurs $x_1, ..., x_n$de dimension D. Le but est de les représenter par un point x_0.On veut, comme pour les moindres carrés, minimiser une erreur, qui est ici l'erreur entre $x_k$etx_0Comme pour la méthode des moindres carrés, on veut alors minimiser la quantité suivante :

$$\frac{1}{n}\sum_{i=1}^{n} \parallel x_k - x_0 \parallel^2$$

Pour faire une représentation géométrique, il faut choisir une distance entre deux points de l'espace. La distance utilisée par l'ACP dans l'espace ou sont représentés les unités, est la distance euclidienne classique. La distance entre deux unités u_i et $u_{i'}$ est égale à :

$$d^2(u_i, u_{i'}) = \sum_{j=1}^{p}(x_{ij}, x_{i'j})^2$$

Avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale.à cette distance on associe un produit scalaire entre deux vecteurs :

$$\langle \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} \rangle = \sum_{j=1}^{p} x_{ij}x_{ij'} = {}^tU_iU_{i'}$$

ainsi que la norme d'un vecteur :

$$\| \overrightarrow{ou_i} \|^2 = \sum_{j=1}^{p} x_{ij}^2 = {}^tU_iU_i$$

Le choix de l'origine pour l'ACP est le centre de gravité du nuage de points,Pour définir ce centre de gravité, il faut choisir un système de pondération des unités.
Le centre de gravité G du nuage des individus est alors le point dont les coordonnées sont les valeurs moyennes des variables

$$\text{G}= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix}$$

Inertie totatle

On note Inertie totale du nuage des individus : $I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, U_i)$
Ce moment d'inertie totale est intéressant car c'est une mesure de la dispersion du nuage des individus par rapport à son centre de gravité. Si ce moment d'inertie est grand,cela signifie que le nuage est très dispersé, tandis que s'il est petit, alors le nuage est très concentré sur son centre de gravité.

Espace des individus

On recherche des sous-espaces représentant au mieux ce nuage de point en respectant 2 critères : le critère de proximité et la fidélité des distances.C'est le sous-espace passant par G qui optimise ces deux critères.
Soit H le sous-espace passant par G, on distingue deux types d'inertie :

- Inertie expliquée par H :

$I_{exp}(H) = \frac{1}{n}\sum_{i=1}^{n} d^2(G, \hat{U}_i)$

- Inertie résiduelle autour de H

$I_{rs}(H) = \frac{1}{n}\sum_{i=1}^{n} d^2(\hat{U}_i, U_i)$

Et pour choisir H, il faut minimiser I_{rs} et maximiser I_{exp}

L'espace des variables

On fait un changement d'origine : $G = 0$ (centrage des variables), La recherche des sous-espaces Hk se fait de proche en proche pour $k = 1p$:Une fois U_1 déterminé, on peut démontrer que le sous-espace $H2$ s'ajustant au mieux au nuage de points contient nécessairement U_1.
Pour déterminer le sous-espace H2, on recherche U_2 tel que U_2 perpendiculaire à U_1 et tel que

la droite portée par U_2, passant par 0, ait une inertie maximale.
On peu démontrer que les vecteurs $U_1, U_2, \ldots, U_p$ peuvent s'obtenir à partir de la matrice d'inertie C (covariance ou corrélation) des valeurs de X, cette matrice est tel qu'il existe V vecteurs propres et λ valeurs propres, Ces vecteurs sont orthogonaux deux à deux et unitaires (de longueur égale à 1). Ils peuvent être rangées par ordre décroissant des valeurs propres associées : le premier vecteur propre V_1 est associé à la valeur propre la plus élevé λ_1 . Ces vecteur sont les vecteurs U_1à U_p recherchés, d'où les droites engendrées par ces vecteurs propres sont appelées respectivement le 1er, 2ème, et pième axe principal d'inertie du nuage.

7.1.6 ACP robuste[7]

Ainsi que nous l'avons indiqué en introduction, une difficulté majeure de l'ACP provient de sa formalisation au moyen d'un critère des moindres carrées qui reste très sensible à la présence de valeurs aberrantes. Une solution consiste à utiliser une technique itérative. Dans une première étape, la présence éventuelle des valeurs aberrantes n'est pas prise en compte et l'ACP est effectuée sur l'ensemble des données disponibles. Puis, grâce au modèle ACP obtenu, les données sont analysées (analyse de leurs projections dans l'espace résiduel afin de déceler la présence de valeurs aberrantes ; ces dernières sont alors retirées et l'ACP est reconduite sur les données restantes. Le processus peut être itérée plusieurs fois ; il est assez efficace lorsque peu de données sont contaminées mais échoue dans la plupart des cas. Pour tolérer la présence de valeurs aberrantes, une autre méthode couramment utilisée dans le domaine de la statistique consiste à remplacer l'estimation standard de la moyenne et de la variance par une estimation robuste [2], [3]. Une autre solution consiste à utiliser une approche par projection dans laquelle les vecteurs propres de la matrice de covariance sont progressivement estimée de façon à réduire l'influence des valeurs aberrantes.
Au lieu de minimiser le carré de la norme des erreurs, nous allons les minimiser par rapport à une fonction ρ

7.2 Annexe B

7.2.1 Jeux de Données

Les données utilisées durant mon stage proviennent de :
- Rotheau2009(Données VL et PL)
- Haguenau2009(Données VL et PL)
- Hagueneau2011(Données VL et PL)
- Erstein2011(Données VL et PL)
- Matzenheim2011(Données VL et PL)
- RN59 2012(Données VL et PL)

Les jeux de donnés de Hagueneau2009 et de Rotheau2009 ont été utilisés pour la première partie de mon stage dans la suite du travail commencé par Marie-Paule Ehrhart[6], pour départager les méthodes d'estimations.
Pour les données de Rn59, j'ai procédé à une modification de 4 passages en changeant leurs étiquettes, dans le but de tester l'algorithmes EM.
Dans les récents jeux de données (Erstein2011, Matzenhein2011, RN59, Haguenau2011)on avait pas seulement les données sur la vitesse et le L_{Amax}, il y avait également des données par bandes de tiers d'octaves L3Fmax, ces données qui ont été principalement utilisé pour la partie de l'Isomap, du K-means, ainsi que pour la dernière partie du remplacement de l'indice L_{Amax} par le SEL ou L_{AE}

7.2.2 Classification de l'opérateur et de K-means

Après avoir effectué un K-means sur les données issus d'un Isomap de Hagueneau2011, on voulait regarder la différence entre les résultats obtenus par l'opérateur lors du dépouillement et celle du K-means.
C-dissous, les données de l'opérateur classifié sous R :

```
which((classe[1:194])==1)
  [1] 1 3 6 7 8 9 10 11 12 13 14 15 16 17 19 20 21 23
 [19] 24 25 26 27 30 31 34 35 37 38 39 40 41 42 43 44 45 46
 [37] 47 48 49 50 52 53 54 55 56 57 59 61 64 65 66 67 68 69
 [55] 70 72 74 75 77 78 80 81 82 83 84 85 86 87 88 89 90 91
 [73] 92 93 94 95 96 97 100 102 103 104 106 107 108 109 110 112 113 114
 [91] 115 116 117 120 122 124 125 126 127 128 129 130 131 132 134 135 136 137
[109] 138 142 143 146 148 149 150 154 155 159 160 161 162 163 164 166 167 168
[127] 169 172 173 174 177 178 179 180 181 182 183 184 185 186 187 189 190 192
[145] 193 194
>which((classe[1:194])==2)
 [1] 2 4 5 18 22 28 29 32 33 36 51 58 60 62 63 71 73 76 79
[20] 98 99 101 105 111 118 119 121 123 133 139 140 141 144 145 147 151 152 153
[39] 156 157 158 165 170 171 175 176 188 19
```

Les figures 7.1 et 7.2 donnent la sortie des données classifiés par le K-means :

```
  2   4   5  18  22  28  29  32  33  36  51  58  60  62  63  71  73  76  79
 98  99 101 105 111 118 119 121 123 133 139 140 141 144 145 147 151 152 153
156 157 158 170 171 175 176 188 191
```

FIGURE 7.2 – Classe PL issus du K-means Haguenau2011

```
  1   3   6   7   8   9  10  11  12  13  14  15  16  17  19  20  21  23
 24  25  26  27  30  31  34  35  37  38  39  40  41  42  43  44  45  46
 47  48  49  50  52  53  54  55  56  57  59  61  64  65  66  67  68  69
 70  72  74  75  77  78  80  81  82  83  84  85  86  87  88  89  90  91
 92  93  94  95  96  97 100 102 103 104 106 107 108 109 110 112 113 114
115 116 117 120 122 124 125 126 127 128 129 130 131 132 134 135 136 137
138 142 143 146 148 149 150 154 155 159 160 161 162 163 164 165 166 167
168 169 172 173 174 177 178 179 180 181 182 183 184 185 186 187 189 190
192 193 194
```

FIGURE 7.3 – Classe VL issus du K-means Haguenau2011

7.2.3 Comparaisons des sorties K-means et Isomap sur les données avec et sans vecteur vitesse

```
> ind.cl1<-which(km.data.ss_vit1$cluster==2)
  [1]   1   3   6   7   8   9  10  11  12  13  14  15  16  17  19  20  21  23
 [19]  24  25  26  27  30  31  34  35  37  38  39  40  41  42  43  44  45  46
 [37]  47  48  49  50  52  53  54  55  56  57  59  61  64  65  66  67  68  69
 [55]  70  72  74  75  77  78  80  81  82  83  84  85  86  87  88  89  90  91
 [73]  92  93  94  95  96  97 100 102 103 104 106 107 108 109 110 112 113 114
 [91] 115 116 117 120 122 124 125 126 127 128 129 130 131 132 134 135 136 137
[109] 138 142 143 146 148 149 150 154 155 159 160 161 162 163 164 165 166 167
[127] 168 169 172 173 174 177 178 179 180 181 182 183 184 185 186 187 189 190
[145] 192 193 194
> ind.cl2<-which(km.data.ss_vit1$cluster==1)
 [1]   2   4   5  18  22  28  29  32  33  36  51  58  60  62  63  71  73  76  79
[20]  98  99 101 105 111 118 119 121 123 133 139 140 141 144 145 147 151 152 153
[39] 156 157 158 170 171 175 176 188 191
```

Remarques On remarque une différence d'une observation "36", entre la classification avec et sans vitesse.

7.3 dBEuler

dBEuler est une production depuis 2006.La dernière version est la 1.6 portant sur la corrections de bugs et évolutions mineurs.C'est sous l'environnement Scilab qu'a été développée l'application dBEuler à partir de 2003 par M.Dutilleux et par de nombreux stagiaires et vacataires.

7.3.1 Fonctionnement de dBEuler

dBEuler fonctionne de maniére assez linéaire.étape par étape. La première action à son ouverture consiste à créer ou ouvrir une campagne. Dans un second temps, l'opérateur importe les enregistrements audio ou alors il peut les enregistrer grâce aux outils d'enregistrements directement disponibles dans le logiciel. L'étape suivante consiste à traiter les calibrages et pré-traiter les enregistrements, soit en pondération A, soit en pondération A avec tiers 2, dans tous les cas dBEuler analyse de toute façon par la suite en tiers d'octave. Dés lors, on passe à l'étape de dépouillement à proprement parler. C'est-à-dire que l'on va traiter chaque enregistrement, découper les passages et stocker les données pertinentes. Une fois ce travail achevé, l'opérateur peut demander au logiciel l'analyse statistique des données mesurées, voir la droite de régression ainsi calculée par catégorie de véhicule sur les couples vitesse, L_{Amax} et supprimer les points aberrants 3 si besoin. Finalement, le logiciel est capable de créer un rapport de la campagne dans un format bien défini.

Table des figures

Liste des tableaux

Bibliographie

[1] S 31-119. *Acoustique - Caractérisation in situ des qualités acoustiques des revêtements de chaussées - Mesurages acoustiques au passage (norme annulée).* AFNOR, Octobre 1993.

[2] Frédéric Bertrand. *Tests de normalité - Distributions univariées et multivariées.* IRMA, Université de Strasbourg, 2011.

[3] C. Bishop. *Neural Networks for Pattern Recognition.* Information science and Statistics. Oxford University Press, Oxford, Angleterre, 1995.

[4] Pierre Charbonnier. *Classification et reconnaissance des formes.* Master IRIV (2ème année) - ENSPS - Université de Strasbourg, 2011.

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

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

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

[8] Meunier. *Reconnaissance de formes.* IAR-6002, 2002.

[9] Gilbert Saporta. *Probabilités, Analyse des données et Statistique.* Editions Technip, 2006.

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

[11] Renaud Wintzer. Rapport de stage : Evolution de dBEuler, 2010.

[12] Ruben H. Zamar and Matias Salibian-Barrera. Bootstrapping robust estimates of regression. *The Annals of Statistics*, 30(2) :556–582, 2002.