Université de Montréal

Modélisation statistique de l'érosion de cavitation d'une turbine hydraulique selon les paramètres d'opération

par

Paule-Marjolaine Bodson-Clermont

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

22 mars 2015

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Modélisation statistique de l'érosion de cavitation d'une turbine hydraulique selon les paramètres d'opération

présenté par

Paule-Marjolaine Bodson-Clermont

a été évalué par un jury composé des personnes suivantes :

M. Alejandro Murua

(président-rapporteur)

M. Jean-François Angers

(directeur de recherche)

M. Pierre Lafaye de Micheaux

(membre du jury)

Mémoire accepté le

14 septembre 2015

SOMMAIRE

Dans une turbine hydraulique, la rotation des aubes dans l'eau crée une zone de basse pression, amenant l'eau à passer de l'état liquide à l'état gazeux. Ce phénomène de changement de phase est appelé cavitation et est similaire à l'ébullition. Lorsque les cavités de vapeur formées implosent près des parois, il en résulte une érosion sévère des matériaux, accélérant de façon importante la dégradation de la turbine. Un système de détection de l'érosion de cavitation à l'aide de mesures vibratoires, employable sur les turbines en opération, a donc été installé sur quatre groupes turbine-alternateur d'une centrale et permet d'estimer précisément le taux d'érosion en kg/10 000 h.

Le présent projet vise à répondre à deux objectifs principaux. Premièrement, étudier le comportement de la cavitation sur un groupe turbine-alternateur cible et construire un modèle statistique, dans le but de prédire la variable cavitation en fonction des variables opératoires (telles l'ouverture de vannage, le débit, les niveaux amont et aval, etc.). Deuxièmement, élaborer une méthodologie permettant la reproductibilité de l'étude à d'autres sites. Une étude rétrospective sera effectuée et on se concentrera sur les données disponibles depuis la mise à jour du système en 2010.

Des résultats préliminaires ont mis en évidence l'hétérogénéité du comportement de cavitation ainsi que des changements entre la relation entre la cavitation et diverses variables opératoires. Nous nous proposons de développer un modèle probabiliste adapté, en utilisant notamment le regroupement hiérarchique et des modèles de régression linéaire multiple.

Mots-clé : cavitation, turbine Francis, mélanges de lois, statistique, opération, regroupement hiérarchique, régression linéaire multiple

SUMMARY

Cavitation erosion which results from repeated collapse of transient vapor cavities on solid surfaces is a constant problematic in hydraulic turbine runners and continues to enforce costly repair and loss of revenues. A vibratory detection system of cavitation erosion was installed 10 years ago for continuous monitoring of 4 hydropower units. A new hardware version of the system was developed and installed in 2010. This new system configuration is more reliable and allows more accurate evaluation of the cavitation erosion of the runners in kg/10 000 h.

The first objective of this study is to investigate cavitation behavior upon one generating unit and to build a statistical model which will allow prediction of instant cavitation related to operating variables, such as gate opening, water flow, headwater level, tailwater levels, etc. The second objective is to develop a methodology for the reproducibility of the studies to other sites. A retrospective study will be conducted and we will mainly focus on data available since the system update in 2010.

The preliminary analysis enhanced the complexity of the phenomenon. Indeed, changes in the relationship between cavitation and various operating variables were observed and could be due to a seasonal behavior or different operating conditions. Using hierarchical clustering and regression models, we formalize this heterogeneity by developing a model which includes operating variables such as active power, tailwater level and gate opening.

Keywords : Cavitation, Francis turbine, mixture model, operation, hierarchical clustering, multiple linear regression

TABLE DES MATIÈRES

REMERCIEMENTS

Je souhaite d'abord remercier mon directeur de recherche, M. Jean-François Angers pour ses conseils avisés et pour sa disponibilité. Je le remercie particulièrement pour le temps qu'il m'a accordé, tant pour la préparation de mes présentations scientifiques que pour les rencontres de suivi pour les travaux de ce mémoire. Son expérience et son intuition m'ont été d'un grand secours tout au long de ces travaux, spécialement sur le plan de l'organisation des idées et pour ma compréhension de la démarche scientifique.

D'autre part, j'aimerais souligner l'aide inestimable de M. François Lafleur, chercheur à l'IREQ, qui m'a donné l'opportunité de vivre une expérience de recherche particulièrement stimulante. De l'élaboration de la problématique à l'implantation des résultats de ce mémoire à Hydro-Québec, j'ai toujours reçu un soutien sans faille, et j'ai bénéficié de sa compétence et de son expertise autant d'un point de vue technique que dans l'élaboration de la méthodologie de recherche. Je remercie aussi M. Denis Thibault, chercheur à l'IREQ et responsable du projet PréDDIT, ainsi que M. Luc Perreault, pour son aide sur le plan statistique et pour la rédaction.

Je remercie chaleureusement Audrey-Anne, Alexandre et Janie ainsi que tous mes collègues qui, par leur bonne humeur, leur motivation et leur sens de l'humour, ont contribué à rendre mes journées plus agréables, faisant du département de mathématiques et statistique un des meilleurs environnements de travail que j'ai connu.

Je souhaite aussi exprimer ma gratitude à ma famille et à tous mes amis qui m'ont de près ou de loin encouragé, prêté une oreille attentive, fait rire et qui ont su me changer les idées, me permettant d'aborder la vie, et mon mémoire, avec un œil différent. Je remercie particulièrement mes parents pour leurs encouragements et leur soutien sans faille ainsi que Gabriel, pour sa patience à toute épreuve et son talent inné pour me faire rire. Merci aussi à M. Jacques Zoon, musicien extraordinaire et professeur de flûte de talent, qui a su trouver les mots qu'il faut en m'encourageant à me réinventer.

Finalement, je remercie Mitacs, la Faculté des études supérieures et post-doctorales de l'Université de Montréal et la Société belge de bienfaisance pour leur appui financier tout au long de mes études.

INTRODUCTION

Avec la demande croissante d'énergie renouvelable et l'ouverture des marchés de l'énergie, nous observons une augmentation de la sollicitation des équipements et un changement dans le mode opératoire des centrales hydroélectriques. Il devient ainsi primordial pour les producteurs de caractériser la durée de vie des turbines et de cibler précisément les délais pour l'arrêt, la maintenance et la réparation de celles-ci. L'érosion de cavitation, phénomène dû à l'implosion de bulles d'air près des parois, est un des modes principaux de dégradation des turbines et un facteur capital de l'efficacité et de la durée de vie des turbines. Toutefois, sa complexité pose de multiples défis pour l'étude du phénomène sur le terrain.

Plusieurs chercheurs ont étudié la cavitation à l'aide de simulations numériques ou avec des simulations en laboratoire (voir Giroux *et al.*, 2011; Quian *et al.*, 2007). En revanche, peu d'études ont utilisé l'approche statistique afin de modéliser le comportement de la cavitation, puisque classiquement, la collecte de données d'érosion se fait en situation d'arrêt complet des turbines, ce qui se produit rarement. Le système de mesures vibratoires mis au point à l'Institut de recherche d'Hydro-Québec (IREQ) a l'avantage de surveiller la cavitation sur une base régulière (toutes les 12 minutes environ) et une base de données plus complète de l'érosion de cavitation et de son évolution est donc disponible.

Ce projet de recherche vise donc à aborder la caractérisation du phénomène de l'érosion de cavitation par le biais d'un modèle probabiliste adapté, permettant d'adopter un point de vue différent pour enrichir la compréhension du phénomène de cavitation. Nous cherchons à comprendre de quelle manière modéliser l'érosion de cavitation en nous appuyant sur les conditions opératoires de quatre groupes turbine-alternateur comme l'ouverture de vannage, le débit et la puissance, et à en quantifier l'incertitude. La mise au point d'un tel outil servira à prédire la progression de la dégradation due à la cavitation et pourrait éventuellement permettre de repousser les délais de réparation et de maintenance, se traduisant par une diminution directe de perte de revenus causée par l'arrêt prolongé des turbines.

Une des difficultés majeures de l'étude du phénomène de cavitation est la nécessité d'attendre l'arrêt complet des turbines pendant plusieurs semaines pour obtenir des données. L'idée de Paul Bourdon, chercheur à l'IREQ, a été d'établir la relation entre la propagation d'ondes vibratoires générées par la cavitation d'entrée et le taux d'érosion de cavitation de la turbine (voir Bourdon, 2000). Suite à ses recherches, un système de mesures vibratoires employable sur les machines en opération a donc été mis au point et validé. Celui-ci permet d'estimer précisément l'érosion de cavitation instantanée en kg/10 000 h en utilisant, entre autres, la fonction vibratoire de transmissibilité et les valeurs quadratiques moyennes (VQM)(voir Bourdon, 2000; Lafleur, 2011).

Notons que la valeur quadratique moyenne est une mesure des amplitudes de vibration qui tient compte de l'évolution du signal vibratoire dans le temps, ainsi que de l'énergie vibratoire et du potentiel de détérioration de la vibration. La fonction vibratoire de transmissibilité, quant à elle, permet de mesurer la modification du signal de vibration dans l'eau ou l'air. Elle ajuste ainsi le signal de vibration perçu à l'accéléromètre pour l'exprimer avec sa « vraie » valeur à la source de cavitation.

Mis en place en 1999 sur quatre groupes turbine-alternateur d'une certaine centrale (ci-après nommée centrale W), ce système a été mis à jour en 2010 pour pallier à certaines problématiques liées à l'acquisition des données. Les évaluations de l'érosion de cavitation sont depuis plus adéquates, bien que des travaux se poursuivent pour améliorer le modèle physique (voir Lafleur, 2012).

Le projet de recherche développé dans ce mémoire souhaite répondre à deux objectifs principaux :

(1) étudier le comportement de la cavitation sur un groupe turbine-alternateur cible et construire un modèle statistique dans le but de prévoir la variable cavitation instantanée en fonction des variables opératoires (par exemple l'ouverture de vannage, le débit, les niveaux amont et aval, etc.) ;

(2) élaborer une méthodologie permettant la reproductibilité de l'étude à d'autres sites.

La réalisation de ces objectifs permettra dans le meilleur des cas de nous doter d'outils pour prédire le taux d'érosion sur le groupe turbine-alternateur cible sans utiliser le système vibratoire. D'autre part, la méthodologie pourra être reproduite pour les trois autres groupes turbine-alternateur munis du système vibratoire. Ce dernier pourrait ainsi éventuellement être installé sur d'autres sites du parc hydro-électrique pour étudier l'érosion de cavitation dans des conditions opératoires différentes.

Dans le premier chapitre, nous présentons le phénomène de cavitation, le cadre de l'étude et les résultats de l'analyse exploratoire. Cette analyse sert en particulier à valider les bases de données et à explorer la relation entre la cavitation et les variables opératoires. Le comportement hétérogène de la cavitation est par ailleurs mis en évidence et son lien avec une certaine saisonnalité est explicité.

Le deuxième chapitre est consacré à l'analyse du comportement de la cavitation basé sur les modes opératoires. Les spécialistes étant réticents à s'appuyer sur la composante temporelle pour l'analyse, la piste des modes opératoires est plutôt retenue pour la suite de l'étude. En effet, l'analyse exploratoire souligne l'impact du fonctionnement des autres groupes turbine-alternateur partageant la même chambre d'équilibre sur la cavitation du groupe à l'étude (groupe 7). L'objectif étant ainsi de grouper les modes opératoires liés à des conditions opératoires similaires pour le groupe turbine-alternateur 7, nous utilisons le regroupement hiérarchique pour classifier les données de cavitation en grappes. Le comportement de la cavitation par grappe sera par la suite étudié.

L'hétérogénéité de la cavitation est traitée dans le chapitre 3. En effet, la distribution de la cavitation par grappe semble évoluer selon une distribution plurimodale. Les modèles de type mélange de lois de probabilités, bien adaptés pour représenter ce type de phénomène hétérogène, y sont présentés. Ce type de méthode a déjà été utilisé à l'IREQ dans le contexte de modélisation des séries de pointes et volumes de crues printanières, en adoptant la perspective bayésienne pour l'estimation des paramètres et en modélisant la dépendance au temps en utilisant les chaînes de Markov cachées (HMM)(voir Evin *et al.*, 2011). Ici, nous développons plutôt le modèle de mélanges de lois pour la loi de Burr à trois paramètres, notamment pour tenir compte de l'asymétrie de certaines distributions. La mise en oeuvre du modèle est réalisée à l'aide de l'algorithme Espérance-Maximisation. Par la suite, les résultats de l'ajustement du modèle de mélanges avec une, deux ou trois composantes est présenté, et nous estimons la cavitation par l'espérance du modèle le plus adéquat.

Dans le chapitre 4, nous présentons les résultats de quatre modèles de régression linéaire par grappe. Afin d'affiner l'estimation de la cavitation pour chaque grappe, nous intégrons les variables opératoires dans un modèle de régression avec et sans transformation logarithmique de la cavitation et avec et sans interactions. La méthode de sélection descendante est utilisée pour la sélection des variables des quatre modèles dans chacune des grappes et l'analyse des résidus est présentée.

Le chapitre 5 est consacré à la validation des modèles. Nous y résumons les modèles considérés et exposons les méthodes de validation croisée pour tester la

robustesse du choix des prédicteurs et la robustesse des paramètres. Les meilleurs modèles de prédiction et les critères de sélection sont aussi présentés.

Finalement, dans la conclusion, nous apportons certaines recommandations quant au modèle final sélectionné. Les limites de l'étude sont soulignées et nous donnons quelques pistes d'intérêt pour les travaux futurs.

Chapitre 1

CADRE DE L'ÉTUDE ET ANALYSE EXPLORATOIRE

Dans ce chapitre, nous présentons d'abord le phénomène de cavitation et son impact sur l'érosion des turbines hydrauliques. La section suivante décrit le cadre de l'étude, soient les variables à l'étude, les différentes bases de données et leurs caractéristiques, ainsi que la mise en oeuvre de la base de données de travail. La dernière section est consacrée aux résultats de l'analyse exploratoire de la cavitation du groupe à l'étude (le groupe 7 de la centrale W) pendant l'année 2011 : le comportement de la cavitation y est étudié et son lien avec les variables opératoires et le fonctionnement des autres groupes turbine-alternateur de la centrale est exploré. Ces résultats serviront de base pour la construction du modèle statistique développé dans les chapitres suivants.

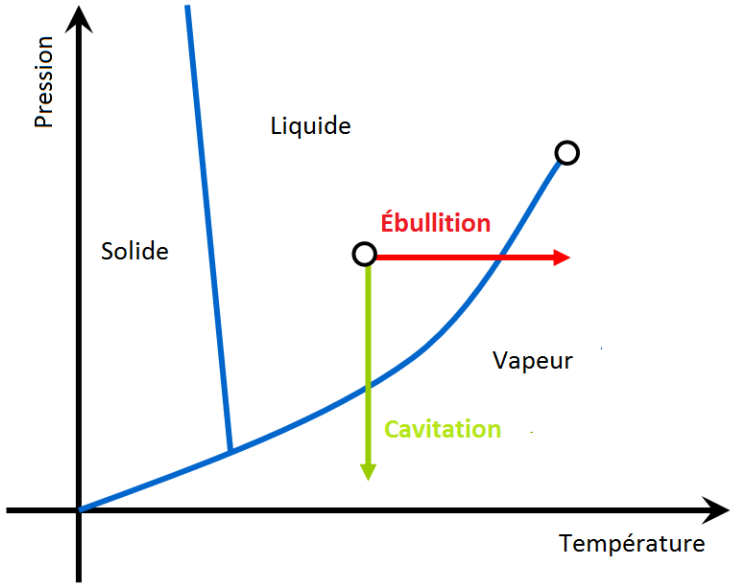
1.1. Phénomène de cavitation

1.1.1. Définition

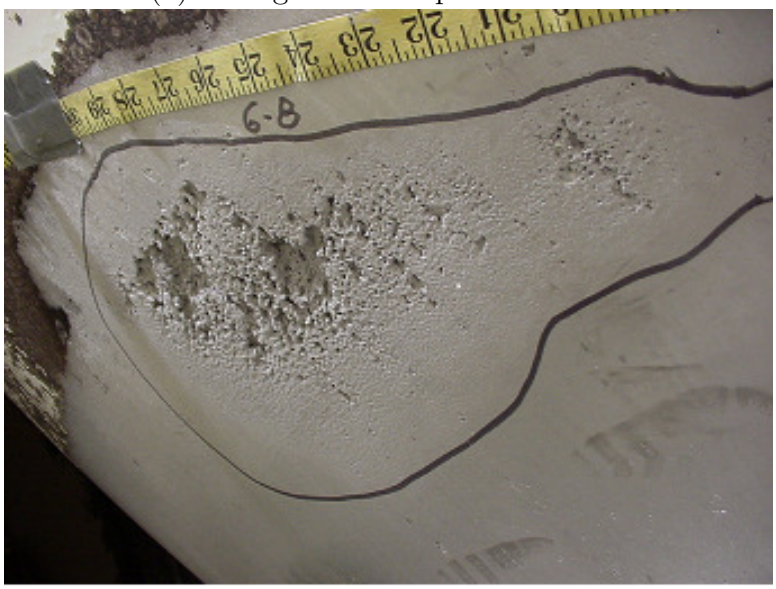
Dans une turbine hydraulique, la rotation des aubes dans l'eau crée une zone de basse pression, amenant l'eau à passer de l'état liquide à l'état gazeux sans apport extérieur de chaleur (voir figure 1.1(a)). Ce phénomène de changement de phase est appelé cavitation et est similaire à l'ébullition. Lorsque les bulles de vapeur formées implosent près des parois de la turbine, il en résulte une érosion sévère des matériaux, accélérant de façon importante la dégradation de la turbine (voir figure 1.1(b)).

1.1.2. Types de cavitation

Les turbines à l'étude à la centrale W sont de type Francis. Dans ces turbines, différentes formes de cavitation se manifestent avec différents niveaux d'agressivité. Les principales formes sont la torche de faible charge, la torche de forte



(a) Changements de phase de l'eau



(b) Exemple d'érosion de cavitation sur une aube

FIGURE 1.1. Phénomène de cavitation et impact sur l'érosion des turbines

charge et la cavitation d'entrée. Les problèmes d'érosion les plus sévères dans les turbines Francis sont généralement liés à la cavitation d'entrée (voir Bourdon, 2000). Cette dernière se caractérise par une poche de vapeur, visible à l'extrados des aubes (voir figure 1.2). Cette poche est généralement instable et lâche des cavités transitoires qui, lorsqu'elles implosent, font apparaître une onde de choc de forte intensité qui excite mécaniquement la structure.

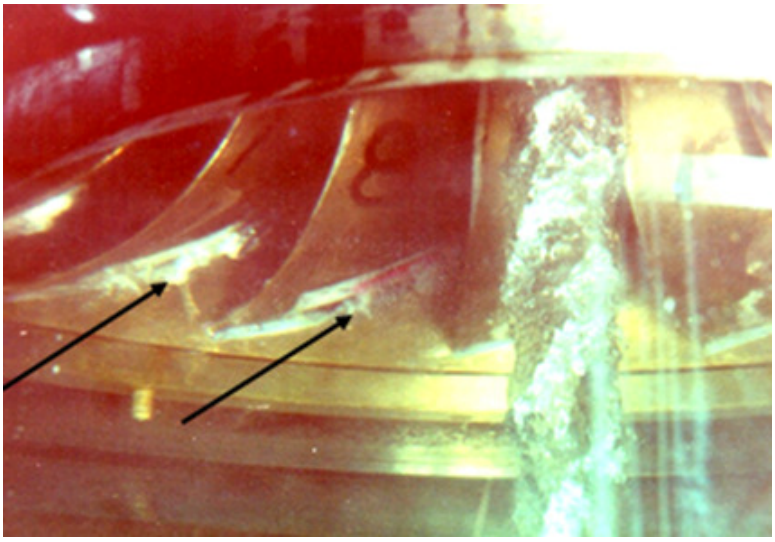


FIGURE 1.2. Cavitation d'entrée très agressive

1.1.3. Complexité du phénomène

Différentes difficultés se présentent lorsqu'on tente de mesurer ou prédire l'érosion de cavitation, principalement liées à la complexité du phénomène. L'agressivité de cavitation mesurée par le fabricant n'est pas connue aux points d'opération extérieurs à la zone d'opération optimale. Quoiqu'il serait en théorie possible d'exploiter la turbine dans cette zone pour minimiser la cavitation, les besoins des exploitants et les conditions hydriques imposent souvent d'opérer à l'extérieur de ces conditions optimales, d'où la nécessité de mieux caractériser le phénomène dans ces zones d'opération.

La complexité du phénomène d'érosion par cavitation se traduit aussi par une distribution aléatoire de volumes de cratères sur la surface (voir figure 1.3). Notons en outre que le taux d'évolution de perte de métal varie dans le temps et selon la résistance des matériaux. Ainsi, la détection de l'érosion de cavitation pendant l'opération des turbines reste un phénomène complexe, mais elle est toutefois primordiale pour la maintenance du parc hydro-électrique.

1.1.4. Impact sur les coûts

L'inspection et la réparation suite à l'érosion par cavitation est un processus long et coûteux, qui engendre un arrêt complet des turbines de l'ordre de quatre à six semaines et donc une perte de revenus élevée. La récolte d'informations à ce sujet sur le terrain est de ce fait difficile et nous disposons de peu de données mesurées directement.

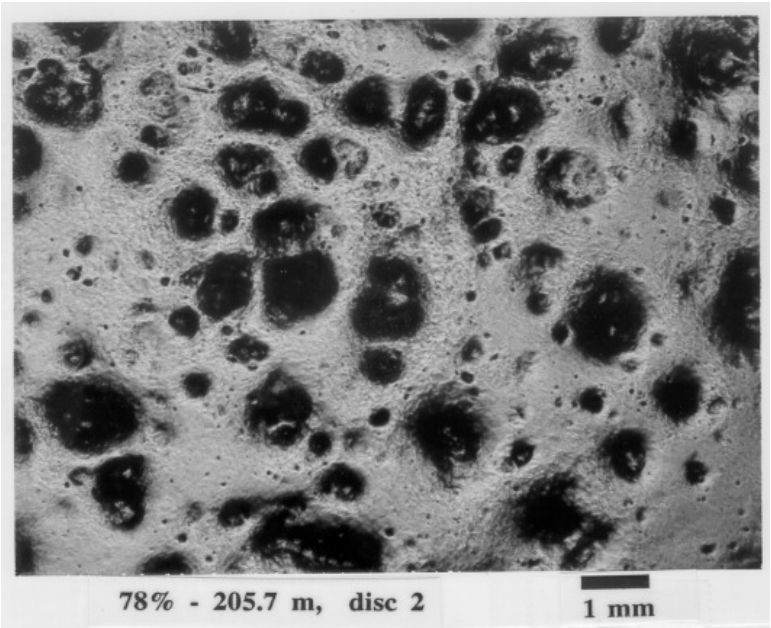


FIGURE 1.3. Exemple d'érosion de cavitation sur une aube lors d'essais de cavitation (Inox 316L soumis à 40 minutes de cavitation à 266 MW)

1.2. CADRE DE L'ÉTUDE

Afin de faciliter la mise en place d'une méthodologie précise, il est établi que nous ciblerons les données d'un groupe turbine-alternateur spécifique (groupe 7) et d'une période spécifique (année 2011). Mentionnons que la centrale W dispose de directives d'exploitation quant à l'ordre de démarrage des groupes, qui priorisent le démarrage du groupe 7 en premier et son arrêt en dernier. Nous disposons donc d'un nombre de données plus élevé pour ce groupe et de configurations de fonctionnement plus diverses, ce qui explique notre choix. Par ailleurs, le système vibratoire ayant été mis à jour en février 2010, l'année 2011 est la première à contenir des données adéquates complètes. L'année 2012 servira à la validation.

1.2.1. Variables

Les variables explicatives retenues sont :

courant (A): : courant électrique produit par la rotation de la turbine ;

débit (m^3/s): : mesure du volume de l'eau par unité de temps entrant dans la turbine ;

amont (m): : niveau de l'eau au réservoir ;

aval (m): : niveau de l'eau à la sortie de la turbine, dans la chambre d'équilibre ;

ouverture de vannage (%): : pourcentage d'ouverture des aubes de la turbine, contrôlant le débit d'eau à l'entrée ;

puissance active (MW): : composante de la puissance faisant appel au courant en phase avec la tension (1.2.1). La puissance active se calcule selon l'équation suivante :

$$\text{Puissance active} = \text{Tension} * \text{Courant actif} = U * I * \cos(\phi), \tag{1.2.1}$$

où ϕ est l'angle entre la tension et le courant ;

puissance réactive (Mvar): : composante de la puissance qui s'exprime en VAR, ou volt-ampère-réactif (1.2.2). La puissance réactive se calcule selon l'équation suivante :

$$\text{Puissance réactive} = \text{Tension} * \text{Courant réactif} = U * I * \sin(\phi), \tag{1.2.2}$$

où ϕ est l'angle entre la tension et le courant ;

puissance effective de stabilité (MW): : puissance maximale que la turbine peut atteindre transitoirement, c'est-à-dire sans tenir compte des limites de l'alternateur ;

tension (kV): : exprime la différence de potentiel électrique entre deux points d'un circuit.

La variable réponse (ou variable d'intérêt) est la variable *taux d'érosion de cavitation* (kg/10 000 h) qui, par abus de langage, est nommée cavitation. Deux variables ont été ajoutées : la variable *hauteur de chute* (m), qui correspond au niveau amont moins le niveau aval, ainsi que le fonctionnement des sept autres groupes partageant la chambre d'équilibre (Marche-Arrêt). Notons que la variable *température de l'eau* n'a pas été utilisée ici puisque c'est une variable lente qui ne varie pas dans un court intervalle de temps. L'objectif étant de modéliser les changements d'érosion de cavitation de façon ponctuelle (qui varient en fonction des conditions opératoires toutes les 12 minutes environ), cette variable n'a pas été intégrée.

L'unité expérimentale est le groupe turbine-alternateur 7. Toutes les mesures ont été prises sur la même unité expérimentale à intervalles de temps non réguliers, qui dépendent du fonctionnement des autres groupes munis du système de mesures vibratoires (5-6-8).

1.2.2. Chambre d'équilibre

La centrale W présente une configuration spécifique : huit groupes turbine-alternateur partagent en effet la même chambre d'équilibre, bassin où se déverse

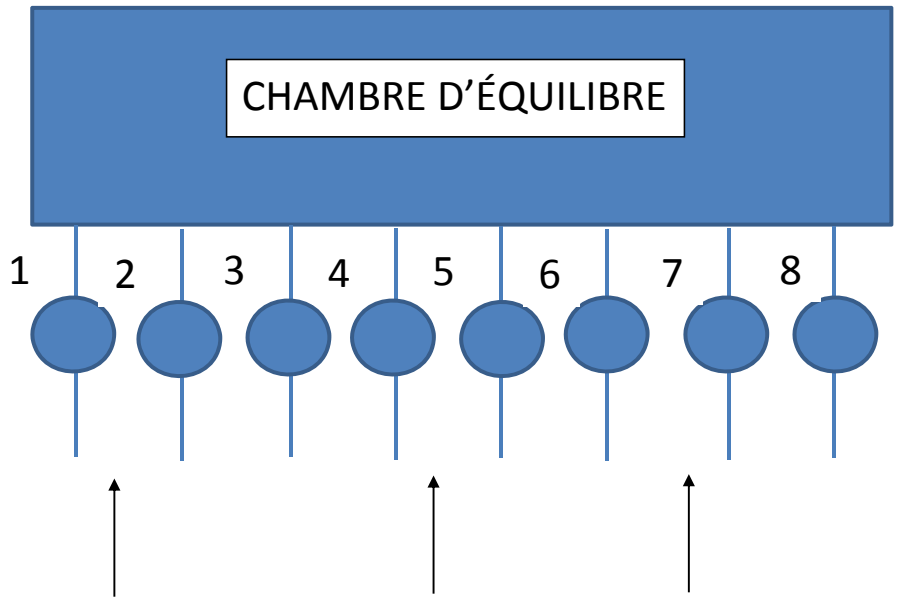


FIGURE 1.4. Schématisation de la chambre d'équilibre à la centrale W

l'eau à la sortie des turbines. Le niveau aval de cette chambre d'équilibre est donc influencé par le fonctionnement ou l'arrêt des autres groupes. De ces huit groupes, quatre sont munis du système de mesures vibratoires qui permet de collecter des données sur l'érosion (groupes 5 à 8). La configuration précise est telle qu'on peut la voir à la figure 1.4.

1.2.3. Construction de la base de données

Trois bases de données principales ont été utilisées pour construire la base de travail : la base de données BD1 pour les données d'érosion instantanée de cavitation, la base de données BD2 pour les variables opératoires et la base de donnée BD3 pour les valeurs nominales des conditions d'exploitation de chaque turbine.

1.2.3.1. *BD1 : données de cavitation*

La première base de données (BD1) fournit les données d'érosion instantanée et est tirée du logiciel Caviciel (voir Lafleur, 2012). Ce logiciel est directement lié au système de détection des mesures vibratoires qui collecte les signaux de vibration sur quatre groupes turbine-alternateur à la centrale W. Il analyse ensuite la valeur quadratique moyenne (VQM) et la modulation des vibrations au palier guide inférieur (voir figure 1.5). Suite à un calcul d'une durée approximative de trois minutes tenant compte de la fonction de transmissibilité dans l'eau, une

estimation précise de l'érosion instantanée en kg/10 000 h est obtenue. Notons que le système lit les données de vibration un groupe à la fois en alternance. Si un groupe est arrêté, il passe au prochain groupe directement. L'intervalle de temps entre deux lectures sur un même groupe dépendra ainsi du fonctionnement des trois autres groupes (5-6-8) sur lesquels le système vibratoire est installé. Cet intervalle varie en général de 9 à 35 minutes si le groupe 7 n'est pas arrêté pour une longue période. Un exemple de fichier de données pour le groupe 7 est disponible au tableau 1.1. La figure 1.6 présente quant à elle un exemple de sortie graphique.

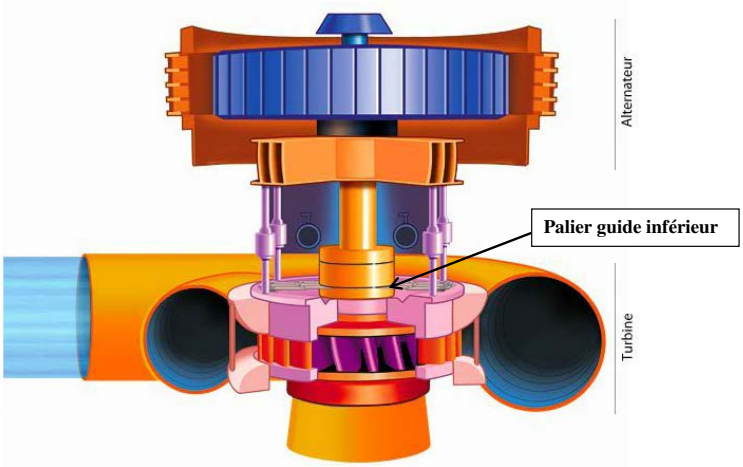


FIGURE 1.5. Vue en coupe d'un groupe turbine-alternateur

TABLEAU 1.1. Exemple de données extraites de la BD1 pour le groupe 7

Date et heure	Érosion instantanée
2011/10/31 04 :47 :16	13,713540
2011/10/31 04 :59 :03	14,778630
2011/10/31 05 :10 :51	13,120590
2011/10/31 05 :22 :40	12,043540
2011/10/31 05 :34 :28	11,185230
2011/10/31 05 :46 :15	11,925970
2011/10/31 06 :00 :23	9,695453
2011/10/31 06 :13 :20	6,279032
2011/10/31 06 :28 :37	6,979538
2011/10/31 06 :43 :53	6,094498
2011/10/31 06 :59 :10	4,717667

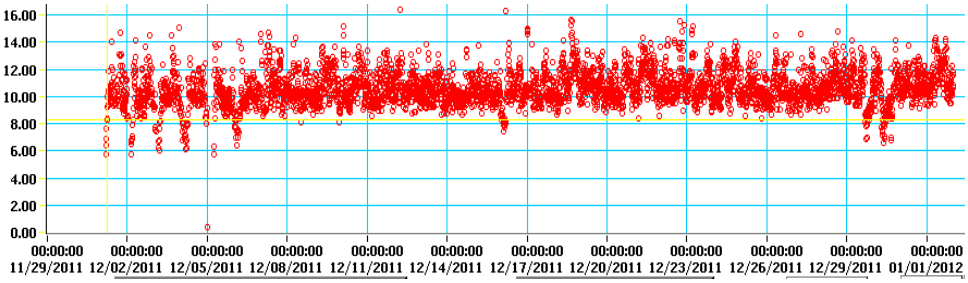


FIGURE 1.6. Exemple de sortie graphique de la BD1 pour la cavitation du groupe 7

1.2.3.2. *BD2 : Variables opératoires*

Les neuf variables opératoires à l'étude sont échantillonnées toutes les cinq minutes et sont disponibles dans la base de données de production nommée BD2. Un exemple de sortie est disponible au tableau 1.2. La hauteur de chute est ensuite calculée en soustrayant le niveau aval au niveau amont.

1.2.3.3. *BD3 : Zones de puissance*

La base de données BD3 contient pour sa part les informations nominales de la turbine étudiée, c'est-à-dire les caractéristiques telles que définies par le constructeur. Plus précisément, nous nous intéressons aux valeurs de puissance optimale selon la hauteur de chute. Sur la base de cette information, une classification en six zones de puissance a été ajoutée selon la hauteur de chute observée (voir figure 1.7). L'information à propos des conditions d'exploitation permises, à éviter et optimales est ainsi mise en relief.

La zone 1 correspond à la zone permise inférieure et les zones 2 et 6 correspondent aux zones à éviter inférieure et supérieure respectivement. La zone permise supérieure, quant à elle, est la zone d'opération principale et contient la majorité des données. Nous la divisons en trois zones (zones 3 à 5). Les frontières correspondant à la zone 4 centrale ont été déterminées arbitrairement sur la base d'une première investigation d'un sous-groupe des données. Cette zone 4, correspondant à la valeur de la puissance optimale $\pm$ 15 MW, est nommée ci-après « zone optimale ». La zone 3 correspond donc à la zone d'opération principale sous la zone optimale, et la zone 5 à celle au-dessus de la zone optimale.

1.2.3.4. *Mise en oeuvre de la BD de travail*

Pour chaque observation de cavitation, disponible à intervalle de temps irrégulier, nous avons associé la série d'observations de variables opératoires au temps le plus proche. Cette dernière étant échantillonnée aux cinq minutes, la distance

TABLEAU 1.2. Exemple de données de variables opératoires extraites de la BD2

Date	Courant	Débit turbiné	Niveau amont	Niveau aval	Ouverture	Production Mvar	Production MW	Puissance effective de stabilité	Tension
2011-12-05 10 :10	11,846251	232,9786395	175,042999	37,09	72,5625	42,390003	291,870026	358,011841	14,175
2011-12-05 10 :15	11,846251	232,9608211	175,033203	37,07	72,5625	40,230003	291,870026	357,617828	14,175
2011-12-05 10 :20	11,846251	233,4554652	175,033203	37,099998	72,5625	42,390003	292,410004	357,688141	14,175
2011-12-05 10 :25	11,846251	234,722748	175,042999	37,07	72,5625	42,390003	294,029999	357,825806	14,175
2011-12-05 10 :30	11,846251	234,0338711	175,042999	37,059998	73,125	42,390003	293,220001	357,673492	14,175
2011-12-05 10 :35	11,846251	232,8910687	175,033203	37,029999	72,3125	41,040001	291,870026	357,308807	14,175
2011-12-05 10 :40	11,846251	233,171005	175,033203	37,189999	72,3125	41,040001	291,870026	357,399597	14,175
2011-12-05 10 :45	11,846251	232,8625941	175,033203	37,139999	72,3125	43,200001	291,600006	357,311737	14,175

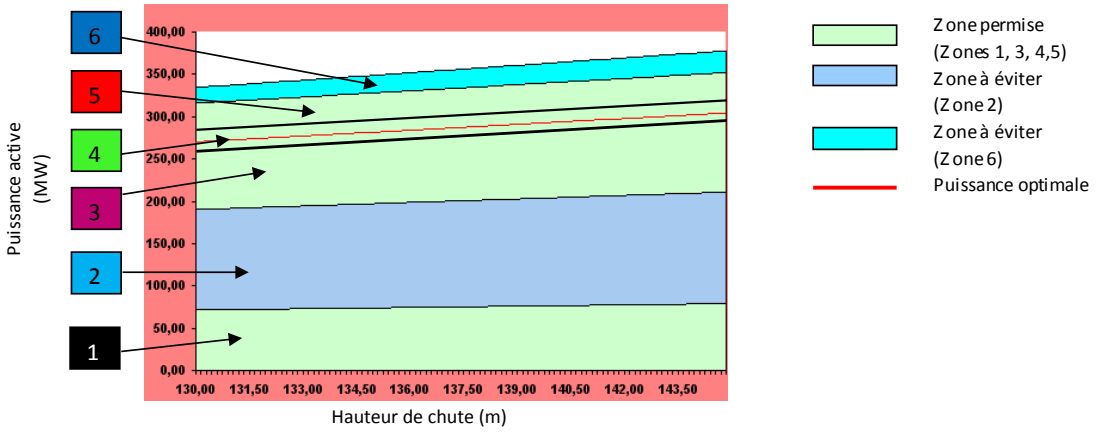


FIGURE 1.7. Classification en zones de puissance selon les zones permises et les zones à éviter

maximale entre la donnée de cavitation et les variables opératoires est de 2 minutes 30 secondes. Selon les experts, l'état d'un groupe turbine-alternateur ne devrait pas varier dans un si court laps de temps et nous considérons valide la correspondance entre la cavitation et les données opératoires. Le temps de référence conservé pour la base de données de travail est celui associé à la cavitation (tiré de la BD1). Un identifiant *semaine* (1 à 53), ainsi qu'un identifiant *zone de puissance* (Zones 1 à 6) ont été ajoutés. La définition de la semaine utilisée est une semaine de 7 jours exactement. La semaine 53 correspond au dernier jour de l'année (ou aux deux derniers jours de l'année si l'année est bissextile).

1.3. ANALYSE EXPLORATOIRE

L'analyse exploratoire est effectuée dans l'objectif d'avoir un portrait général de la cavitation du groupe 7 et de son lien avec les variables explicatives. En effet, nous souhaitons mettre en relief certaines caractéristiques qui permettront d'élaborer une piste de modélisation. Pour ce faire, les graphiques suivants ont été tracés pour chaque semaine de l'année 2011 : histogramme de la cavitation, série temporelle de la cavitation et des variables opératoires, nuages de points de chaque variable opératoire en lien avec la cavitation, groupée en fonction des zones de puissance, et série temporelle du fonctionnement des huit groupes partageant la chambre d'équilibre (1 à 8). Pour explorer le lien entre les différentes variables, nous avons aussi extrait les tableaux de corrélation entre les variables et les nuages de points matriciels.

TABLEAU 1.3. Tableau des corrélations : semaine 2 (8 au 15 janvier 2011)

	Cavit.	Courant	Débit	Amont	Aval	Ouv.	Mvar	MW	PuissEff	Tension	hChute
Cavit.	1,000	-0,598	-0,621	0,345	-0,611	-0,622	-0,080	-0,621	0,601	-0,115	0,606
Courant		1,000	0,978	-0,298	0,745	0,968	0,215	0,988	-0,720	0,149	-0,729
Débit			1,000	-0,399	0,843	0,996	0,192	0,995	-0,821	0,193	-0,830
Amont				1,000	-0,619	-0,416	0,497	-0,346	0,672	0,288	0,672
Aval					1,000	0,862	0,071	0,794	-0,991	0,201	-0,998
Ouv.						1,000	0,186	0,986	-0,841	0,197	-0,850
Mvar							1,000	0,216	-0,020	0,902	-0,022
MW								1,000	-0,770	0,197	-0,779
PuissEff									1,000	-0,165	0,994
Tension										1,000	-0,164
hChute											1,000

1.3.1. Lien entre les variables

Afin de mieux comprendre le lien entre les différentes variables opératoires, des tableaux de corrélation sont produits pour chaque semaine, ainsi que les nuages de points matriciels. Deux groupes de variables liées se profilent : le premier groupe réunit les variables *courant, débit, ouverture et puissance active*. Le deuxième groupe contient les variables *aval, hauteur de chute et puissance effective de stabilité*. Dans chaque groupe, chacune des variables est liée de manière similaire à la cavitation et c'est pourquoi les résultats seront présentés pour une seule variable par groupe, soient la puissance active pour le groupe 1 et la hauteur de chute pour le groupe 2. Notons que différentes abréviations sont utilisées ci-après pour certaines variables : ouverture de vannage (*Ouv.*), puissance réactive (*Mvar*), puissance active (*MW*), puissance effective de stabilité (*PuissEff*), hauteur de chute (*hChute*).

Un exemple de ces relations est présenté dans le nuage de points matriciel (voir figure 1.8) pour la semaine 2 (8 au 15 janvier 2011). Le groupe 1 correspond aux cercles rouges et le groupe 2 aux cercles bleus. Les couleurs du graphique correspondent aux zones de puissance telles que spécifiées à la figure 1.7. Pour la semaine 2, les corrélations pour le groupe 1 varie entre 0,968 et 0,996 (voir tableau 1.3). Les corrélations pour le groupe 2 sont respectivement de -0,991 (*puissance effective de stabilité* et *aval*), -0,998 (*hauteur de chute* et *aval*) et 0,994 (*puissance effective de stabilité* et *hauteur de chute*).

Les relations entre les variables évoluent au cours de l'année et c'est pourquoi l'analyse exploratoire s'attarde à une caractérisation par saison. Les groupes de variables présentés ci-haut ont cependant un lien stable au cours de l'année, quoique moins fort que ceux observés sur la durée d'une semaine. Le tableau de

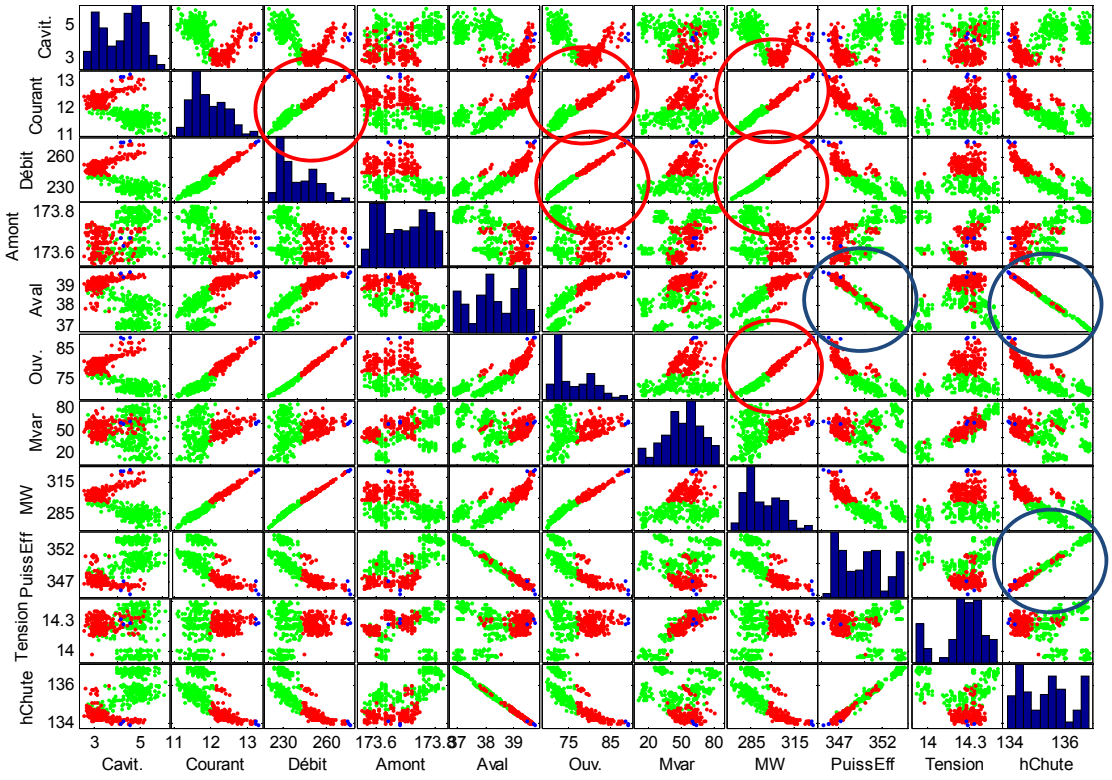


FIGURE 1.8. Nuage de points matriciel : semaine 2 (8 au 15 janvier 2011)

corrélation ci-après donne une mesure du lien entre ces variables pour l'ensemble des données étudiées (voir tableau 1.4).

TABLEAU 1.4. Tableau des corrélations : année complète 2011

	Cavit.	Courant	Débit	Amont	Aval	Ouv.	Mvar	MW	PuissEff	Tension	hChute
Cavit.	1,000	0,160	-0,111	0,348	-0,664	-0,220	-0,504	0,075	0,705	-0,054	0,718
Courant		1,000	0,918	0,140	-0,026	0,840	-0,105	0,975	0,110	0,559	0,095
Débit			1,000	-0,041	0,285	0,978	0,186	0,965	-0,229	0,674	-0,251
Amont				1,000	-0,087	-0,119	-0,117	0,107	0,597	-0,033	0,600
Aval					1,000	0,412	0,566	0,071	-0,834	0,235	-0,849
Ouv.						1,000	0,264	0,898	-0,366	0,679	-0,394
Mvar							1,000	0,058	-0,511	0,302	-0,517
MW								1,000	0,017	0,634	0,000
PuissEff									1,000	-0,204	0,986
Tension										1,000	-0,207
hChute											1,000

1.3.2. Résultats

L'analyse exploratoire met en relief le comportement hétérogène de la cavitation, qui semble lié à une certaine saisonnalité ou au mode opératoire des autres

groupes partageant la chambre d'équilibre (Marche-Arrêt). Deux périodes en particulier présentent des différences, que nous nommons saison hivernale (décembre à mars) et saison estivale (juin à octobre). Nous observons aussi deux zones de transition (avril-mai et novembre) où le comportement de la cavitation est plus difficile à cerner.

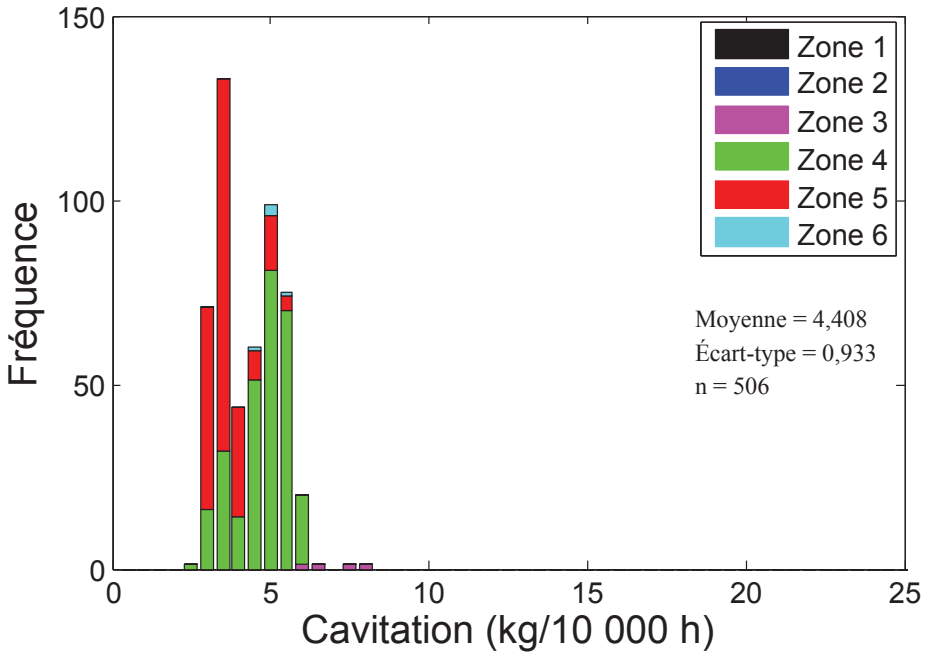
Lorsque nous observons les histogrammes de la cavitation par semaine, nous remarquons que la cavitation présente typiquement une distribution plurimodale, caractérisée différemment selon la saison (voir figures 1.9(a) et 1.10(a)). En effet, le comportement « Hiver » présente une étendue plus serrée et se concentre dans des zones basses de cavitation (voir figure 1.9(b)). Le comportement « Été », quant à lui, présente une étendue de cavitation beaucoup plus grande (4 à 20 kg/10 000 h typiquement), et est caractérisé par un comportement cyclique où apparaissent des pics de haute cavitation à intervalles réguliers (voir figure 1.10(b)).

Un autre aspect de la caractérisation saisonnière concerne le lien entre la cavitation et les variables explicatives. Les nuages de points nous permettent en effet de déceler certaines tendances : alors que la puissance active semble mieux expliquer la cavitation pendant l'hiver, la hauteur de chute caractérise mieux la cavitation pendant l'été (voir figures 1.11 et 1.12). Nous observons aussi une rupture dans la direction de la relation entre la cavitation et la puissance active pendant l'hiver, assez bien mise en valeur par les zones de puissance.

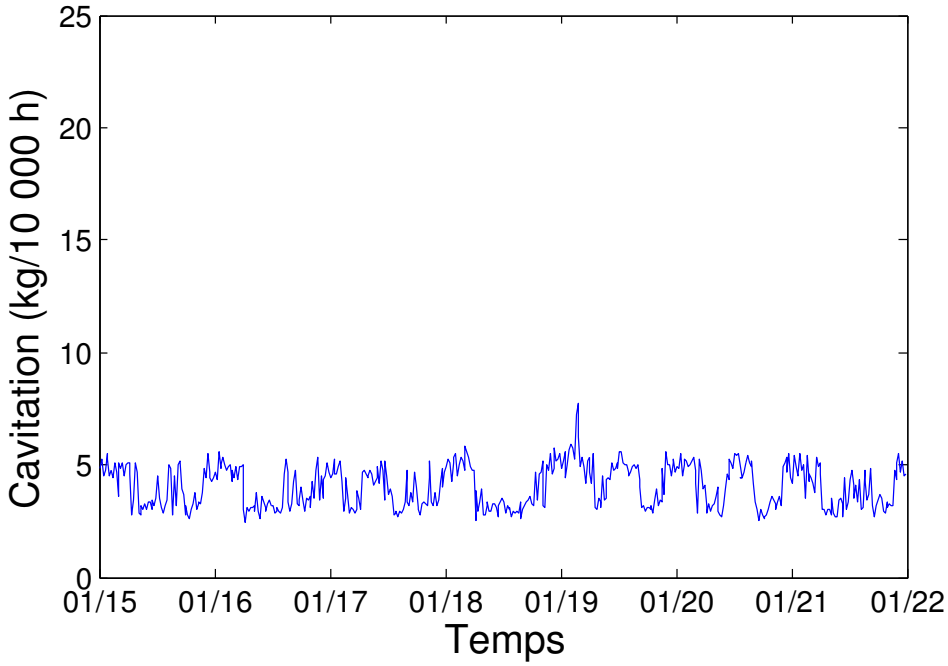
Comme la hauteur de chute paraît être une variable significative, en particulier pendant la période estivale, nous avons choisi d'approfondir l'étude du fonctionnement des autres groupes turbine-alternateur. En effet, huit groupes partagent la même chambre d'équilibre à la centrale W, tel que vu à la section 1.2.2 (voir figure 1.4). Ainsi, le fonctionnement de l'un ou l'autre des groupes influence directement le niveau aval et possiblement la cavitation. Comme le niveau amont est assez stable sur la durée d'une semaine, nous supposons donc que le fonctionnement des groupes influence principalement la hauteur de chute.

Il est important de noter qu'un seul limnimètre mesure le niveau aval et est positionné vis-à-vis le groupe 4 (voir figure 1.4). Cette mesure est celle utilisée dans cette étude. Cependant, selon le mode des fonctionnements des groupes, nous devons garder en tête que le niveau aval peut varier d'un endroit à l'autre de la chambre d'équilibre.

Nous avons choisi d'observer la cavitation du groupe 7, ainsi que le fonctionnement des groupes 1 à 8, avec le critère tel que spécifié au tableau 1.5. Notons que le critère de marche à vide est basé sur le tiers du débit optimal selon les données nominales de la turbine.

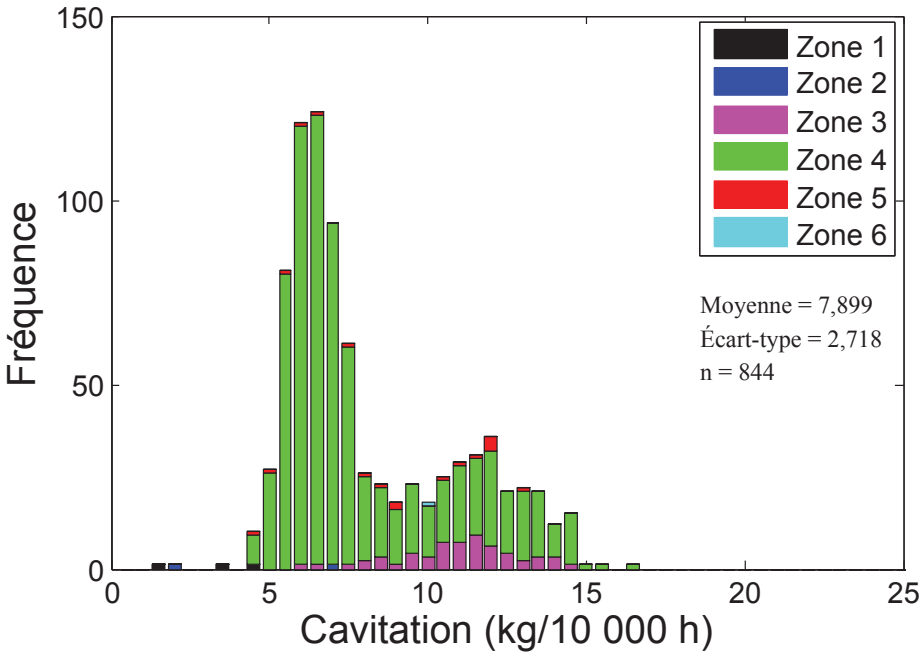


(a) Histogramme de la cavitation
semaine 3 (15 au 21 janvier 2011)

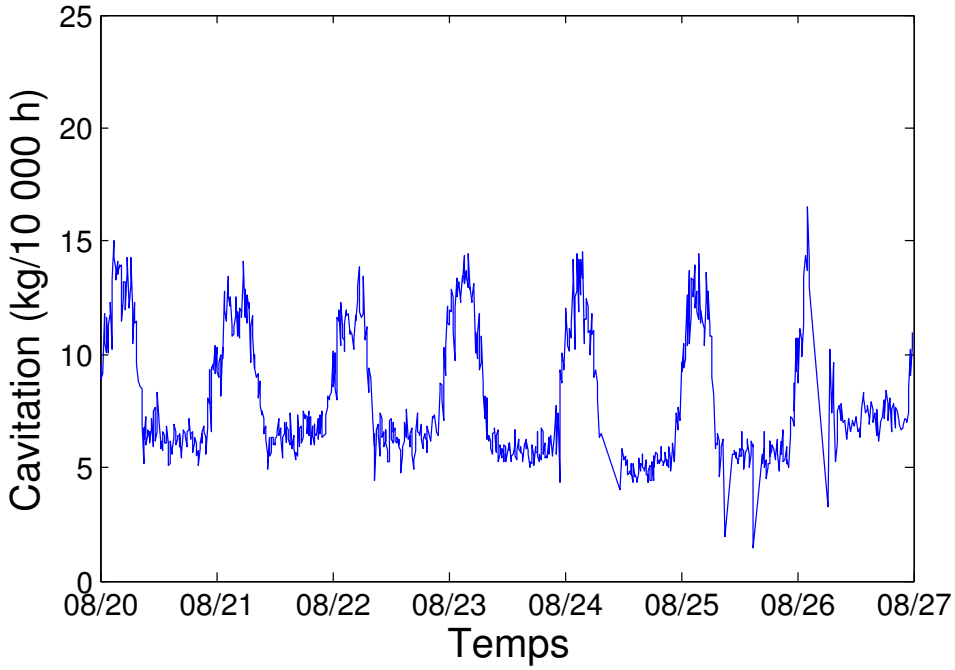


(b) Série temporelle de la cavitation :
semaine 3 (15 au 21 janvier 2011)

FIGURE 1.9. Comportement type de la cavitation pendant la saison hivernale



(a) Histogramme de la cavitation :
semaine 34 (20 au 26 août 2011)



(b) Série temporelle de la cavitation :
semaine 34 (20 au 26 août 2011)

FIGURE 1.10. Comportement type de la cavitation pendant la saison estivale

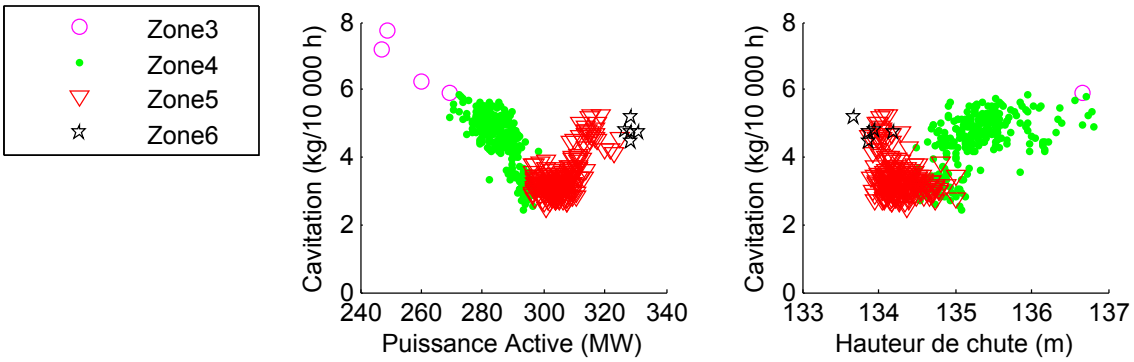


FIGURE 1.11. Relation entre la cavitation et deux variables opératoires : puissance active et hauteur de chute (Hiver : 15 au 21 janvier 2011)

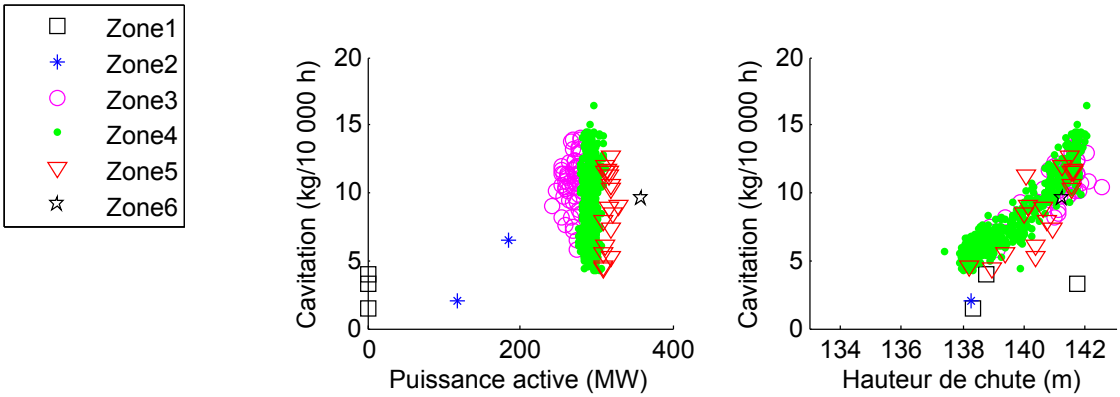


FIGURE 1.12. Relation entre la cavitation et deux variables opératoires : puissance active et hauteur de chute (Été : 20 au 26 août 2011)

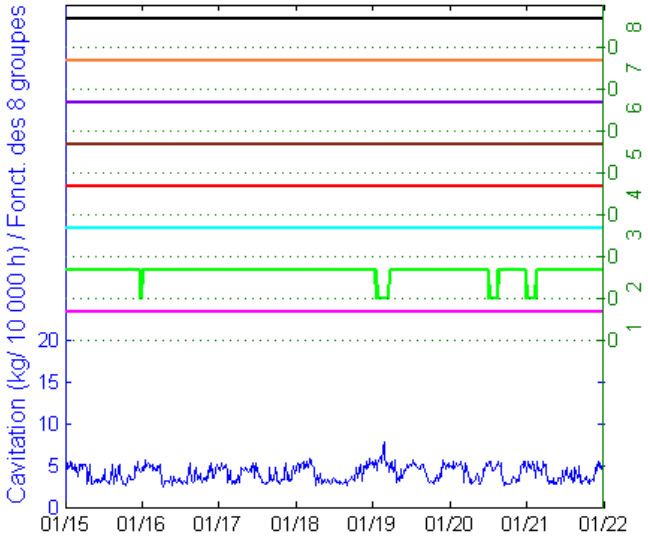
TABLEAU 1.5. Critère de fonctionnement des groupes turbine-alternateur

État du groupe	Critère
Arrêt = 1	Débit = 0 & MW = 0
Marche à vide = 2	0 < Débit < 70 & MW = 0
Marche = 3	Sinon

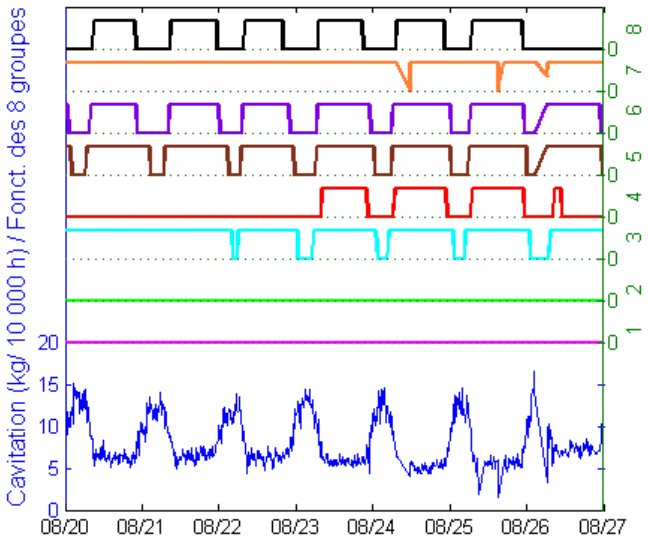
La figure 1.13 présente le comportement de la cavitation du groupe 7 en parallèle avec le fonctionnement des huit groupes de la chambre d'équilibre. La série temporelle dans le bas des deux figures représente les mesures de cavitation du groupe 7. Les conditions d'opération (Marche-Arrêt) des groupes 1 à 8 sont schématisées dans la partie supérieure du graphique par les lignes de couleur. Si

la ligne colorée est confondue avec la ligne pointillée, le groupe est à l'arrêt, sinon le groupe est en fonctionnement. Nous pouvons voir que, pendant la période hivernale, les groupes sont presque toujours en fonctionnement et peu d'arrêts sont observés, contrairement à la période estivale. Celle-ci est en effet caractérisée par des arrêts simultanés de plusieurs groupes à la fois. Nous observons ainsi un comportement très intéressant : il semble y avoir un lien entre l'arrêt de certains groupes pendant l'été et la présence de pics de cavitation sur le groupe 7 (voir figure 1.13(b)). À l'opposé, les fluctuations de cavitation pendant l'hiver ne semblent pas être causées par un arrêt des autres groupes : nous n'observons pas les pics de cavitation malgré certains arrêts (voir figure 1.13(a)).

Quoique les deux comportements « Hiver » et « Été » semblent bien caractériser chacune des saisons, un regard sur l'évolution de chacun des graphiques, par période d'une semaine et sur l'ensemble de l'année, souligne l'idée que des zones de transition sont présentes. Les experts préférant s'appuyer sur la piste du mode opératoire (Marche-Arrêt de tous les groupes), plutôt que sur la composante temporelle, la piste retenue pour la suite de la modélisation sera l'opération de ces huit groupes turbine-alternateur. En effet, comme les conditions d'exploitation sont très variables d'une année à l'autre, l'idée de caractériser la cavitation en se basant sur la périodicité du comportement opératoire semblait trop aléatoire. Le chapitre suivant s'attardera à développer le lien entre les modes opératoires et la cavitation du groupe 7.



(a) semaine 3 (15 au 21 janvier 2011)



(b) semaine 34 (20 au 26 août 2011)

FIGURE 1.13. Fonctionnement des 8 groupes et cavitation du groupe 7

Chapitre 2

ANALYSE SELON LES MODES OPÉRATOIRES

Comme nous l'avons vu au chapitre précédent, le fonctionnement des autres groupes partageant la chambre d'équilibre semble avoir un impact sur les variations de la cavitation du groupe 7. Par exemple, un arrêt simultané des autres groupes correspond à un pic de cavitation du groupe 7. Dans ce chapitre, nous aborderons la problématique sous un angle différent : nous tentons en effet de nous éloigner de la composante temporelle pour étudier le comportement de la cavitation en lien avec les modes opératoires de la centrale, c'est-à-dire les différentes configurations de Marche-Arrêt des groupes dans la chambre d'équilibre. Pour ce faire, nous regrouperons les différents modes opératoires en grappes similaires du point de vue des variables explicatives à l'aide du regroupement hiérarchique. Une étude du comportement de la cavitation à l'intérieur de chacune des grappes sera ensuite présentée.

Notons que les zones de puissance étudiées pour la suite de la modélisation correspondent aux zones 3 à 5 explicitées à la section 1.2.3.3 (zone d'opération principale). Les valeurs de cavitation correspondant aux zones 1, 2 et 6 sont peu nombreuses : nous en dénombrons $n = 215$ pour l'année 2011. Le comportement dans ces zones d'opération fera l'objet d'une étude ultérieure et la suite de ce mémoire s'intéresse plutôt à la modélisation dans la zone d'opération principale. Le nombre de données de cavitation pour l'année 2011 passe donc de 35 582 pour l'ensemble des six zones à 35 367 pour la zone d'opération principale (zones 3 à 5).

2.1. Mode opératoire

2.1.1. Définition

Commençons par définir exactement ce que nous entendons par mode opératoire. La configuration de la centrale hydroélectrique où se trouve le groupe turbine-alternateur étudié (groupe 7) comprend une chambre d'équilibre, c'est-à-dire un bassin dans lequel se déverse l'eau qui traverse les turbines hydroélectriques. Huit groupes turbine-alternateur se déversent dans cette chambre d'équilibre (voir figure 1.4). Dans les chapitres suivants, ce que nous appellons le mode opératoire correspond à la configuration Marche-Arrêt de ces huit groupes. Nous considérons un groupe à l'arrêt si le débit et la puissance active égalent 0. Le groupe est considéré en marche dans les autres cas. Notons que l'état de marche inclut aussi la marche à vide (la turbine est en fonctionnement, mais ne produit pas de puissance). La notation utilisée contient seulement les groupes en marche. Par exemple, pour le mode opératoire « 357 », les groupes 3, 5 et 7 sont en fonctionnement, alors que les groupes 1, 2, 4, 6 et 8 sont à l'arrêt.

2.1.2. Caractéristiques pour l'année 2011

Des directives d'exploitation imposent un ordre de démarrage et d'arrêt pour la centrale étudiée. La priorité de démarrage des groupes dans la chambre d'équilibre d'intérêt s'établit comme suit : 7-3-5-6-1-4-2-8 (voir Moreau, 2012). L'arrêt des groupes se fait selon l'ordre inverse. Le nombre de possibilités de modes opératoires est donc réduit : toutes les combinaisons de Marche-Arrêt ne sont pas possibles. Pour l'année 2011, 76 modes opératoires sont observés et chaque donnée de cavitation du groupe 7 est associée au mode opératoire observé à l'instant correspondant. La liste complète des modes opératoires et leur fréquence pour l'année 2011 est disponible à l'annexe A. Le tableau 2.1 liste les quatre modes opératoires les plus fréquemment observés lorsque le groupe 7 présentait de la cavitation. Le mode opératoire le plus fréquent ($N = 11\,428\,(32,31\%)$) est le mode « 12345678 », c'est-à-dire lorsque les huit groupes sont en fonctionnement.

Tableau 2.1. Modes opératoires les plus fréquemment observés (année 2011)

Mode opératoire	Fréquence	Fréquence (%)
« 12345678 »	11428	32,31
« 1345678 »	3277	9,27
« 345678 »	1875	5,30
« 357 »	1749	4,95

2.1.3. Modes opératoires et cavitation

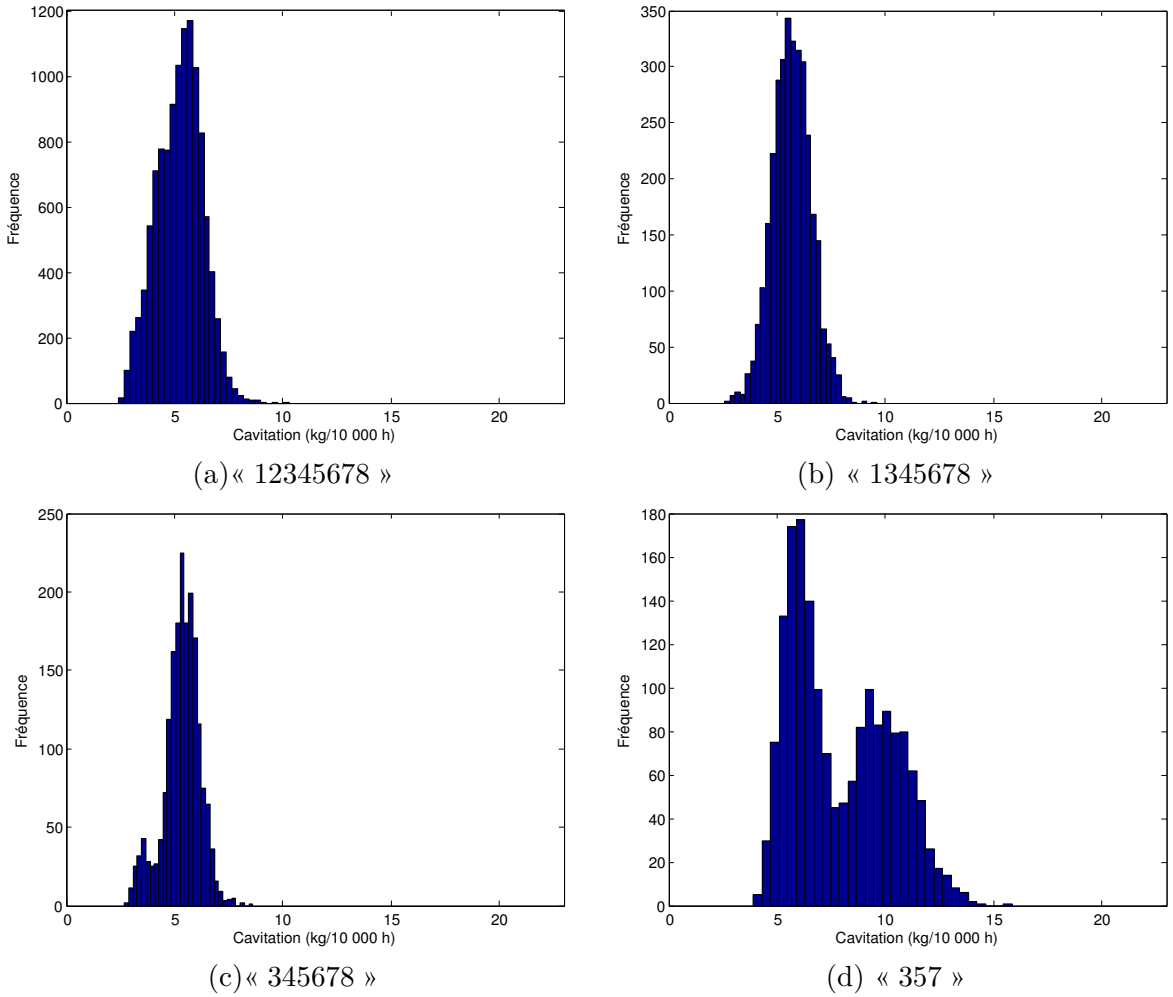
Le comportement de la cavitation du groupe 7 peut varier énormément selon le contexte du mode opératoire. Par exemple, nous pouvons voir à la figure 2.1(d) que la cavitation sous le mode opératoire « 357 »présente une distribution bimodale, alors que la distribution sous le mode « 12345678 » est unimodale et légèrement asymétrique (voir figure 2.1(a)). En gardant en tête l'objectif de modéliser la cavitation du groupe 7 selon les variables explicatives d'opération, le but est ici de caractériser les modes opératoires selon ces mêmes variables explicatives et de les rassembler en sous-groupes présentant des caractéristiques homogènes, appelés *grappes*. Nous présupposons ainsi que la cavitation, à l'intérieur de chaque grappe, sera plus homogène et que son comportement se prêtera plus facilement à la modélisation.

2.2. Regroupement hiérarchique

La méthode de regroupement hiérarchique (*hierarchical clustering*) est une méthode exploratoire permettant de regrouper des observations selon des caractéristiques similaires (voir Mooi et Sarstedt, 2011). Cette méthode, très utilisée par exemple pour segmenter la clientèle en marketing, ou pour l'analyse de l'expression des gènes, a l'avantage de ne requérir aucun présupposé sur le comportement des données. Seules une mesure de distance et une fonction de lien servant à agglomérer les grappes sont nécessaires. De plus, un autre avantage non négligeable de cette méthode est qu'elle ne nécessite pas de spécifier à l'avance le nombre de grappes voulues (voir Ward Jr, 1963), par opposition à d'autres méthodes de partitionnement comme la méthode des k-moyennes par exemple. Pour d'autres méthodes de partitionnement, le lecteur peut se référer à Mooi et Sarstedt (2011).

2.2.1. Caractéristiques d'intérêt des modes opératoires

Afin de regrouper les modes opératoires en grappes homogènes, la première étape consiste à clarifier les caractéristiques qui seront utilisées pour grouper les modes opératoires. Pour chaque mode opératoire, nous utilisons la moyenne, l'écart-type et la médiane des 10 variables explicatives. Nous rappelons que ces dernières sont : le courant, le débit, le niveau amont, le niveau aval, l'ouverture, la puissance active, la puissance réactive, la puissance effective de stabilité, la tension et la hauteur de chute. Il existe plusieurs choix possibles pour décrire la distribution des modes opératoires et ainsi choisir les caractéristiques d'intérêt. Le choix de deux mesures de tendance centrale s'explique ici par la nécessité de



(a)« 12345678 » (b) « 1345678 »

(c)« 345678 » (d) « 357 »

FIGURE 2.1. Histogrammes de la cavitation du groupe 7 pour les modes opératoires les plus fréquents de l'année 2011

tenir compte d'une certaine asymétrie dans la distribution. Une alternative aurait été d'employer le coefficient d'asymétrie ou d'utiliser les quartiles de distribution empirique.

Pour chaque mode opératoire, un nombre variable d'observations est disponible. Par exemple, pour le mode opératoire le plus fréquemment utilisé « 12345678 », 11 428 observations ont été utilisées pour calculer les 30 statistiques. La fréquence du mode opératoire a par ailleurs été ajoutée comme caractéristique d'intérêt (ici, $11\,428/35\,367 = 32,31\,\%$ par exemple). L'objectif de cet ajout est de mettre en relief les modes opératoires rares, qui sont plus sujets à ne pas respecter l'ordre de démarrage des groupes (voir section 2.1.2) et donc à produire des conditions de cavitation inhabituelles. Ainsi, à chacun des 76 modes opératoires observés pour l'année 2011 dans la zone d'opération principale du groupe 7 correspond une observation multidimensionnelle de 31 caractéristiques.

2.2.2. Principe

Le regroupement hiérarchique est une procédure itérative, qui permet de grouper les éléments en grappes en minimisant la distance entre ceux-ci. Supposons qu'on cherche à agglomérer n éléments. L'élément de départ est le singleton, c'est-à-dire qu'à la première étape, chaque élément (ici chaque mode opératoire) est considéré comme une grappe. L'étape suivante consiste à faire passer le nombre de grappes de n à $n-1$ en groupant deux de ces éléments en une seule grappe et ce, en s'assurant de perdre le moins d'information possible selon un certain critère choisi. Les $n-1$ grappes restantes sont ensuite examinées pour déterminer si un troisième élément doit être ajouté à la première paire, ou si deux autres éléments doivent être réunis pour former la grappe suivante et ainsi, réduire le nombre de grappes à $n-2$. Le processus est poursuivi jusqu'à la formation d'une seule grappe, si désiré.

La décision d'agglomérer une grappe avec l'une ou l'autre des grappes requiert de se doter d'une distance pour mesurer la similarité (ou dissimilarité) entre deux éléments et d'une fonction de lien permettant d'agglomérer les éléments d'une grappe. Il existe différentes mesures de similarité dans la littérature, telles que les distances euclidienne, de Mahanalobis, cityblock, de Chebychev, etc. (voir Tan *et al.*, 2006, chap. 9). Les différentes fonctions de lien comprennent entre autres les fonctions simple (plus petite distance), complète (plus grande distance), centroïde (distance entre les barycentres) et le critère de Ward (basé sur l'inertie). La figure 2.2 présente graphiquement des exemples de distances et de fonctions de lien.

2.2.3. Mesure de similarité

Dans le cas du regroupement hiérarchique des modes opératoires, nous avons choisi la mesure de similarité usuelle, c'est-à-dire la distance euclidienne, qui correspond à :

$$d_{euc} = ||a - b|| = \sqrt{\sum_{k}(a_k - b_k)^2}, \tag{2.2.1}$$

avec $k = 1, ..., 31$. Afin de retirer l'influence des unités sur la distance, celle-ci a été calculée pour chaque paire d'éléments après la standardisation des caractéristiques.

2.2.4. Fonction de lien

Tel que vu dans Ferreira et Hitchcock (2009), le choix de la fonction de lien dépend fortement de la forme des données. Par exemple, la méthode simple est reconnue pour mal performer, sauf dans le cas de grappes en forme de longues

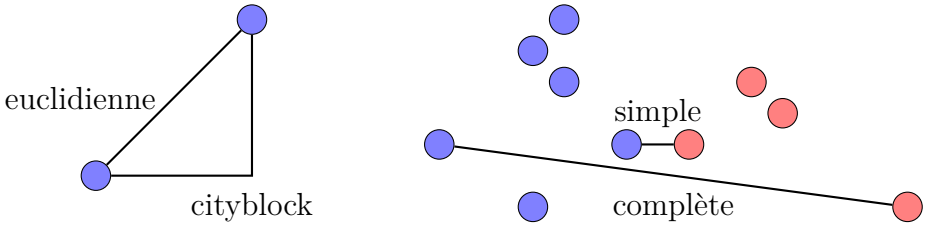


FIGURE 2.2. Exemples de mesures de similarité et de fonctions de lien

chaînes. De plus, cette méthode est très sensible aux valeurs aberrantes, tout comme la méthode complète (voir Kuiper et Fisher, 1975). Par ailleurs, un des problèmes avec la méthode centroïde est la possibilité d'inversion. Ceci apparaît lorsque la distance entre une paire de grappes est plus petite que celle d'une autre paire qui a été fusionnée précédemment. La fusion n'est donc pas une fonction monotone, et ceci rend les résultats difficiles à interpréter (voir Morgan et Ray, 1995).

Dans le cas du regroupement hiérarchique des modes opératoires, la fonction de lien choisie est le critère de Ward (voir Ward Jr, 1963). Celui-ci s'appuie non pas sur un critère de distance pour grouper les éléments, mais sur la minimisation de la variance intra-grappe (ou inertie). En général, cette méthode performe le mieux si on souhaite mettre en évidence des grappes de taille à peu près égale, comme c'est le cas ici (voir Ferreira et Hitchcock, 2009; Kuiper et Fisher, 1975).

Soit $\boldsymbol{x_{ij}}$ l'observation multidimensionnelle j dans la grappe i ($\boldsymbol{x_{ij}}$ est un vecteur de dimension 31). À chaque étape, la somme suivante est calculée pour toute paire de grappes et le choix du regroupement entre deux grappes correspond à la valeur minimale de cette somme :

$$SSE = \sum_{i=1}^{G} \sum_{j=1}^{n_i} (\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_{\boldsymbol{i.}})^T (\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_{\boldsymbol{i.}}), \tag{2.2.2}$$

où $\boldsymbol{x_{ij}}$ est le j^e élément dans la i^e grappe, n_i est le nombre d'éléments dans la i^e grappe, G est le nombre de grappes et $\bar{\boldsymbol{x}}_{\boldsymbol{i.}}$ correspond au vecteur des moyennes pour chaque caractéristique dans la grappe i.

2.2.5. Dendrogramme

Le dendrogramme (voir figure 2.3) sert à représenter visuellement le regroupement hiérarchique. On y voit les différentes grappes et la manière dont elles ont été regroupées à chaque étape de l'algorithme. Chaque lien horizontal représente une étape où deux grappes ont été liées. La hauteur du lien vertical représente la distance entre ces deux grappes. L'étape de base, où chaque observation consiste en une grappe, se lit sur l'axe des x. Le nombre de grappes final se décide en coupant

horizontalement le dendrogramme. Le choix du nombre de grappes est souvent subjectif et guidé par l'expertise du chercheur. Un indicateur pour le choix du nombre de grappes est la distance entre les grappes aux différentes étapes . Ainsi, le nombre de grappes choisi est un compromis entre l'homogénéité des grappes inférieures et la distance entre les grappes choisies. Il est important de s'assurer que les résultats sont aisément interprétables. Le nombre de grappes doit ainsi être assez restreint pour permettre l'utilisation pratique, mais assez grand pour que chacune des grappes comportent leur spécificité propre (voir Mooi et Sarstedt, 2011).

2.2.6. Choix du nombre de grappes

Dans le cas des modes opératoires, nous avons choisi cinq grappes de modes opératoires telles que représentées à la figure 2.3. En effet, si nous choisissons six grappes au lieu de cinq, le nombre de données incluses dans une des grappes est très petit. De fait, la partie gauche de la grappe 4 (en rouge) qui contient les modes opératoires « 23457 », « 237 », « 267 » et « 2347 » ne compte que sept données. En elle-même, la grappe 4 réunit des modes opératoires moins fréquemment observables ($n = 223$). La séparation de cette grappe en d'autres plus petites grappes rend ainsi l'interprétation des résultats plus difficile. Par ailleurs, l'homogénéité des cinq grappes semble adéquate d'un point de vue opératoire. Par exemple, la grappe 1 (mauve) contient majoritairement des modes opératoires avec peu de groupes turbine-alternateur en fonctionnement, alors que la grappe 5 (vert) contient pour sa part des modes opératoires avec un nombre élevé de groupes en marche.

2.2.7. Résumé de la procédure

Le schéma à la figure 2.4 présente les différentes étapes requises pour l'analyse par regroupement hiérarchique et l'application au cas à l'étude.

2.3. Résultats en fonction des grappes

Il est intéressant de noter que la caractérisation par grappe permet de bien délimiter les pics et les creux de la cavitation sur les séries temporelles. La figure 2.5 montre des exemples de certaines semaines au cours de l'année 2011, avec la grappe correspondant au mode opératoire observé au moment de la mesure de cavitation instantanée. En (a), le graphique d'une semaine d'hiver montre que la cavitation est relativement homogène. La majorité des données de cavitation sont comprises entre 3 et 8 kg/10 000 h et correspondent à la grappe 5, constituée

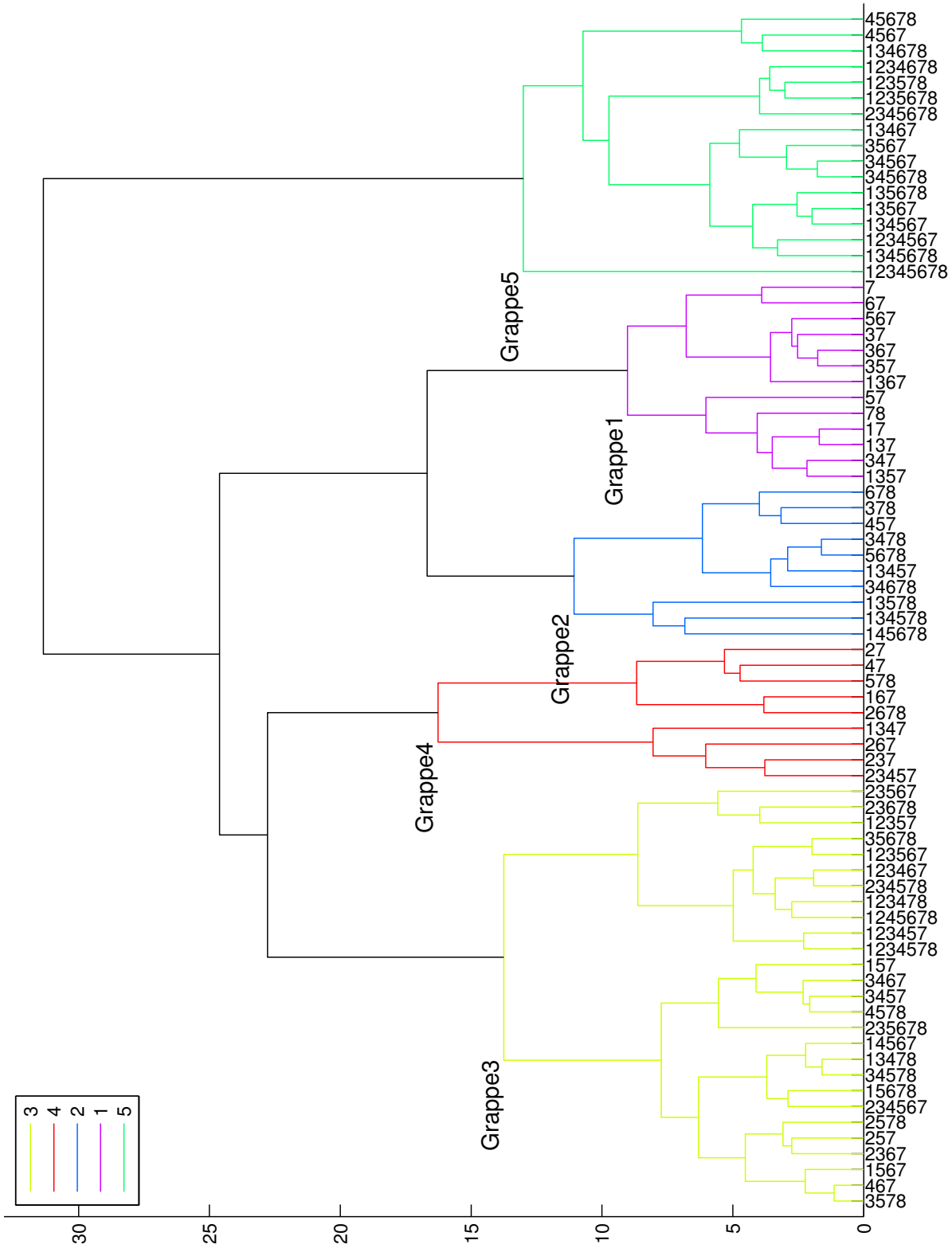


FIGURE 2.3. Dendrogramme des modes opératoires

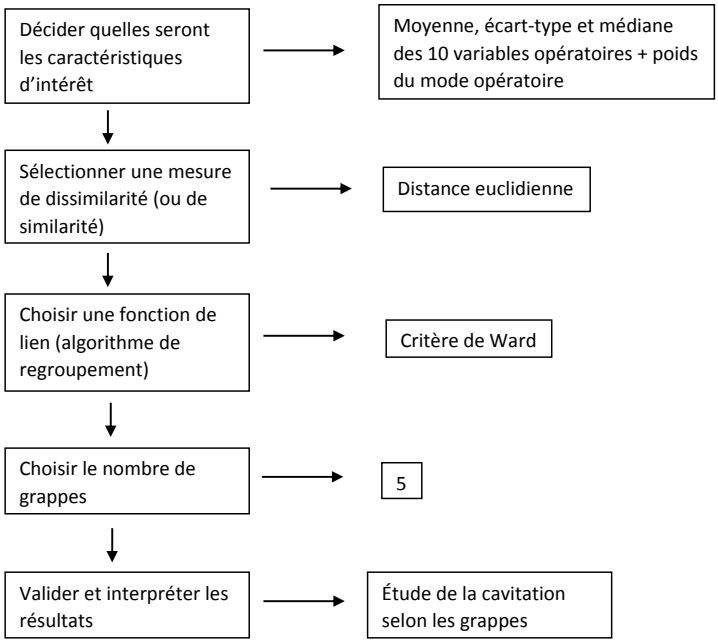
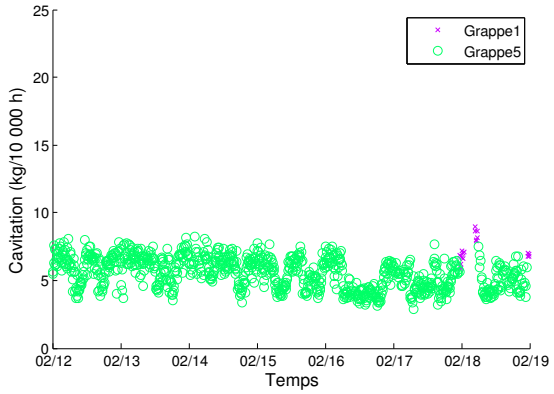


FIGURE 2.4. Procédure d'analyse par regroupement hiérarchique

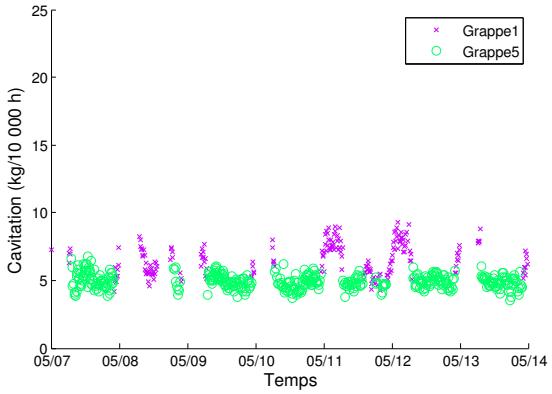
principalement par un nombre élevé de groupes turbine-alternateur en marche simultanément. En (b), nous observons une modification du comportement de la cavitation, notamment caractérisée par l'alternance de modes opératoires inclus dans la grappe 1 et la grappe 5. Ces deux grappes contiennent les modes opératoires les plus fréquemment utilisés (respectivement $n = 23\,901$ pour la grappe 5 et $n = 7456$ pour la grappe 1). Cette alternance est typique de la période printanière (mi-avril à mi-juin).

La période estivale (c), quant à elle, met en relief une variété plus grande de grappes opératoires. Les pics sont généralement représentés par la grappe 1, mais les grappes caractérisant les creux peuvent alterner entre les grappes 2, 3, 4 ou 5. Pour ce qui est de la période automnale (d), nous remarquons en particulier l'absence presque complète de valeurs de cavitation associées à la grappe 5, les valeurs basses de cavitation étant notamment associées aux grappes 2 ou 3. En fait, les valeurs de cavitation associées à la grappe 5 réapparaissent progressivement vers la fin de l'année, au moment où le comportement typiquement « hivernal » se rétablit, c'est-à-dire une cavitation très basse (entre 3 et 8 kg/10 000 h) et très homogène, comme en (a).

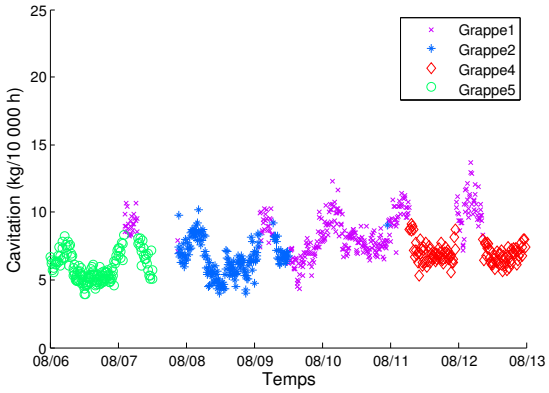
Par ailleurs, les histogrammes de la cavitation par grappes sont présentés à la figure 2.6. La présentation des grappes respecte l'ordre de gauche à droite sur le dendrogramme 2.3. Nous remarquons que la distribution de la cavitation par



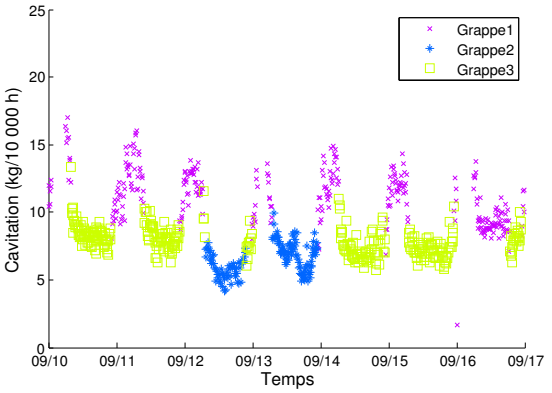
(a) Exemple pour l'hiver : semaine 7



(b) Exemple pour le printemps : semaine 19



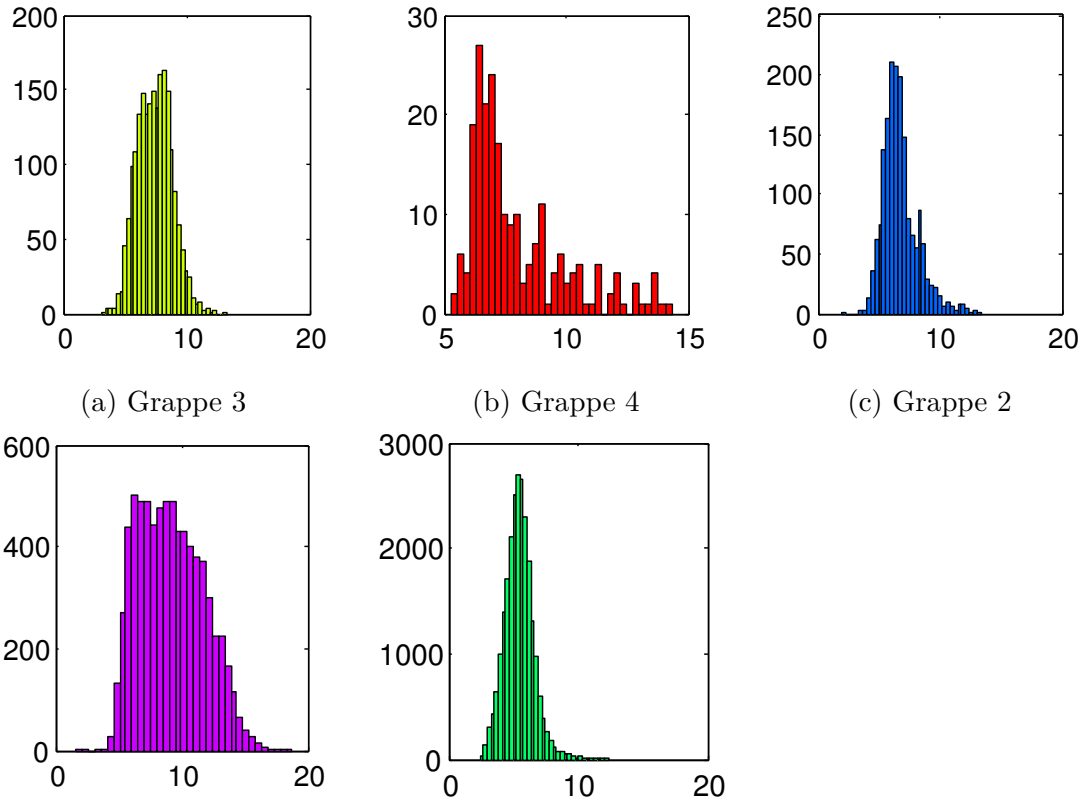
(c) Exemple pour l'été : semaine 32



(d) Exemple pour l'automne : semaine 37

FIGURE 2.5. Caractérisation de la cavitation en fonction d'un regroupement en cinq grappes

grappe ne s'apparente pas à des distributions statistiques usuelles (telles normale, gamma, Gumbel, etc.) pour la plupart des grappes. En effet, à l'exception de la grappe 5, chacune des autres grappes semble présenter une distribution plurimodale, qui laisse penser qu'à l'intérieur de chacune des grappes, les observations de cavitation peuvent provenir d'une sous-population statistique distincte. Notons en outre que les cinq distributions présentent une certaine asymétrie positive. Le chapitre 3 s'attardera à modéliser ces distributions, par le biais de modèles de mélanges de lois.



(a) Grappe 3 (b) Grappe 4 (c) Grappe 2

(d) Grappe 1 (e) Grappe 5

FIGURE 2.6. Histogrammes de la cavitation en fonction d'un regroupement en cinq grappes

Chapitre 3

MÉLANGE DE LOIS ET ALGORITHME ESPÉRANCE-MAXIMISATION

La plurimodalité apparente de certains des histogrammes par grappe (voir figure 2.6) nous amène à explorer la distribution de la cavitation à travers des modèles de mélanges de lois. Dans ce chapitre, nous expliciterons plus spécifiquement les modèles de mélange de lois Burr de type XII, en utilisant l'algorithme Espérance-Maximisation pour l'estimation des paramètres. Ce modèle sera par la suite appliqué aux cinq grappes, avec un nombre variable de sous-populations. Pour déterminer le nombre de sous-populations adéquat dans chacune des grappes, nous utiliserons le critère BIC (*Bayesian Information Criterion*). La dernière étape de la modélisation consistera à estimer la cavitation dans chaque grappe par la moyenne théorique, donnée par l'espérance du mélange de lois optimal trouvé précédemment.

3.1. Choix des méthodes

Un des avantages des modèles de mélange de lois est que ceux-ci offrent une très grande souplesse. En effet, ils permettent de reproduire plusieurs modes, différents types d'asymétrie, ainsi que des comportements asymptotiques variés. La distribution de certaines grappes présentant une forme plurimodale, ces types de modèles ont donc été choisis de manière à représenter formellement une telle hétérogénéité. Par ailleurs, notons que les mélanges de lois ont déjà été utilisés à l'IREQ dans le contexte de modélisation des séries de pointes et volumes de crues printanières (voir Evin *et al.*, 2011). Dans cet article, les auteurs abordent la modélisation de données hydrologiques par le biais de mélange de lois normale, gamma et Gumbel,(avec et sans dépendance markovienne) en adoptant la perspective bayésienne pour l'estimation des paramètres. Dans le cadre de ce mémoire, pour faciliter le traitement du nombre élevé de données (plus de 30 000

par année) et étant donné le temps de calcul nécessaire à l'application de ce type de méthodes, nous avons choisi de développer plutôt l'estimation des paramètres par l'intermédiaire de l'algorithme EM, présenté par Dempster *et al.* (1977).

3.2. Mélanges de lois

L'idée qui sous-tend les modèles de type mélange de lois est que pour la population observée, il existe différentes sous-populations desquelles proviennent les observations. Chacune de ces sous-populations est généralement caractérisée par une loi provenant d'une même famille paramétrique, mais avec des paramètres différents (voir Frühwirth-Schnatter, 2006).

Par exemple, supposons que la population des Y est partagée en plusieurs sous-populations distinctes, $G_1, G_2, ...G_J$ avec une proportion w_j $(j = 1, ..., J)$ associée. La collecte des données d'un échantillon tiré aléatoirement de cette population consiste à enregister les valeurs de Y, mais aussi de l'indicateur G d'appartenance à la sous-population. La probabilité d'appartenance à une certaine sous-population G_j est donnée par w_j. Ainsi, conditionnellement à son appartenance à la sous-population G_j, Y est une variable aléatoire avec une fonction de densité $f_j(Y \mid \boldsymbol{\theta_j})$, où $\boldsymbol{\theta_j}$ est un vecteur tel que $\boldsymbol{\theta_j} = (\theta_{j1}, ..., \theta_{jl})$, avec l le nombre de paramètres caractérisant la densité f_j. La loi de probabilité conjointe de Y et G est donc donnée par :

$$\mathbb{P}(Y, G) = \mathbb{P}(Y \mid G)\,\mathbb{P}(G) = f_j(y \mid \boldsymbol{\theta_j})\,\mathbb{P}(G = G_j).$$

En général, les modèles de mélanges de lois sont utilisés dans un contexte où la collecte des données sur l'appartenance de Y à la sous-population G_j n'a pu être effectuée. Seules les données sur Y sont disponibles. La densité marginale de Y est conséquemment donnée par :

$$\begin{aligned} f_Y(y \mid \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_J}) &= \sum_{j=1}^{J} f_j(y \mid \boldsymbol{\theta_j})\,\mathbb{P}(G = G_j) \\ &= f_1(y \mid \boldsymbol{\theta_1})\,w_1 + \cdots + f_J(y \mid \boldsymbol{\theta_J})\,w_J. \end{aligned} \tag{3.2.1}$$

Soit $\mathbf{y} = (y_1, ..., y_n)$ le vecteur des données observées. En supposant que les probabilités d'appartenance à chaque sous-population G_j sont inconnues, le problème consiste à estimer les $J - 1$ proportions d'appartenance $w_1, w_2, ..., w_{J-1}$ et les paramètres $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J$ des fonctions de densité $f_1, f_2, ...f_J$, sur la base des observations du vecteur $\mathbf{y}$.

Une manière de formaliser le problème est d'introduire le concept de variables latentes, c'est-à-dire de considérer qu'il existe une donnée non observée z_i qui définit l'appartenance de l'observation à l'une ou l'autre des sous-populations G_j

du mélange. Ainsi, la i^e observation complète consiste en deux parties : la variable d'intérêt Y_i et la variable latente $\boldsymbol{Z_i}$. La variable $\boldsymbol{Z_i}$ est un vecteur de variables dichotomiques définies comme suit : $\boldsymbol{Z_i} = (Z_{i1}, ..., Z_{iJ})$ où chacun des Z_{ij} est tel que

$$Z_{ij} = \begin{cases} 1 & \text{si } y_i \text{ appartient à la population } j; \\ 0 & \text{si } y_i \text{ n'appartient pas à la population } j\,; \end{cases}$$

avec $\mathbb{P}(Z_{ij} = 1) = w_j$. Chaque Z_{ij} est donc une variable Bernoulli de paramètre w_j. Notons qu'exactement un de ces Z_{ij} vaut 1 et que les autres valent 0 pour chaque i. Nous pouvons aussi reformuler les $\boldsymbol{Z_i}$ tels que $Z_i = j$ si l'observation y_i provient de la sous-population G_j. Connaissant les proportions w_j et les fonctions de densité f_j, les données sont générées selon le schéma suivant :

$$\begin{aligned} &Z_i \sim Multinômiale(1, w_1, ..., w_J) \\ &Y_i \sim f_{z_i} \qquad \text{et} \\ &\sum_{j=1}^{J} w_j = 1. \end{aligned}$$

Tel que vu à l'équation (3.2.1), et considérant que tous les w_j sont tels que $0 \leq w_j \leq 1$, et que les fonctions de densité $f_1, f_2, ..., f_J$ ont toutes le même support, nous pouvons reformuler la densité marginale de y_i en fonction des z_i comme :

$$\begin{aligned} f_{Y_i}(y_i \mid \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_J}) &= f_{Z_i=1}(y_i \mid \boldsymbol{\theta_1})\mathbb{P}(Z_i = 1) + \ ... + f_{(Z_i=J)}(y_i \mid \boldsymbol{\theta_J})\mathbb{P}(Z_i = J) \\ &= f_1(y \mid \boldsymbol{\theta_1})\, w_1 + \cdots + f_J(y \mid \boldsymbol{\theta_J})\, w_J. \end{aligned}$$

Le nombre total de paramètres à estimer est donc $(l+1) * J - 1$, soit les $J-1$ premiers w_j plus les l paramètres caractérisant la distribution de chaque sous-population. Soit $\boldsymbol{\Psi} = (\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_J}, w_1, ...w_{J-1})$ le vecteur de paramètres à estimer.

La loi de probabilité conjointe pour les données complètes (Y, Z) est donnée par :

$$\mathbb{P}(Y, Z \mid \boldsymbol{\Psi}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[w_j * f_j(y_i \mid \boldsymbol{\theta_j})\right]^{z_{ij}}.$$

De plus, la probabilité de Z_{ij} conditionnellement à Y_i est :

$$\begin{aligned} \mathbb{P}(Z_{ij} = 1 \mid Y_i) &= \frac{w_j\, f_j(y_i \mid \boldsymbol{\theta_j})}{\sum_{j=1}^{J} w_j\, f_j(y_i \mid \boldsymbol{\theta_j})} \\ \mathbb{P}(Z_{ij} = 0 \mid Y_i) &= 1 - \frac{w_j\, f_j(y_i \mid \boldsymbol{\theta_j})}{\sum_{j'=1}^{J} w_{j'}\, f_{j'}(y_i \mid \boldsymbol{\theta_{j'}})}. \end{aligned}$$

3.3. Choix de la loi de Burr

Un des modèles de mélange les plus populaires pour les distributions statistiques continues est le modèle de mélange de lois normales. Cependant, dans le cas de l'érosion de cavitation, les données sont strictement positives, alors que la loi normale admet des valeurs négatives. Cette propriété vient souvent de pair avec une certaine asymétrie de la distribution, comportement que nous pouvons observer par exemple par l'intermédiaire de la distribution de la grappe 5, de prime abord unimodale, et présentant une asymétrie à droite (voir figure 2.6(e)). Ainsi, dans l'optique d'être le plus fidèle possible au caractère des données, l'application du modèle de mélanges de loi est orientée vers le mélange de lois de Burr de type XII.

Notons que la fonction de probabilité de la loi de Burr à trois paramètres telle qu'explicitée par Tadikamalla (1980) est donnée par :

$$f_y(y) = \frac{kc}{\alpha}\left(\frac{y}{\alpha}\right)^{c-1}\left(1+\left(\frac{y}{\alpha}\right)^{c}\right)^{-(k+1)} \text{ avec } \quad \text{y} > 0, \tag{3.3.1}$$

où $c > 0$ et $k > 0$ sont des paramètres de forme et $\alpha > 0$ est un paramètre d'échelle.

Cette dernière est particulièrement flexible et permet de refléter différents types d'asymétrie et d'aplatissement (voir figure 3.1). Originalement introduite par Burr avec deux paramètres (voir Burr, 1942), Tadikamalla ajoute un paramètre d'échelle supplémentaire (voir Tadikamalla, 1980). Par ailleurs, la distribution de Burr de type XII, parfois nommée distribution log-logistique généralisée, inclut de nombreuses distributions usuelles, telles la distribution lomax si $c = 1$ et $\alpha = 1$, et la distribution Weibull à deux paramètres si $k \to \infty$ et $\alpha \to \infty$ avec $\left(\frac{\alpha}{k}\right)^{1/c} = \kappa$, où κ est une constante (voir Shao, 2004). Notons que la loi Burr couvre une étendue plus large de formes de distributions, dont la forme des courbes normale, log-normale, logistique et Gamma (voir Bokhari et Ahmad, 2014).

Le choix de la loi de Burr est aussi motivé par l'exploration des données en terme d'ajustement statistique de lois. En effet, l'idée première est d'ajuster une distribution unimodale à chacune des grappes (voir l'annexe B). Différentes distributions sont testées, et nous retenons celle qui performe le mieux selon le critère BIC (Bayesion information criterion), c'est-à-dire que nous choisissons la distribution qui minimise le critère BIC (Schwarz *et al.*, 1978). Quoique l'ajustement ne soit pas optimal, nous pouvons voir que la loi de Burr est celle qui s'ajuste le mieux dans le cas de la grappe 4, fortement asymétrique (voir figure B.4).

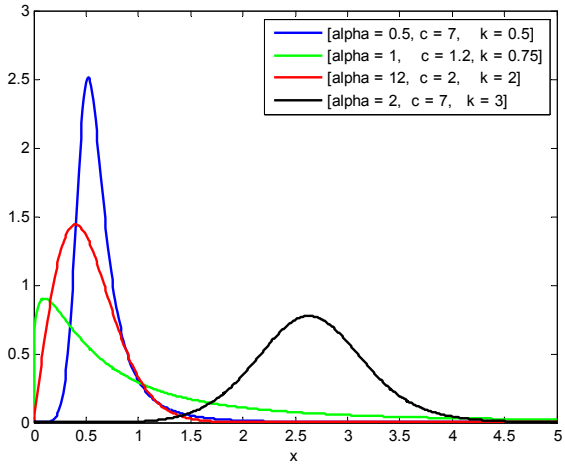


FIGURE 3.1. Exemple de lois de Burr avec différents paramètres

Par ailleurs, l'objectif étant de rendre le processus de modélisation le plus accessible possible, notons que nous nous sommes limités aux distributions disponibles dans Matlab 2013b. Les distributions disponibles dans le logiciel, mais avec un support non adéquat (qui ne pouvait contenir les données de cavitation), ont par ailleurs été exclues. Les neuf distributions ajustées sont donc : Birnbaum-Sauders, gamma, inverse gaussienne, log-logistique, Burr type XII à 3 paramètres, Nakagami, normale, Weibull et valeurs extrêmes généralisées (GEV). L'estimation des paramètres est effectuée par le maximum de vraisemblance (voir annexe B).

Le tableau 3.1 nous permet de constater que les trois lois les plus appropriées en regard de la minimisation du critère BIC sont la loi de Burr, la loi des valeurs extrêmes généralisées et la loi gamma. Quoiqu'elle performe assez bien globalement pour les cinq grappes, la loi GEV ne se classe jamais au premier rang. La loi gamma se classe au premier rang seulement pour la grappe 1. La loi de Burr, quant à elle, se classe au premier rang pour trois des cinq grappes (2, 4 et 5) et paraît être la meilleure en terme d'ajustement global pour les cinq grappes. Notons que les deux autres grappes (1 et 3) présentent quant à elles des distributions clairement plurimodales (voir figure 2.6) et il paraît normal qu'un ajustement unimodal ne soit pas optimal pour elles.

Ainsi, la nécessité d'ajuster un modèle particulièrement flexible aux données et la première exploration des données via l'ajustement unimodal de différentes lois nous incitent à utiliser un modèle de mélanges de lois de Burr à trois paramètres. L'estimation des paramètres est effectuée par le biais de l'algorithme Espérance-Maximisation et est développée dans la section suivante.

TABLEAU 3.1. BIC pour l'ajustement de 9 distributions unimodales

	Grappe 1 N = 7456	Grappe 2 N = 1738	Grappe 3 N = 2049	Grappe 4 N = 223	Grappe 5 N = 23 901
BirnbaumSaunders	34931,2	5993,4	7208,5	897,4	72474,4
Gamma	34889,5	6038,6	7173,9	913,9	72168,1
Inverse-gaussienne	34937,8	5992,6	7209,4	897,0	72483,5
Loglogistique	35417,1	5957,2	7261,1	899,4	72058,4
Burr	35169,0	5934,7	7199,6	839,5	71744,3
Nakagami	34934,7	6121,3	7162,3	932,8	72344,2
Normal	35199,6	6228,3	7175,8	955,2	72872,4
Weibull	35242,0	6474,3	7282,4	976,5	75423,9
GEV	34894,8	5967,8	7181,5	846,2	72534,4

3.4. ESTIMATION DES PARAMÈTRES ET ALGORITHME EM

Développé par Dempster *et al.* (1977), l'algorithme EM, pour Espérance-Maximisation, est une méthode itérative utilisée pour trouver l'estimateur du maximum de vraisemblance dans un modèle pour lequel les données sont incomplètes. Comme les modèles de mélange de lois présentent des défis au niveau de l'estimation des paramètres, la publication de cet article a grandement contribué à stimuler l'intérêt dans l'utilisation des modèles de mélanges pour les données hétérogènes. L'idée est de conceptualiser le problème sous l'angle des données incomplètes : les Z_{ij}, c'est-à-dire les variables déterminant l'appartenance de l'observation y_i à la sous-population G_j, sont considérées comme une variable non observée. L'estimation par le maximum de vraisemblance en est grandement simplifiée.

Tel que son nom l'indique, l'algorithme EM comporte deux étapes, qu'on itère jusqu'à convergence : l'étape Espérance (ou étape E) et l'étape Maximisation (ou étape M), où on maximise l'expression trouvée à l'étape E, ce qui permet de mettre à jour l'estimation des paramètres. Ces deux étapes sont développées ci-après. L'étape Espérance (E) calcule l'espérance du log de la fonction de vraisemblance, en tenant compte des variables observées et de la valeur des paramètres estimés à l'étape précédente r. Cette quantité se note $Q(\Psi, \Psi_r)$ et s'exprime comme suit :

$$Q(\Psi; \Psi_r) = E_{\mathbf{Z}|\mathbf{Y},\Psi_r}[\log \mathbb{P}(\mathbf{Y,Z} \mid \mathbf{\Psi})].$$

L'étape Maximisation (M) permet de trouver les estimateurs du maximum de vraisemblance correspondant à l'étape r (noté Ψ_{r+1}) en maximisant $Q(\Psi; \Psi_r)$, l'espérance de la fonction de log-vraisemblance trouvée à l'étape E, c'est-à-dire :

$$\Psi_{r+1} = \underset{\Psi}{\operatorname{argmax}}\ Q(\Psi; \Psi_r).$$

Les paramètres trouvés à l'étape M sont ensuite utilisés comme nouveau point de départ (Ψ_{r+1}) pour trouver $Q(\Psi, \Psi_{r+1})$ et son argmax mis à jour et nous répétons les étapes E et M jusqu'à convergence.

L'algorithme EM présente plusieurs propriétés attrayantes par rapport à d'autres méthodes itératives. En effet, tel que mentionné dans McLachlan et Krishnan (1997), l'algorithme EM est stable numériquement et chaque itération augmente la vraisemblance. De plus, sous certaines conditions de régularité, il converge globalement. Il est en général facilement applicable, puisqu'il repose sur le calcul de données complètes, et facile à programmer. En effet, il n'implique pas l'évaluation de la vraisemblance des données observées, mais seulement de son espérance. Il requiert peu de mise en mémoire et peut être exécuté sur un ordinateur sans grande puissance. Certaines de nos grappes comprenant un nombre élevé de données (plus de 20 000), cet avantage est non négligeable.

Une application aux données de cavitation de l'algorithme EM avec un mélange de lois normales a été développée précédemment (Bodson-Clermont, 2013). Pour plus de détails sur la procédure, le lecteur peut aussi se référer à McLachlan et Krishnan (1997). Notons que l'algorithme EM est un outil populaire pour les problèmes d'estimation statistique sur des données incomplètes ou sur des problèmes pouvant être posés en supposant une forme similaire, comme les modèles de mélange de lois. Son utilisation est largement répandue dans des champs aussi variés que la classification, l'imagerie médicale ou la génétique.

3.5. Exemple de mélange de lois de Burr

Dans la section ci-après, nous développons l'algorithme EM appliqué à un mélange de lois de Burr à plusieurs sous-populations. Tel que vu à la section précédente, la densité conjointe d'un mélange est donnée par :

$$\mathbb{P}(\mathbf{Y},\, \mathbf{Z} \mid \mathbf{\Psi}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[w_j\, f_j(y_i \mid \boldsymbol{\theta_j})\right]^{z_{ij}},$$

où

$$z_{ij} = \begin{cases} 1 & \text{si } y_i \text{ appartient à la population } j \\ 0 & \text{si } y_i \text{ n'appartient pas à la population } j. \end{cases}$$

Dans le cas d'un mélange de lois de Burr à trois paramètres, la log-vraisemblance s'exprime donc comme :

$$\log \mathbb{P}(Y, Z \mid \mathbf{\Psi}) = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \log \left[w_j\, f_j(y_i \mid \boldsymbol{\theta_j})\right] \tag{3.5.1}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \, \log w_j + \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \, \log \left[f_j(y_i \mid \boldsymbol{\theta_j}) \right],$$

où $f_j(y_i \mid \boldsymbol{\theta_j})$ est la fonction de densité de la population j, telle que donnée à l'équation (3.3.1). Pour ajuster le modèle de mélange de lois de Burr, nous devons ainsi estimer trois paramètres par sous-population ($\boldsymbol{\theta_j} = (k, c, \ \alpha)$) et $J-1$ paramètres de probabilités d'appartenance w_j. Les deux étapes de l'algorithme EM (Espérance et Maximisation) appliquées à la loi de Burr sont développées ci-après.

Pour l'étape E, nous devons calculer l'espérance conditionnelle du log de la fonction de vraisemblance des données complètes, par rapport aux données observées et aux évaluations des paramètres à l'étape précédente. Nous avons que :

$$Q(\psi; \psi_r) = E_{Z|Y,\boldsymbol{\psi_r}} \left[\log \mathbb{P}(Y, Z \mid \boldsymbol{\Psi_r}) \right]$$

où

$$Z_{ij} | \boldsymbol{\Psi_r} \sim Bernoulli(w_j^{(r)}).$$

Notons que :

$$\begin{aligned} E_{Z_{ij}|Y_i,\boldsymbol{\Psi}_r}(Z_{ij}) &= \sum_{z_{ij}} z_{ij} \ \mathbb{P}(Z_{ij} = z_{ij} \mid Y_i, \boldsymbol{\Psi}_r) \\ &= 0 * \mathbb{P}(Z_{ij} = 0 \mid Y_i, \boldsymbol{\Psi}_r) + 1 * \mathbb{P}(Z_{ij} = 1 \mid Y_i, \boldsymbol{\Psi}_r) \\ &= \mathbb{P}(Z_{ij} = 1 \mid Y_i, \boldsymbol{\Psi}_r). \end{aligned}$$

Supposons que nous sommes à l'itération r et posons

$$\mathbb{P}(Y_i = y_i \mid Z_{ij} = 1, \boldsymbol{\Psi}_r) = f_j(y_i \mid Z_{ij}, \boldsymbol{\Psi_r})$$

et

$$\mathbb{P}(Z_{ij} = 1) = w_j^{(r)}.$$

Nous cherchons donc :

$$\mathbb{P}(Z_{ij} = 1 \mid Y_i, \boldsymbol{\Psi}_r) = \frac{\mathbb{P}(Y_i = y_i \mid Z_{ij} = 1, \boldsymbol{\Psi}_r) \, \mathbb{P}(Z_{ij} = 1)}{\mathbb{P}(Y_i = y_i \mid \boldsymbol{\Psi}_r)} \tag{3.5.2}$$

$$= \frac{f_j(y_i \mid Z_{ij}, \boldsymbol{\Psi_r},) \, w_j^{(r)}}{\sum_l f_l(y_i \mid Z_{il}, \boldsymbol{\Psi_r},) w_l^{(r)}} \tag{3.5.3}$$

Nous appelons cette expression $w_{ij}^{(r)}$ les probabilités conditionnelles évaluées selon les valeurs des paramètres de l'étape r. Nous avons donc que :

$$E_{Z_{ij}|Y_i,\boldsymbol{\Psi}_r}(Z_{ij}) = w_{ij}^{(r)}.$$

Ainsi, pour l'étape E, en utilisant l'équation (3.5.1),

$$\begin{aligned}
Q(\psi;\psi_r) &= E_{Z|Y,\boldsymbol{\psi_r}}\left[\log \mathbb{P}(Y,Z \mid \boldsymbol{\Psi})\right] \\
&= E_{Z|Y,\boldsymbol{\psi_r}}\left[\sum_{i=1}^{n}\sum_{j=1}^{J} Z_{ij}\, \log w_j + \sum_{i=1}^{n}\sum_{j=1}^{J} Z_{ij}\, \log\left[f_j(y_i \mid \boldsymbol{\theta_j})\right]\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{J} E_{Z|Y,\boldsymbol{\psi_r}}(Z_{ij})\, \log w_j + \sum_{i=1}^{n}\sum_{j=1}^{J} E_{Z|Y,\boldsymbol{\psi_r}}(Z_{ij})\, \log\left[f_j(y_i \mid \boldsymbol{\theta_j})\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, \log w_j + \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, \log\left[f_j(y_i \mid \boldsymbol{\theta_j})\right].
\end{aligned}$$

Pour un mélange de lois de Burr, l'étape E s'écrit comme :

$$\begin{aligned}
Q(\psi;\psi_r) &= \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, \log w_j + \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, \log\left[\frac{k_j c_j}{\alpha_j}\left(\frac{y_i}{\alpha_j}\right)^{c_j-1}\left[1+\left(\frac{y_i}{\alpha_j}\right)^{c_j}\right]^{-(k_j+1)}\right]. \\
&= \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, \log w_j + \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, \log k_j + \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, \log c_j \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, (c_j-1)\log y_i - \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, (c_j)\log \alpha_j \\
&- \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(r)}\, (k_j+1)\log\left(1+\left(\frac{y_i}{\alpha_j}\right)^{c_j}\right).
\end{aligned}$$

L'étape de maximisation, opérée sur l'expression précédente, permet d'estimer les paramètres $\boldsymbol{\Psi_{r+1}}$, c'est-à-dire $3 * J$ paramètres définissant la distribution de Burr de chaque sous-population et $J-1$ probabilités d'appartenance :

- α_j pour $j = 1, ..., J$
- c_j pour $j = 1, ..., J$
- k_j pour $j = 1, ..., J$
- w_j pour $j = 1, ..., J-1$.

Dans le contexte des modèles de mélanges de lois de dimension finie, la convergence de l'algorithme EM peut parfois être très lente. Des extensions de l'algorithme EM ont été développées pour contourner ce problème. En particulier, Celeux et Diebolt (1986) ont considéré une version modifiée de l'algorithme EM appelée l'algorithme EM stochastique, développé dans la section suivante.

3.6. Extensions de l'algorithme EM

3.6.1. EM stochastique

Dans l'étape E de l'algorithme EM, étant donné la linéarité de la log-vraisemblance complète, chaque z_{ij} est remplacé par les valeurs des probabilité conditionnelles de l'étape r, c'est-à-dire par $w_{ij}^{(r)}$ (voir l'équation 3.6.1). Dans le cadre de l'algorithme EM stochastique, l'étape E consiste plutôt à générer les $Z_i^{(r)}$ (où $Z_i^{(r)}$ peut prendre des valeurs de 1 à J) à partir de ces mêmes probabilités conditionnelles $w_{ij}^{(r)}$. Ainsi, pour chaque y_i, on génère un z_i de la distribution multinomiale avec des paramètres $(1, w_{i1}^{(r)}, w_{i2}^{(r)}, ..., w_{iJ}^{(r)})$. On obtient ainsi une partition des n observations $y_1, y_2, ..., y_n$ selon les J sous-populations du modèle de mélange. Après l'étape maximisation, les probabilités conditionnelles sont mises à jour comme pour l'algorithme EM selon l'équation :

$$w_{ij}^{(r+1)} = \frac{f_j(y_i \mid Z_{ij}, \boldsymbol{\Psi_{r+1}},)\, w_j^{(r+1)}}{\sum_l f_l(y_i \mid Z_{il}, \boldsymbol{\Psi_{r+1}},) w_l^{(r+1)}}. \tag{3.6.1}$$

Cet algorithme permet de prévenir la situation où la séquence reste près d'un point stationnaire instable de la fonction de vraisemblance (voir McLachlan et Krishnan, 1997, chap. 6). Nous évitons ainsi les cas de convergence lente observée dans certains cas de l'application de l'algorithme EM aux modèles de mélanges. Notons par ailleurs que cet algorithme est un cas particulier de l'algorithme EM avec méthodes de Monte-Carlo (MCEM) qui consiste, pour chaque itération r, à générer M fois les $Z_i^{(r)}$ et à approximer l'espérance comme suit :

$$Q(\psi; \psi_r) = \frac{1}{M} \sum_{m=1}^{M} \log \mathbb{P}(\boldsymbol{\Psi} \mid Y, Z_{(m)}^{(r)}). \tag{3.6.2}$$

Pour l'algorithme EM avec méthodes Monte-Carlo (et donc pour l'algorithme EM stochastique), la propriété de monotonicité est perdue. Cependant, si les valeurs initiales sont choisies judicieusement, l'algorithme s'approche d'un maximum avec une probabilité élevée (voir Chan et Ledolter, 1995, pour plus de détails).

3.6.2. Étape M : utilisation de l'extension ECME

Une autre extension de l'algorithme EM a été proposée par Meng et Rubin (1993), dans le cas où l'étape de maximisation à l'étape M est relativement simple

lorsqu'on conditionne la fonction de vraisemblance par rapport à certains paramètres. L'étape M est ainsi remplacée par une suite d'étapes de maximisation conditionnelle, où chacun des paramètres est maximisé de façon unidimensionnelle alors que les autres paramètres sont considérés fixes. Conséquemment, la convergence est typiquement plus lente en terme de nombre d'itérations, mais peut être plus rapide en temps de calcul total. Une extension supplémentaire de cet algorithme, développée par Liu et Rubin (1994), permet de gagner en vitesse presque toujours en termes de nombre d'itérations et de temps de calcul total et se nomme l'algorithme ECME. En effet, en maximisant conditionnellement à la fonction de log-vraisemblance $\log \mathbb{P}(\mathbf{\Psi} \mid Y, Z)$ plutôt qu'à l'approximation donnée par la fonction Q proposée par l'algorithme EM et ECM, la vitesse de convergence est grandement améliorée (voir McLachlan et Krishnan, 1997).

Comme les algorithmes EM et ECM, l'ECME converge habituellement vers un maximum local de la fonction de vraisemblance, tout en simplifiant les étapes computationnelles et en réduisant le temps de calcul. C'est cette dernière extension qui est développée dans le contexte de ce mémoire, en maximisant chacun des paramètres conditionnellement à la fonction de vraisemblance et en fixant les autres paramètres à leur dernière valeur courante. L'application aux données de cavitation permet de vérifier que la méthode fonctionne bien et que les estimations avec cette méthode sont raisonnables (voir section 3.8).

Après avoir générer les $Z_i^{(r)}$ (où $Z_i^{(r)}$ peut prendre des valeurs de 1 à J) à partir des probabilités conditionnelles $w_{ij}^{(r)}$ (voir équation (3.6.1)), l'idée est d'extraire tous les y_i qui appartiennent à la même sous-population, c'est-à-dire ayant les mêmes valeurs de $Z_i^{(r)}$. L'exemple ci-après est construit pour la première sous-population G_1, soit pour les y_i tels que $Z_i = 1$. L'estimation des paramètres pour cette sous-population se fait ensuite sur ce sous-échantillon. Soit n_1 le nombre de y_i tels que $Z_i = 1$. Nous cherchons ainsi le maximum de vraisemblance pour w, α, c et k. L'estimation pour les autres sous-populations $1, ..., J$ se fait de la même manière.

3.6.2.1. *Estimation pour le paramètre w*

L'estimation de la probabilité d'appartenance w_j pour la sous-population G_j consiste à maximiser la log-vraisemblance par rapport au paramètre w_j. À l'étape r, l'étape E est donc donnée pare :

$$Q(\psi; \psi_r) = \log \mathbb{P}(\mathbf{\Psi} \mid Y, Z^{(r)})$$

qui est donnée par l'équation (3.5.1). En tenant compte de la contrainte $\sum_{j=1}^{J} w_j = 1$, le lagrangien s'écrit :

$$\mathcal{L}(w_1, ..., w_J, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} \log w_j + \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} \log\left[f_j(y_i \mid \boldsymbol{\theta_j})\right] - \lambda\,(1 - \sum_{j=1}^{J} w_j).$$

En mettant la dérivée par rapport à w_j égale à 0, nous obtenons :

$$\frac{\partial}{\partial w_j}\mathcal{L}(w_1, ..., w_J, \lambda) = \sum_{i=1}^{n} \frac{z_{ij}}{w_j} + \lambda = 0$$

et donc

$$w_j = -\sum_{i=1}^{n} \frac{z_{ij}}{\lambda}. \tag{3.6.3}$$

En remplaçant w_j par l'expression donnée en (3.6.3), la dérivée par rapport à λ égale à 0 donne :

$$-\frac{\partial}{\partial \lambda}\mathcal{L}(w_1, ..., w_J, \lambda) = 1 - \sum_{j=1}^{J} w_j = 1 - \sum_{j=1}^{J}\left(-\sum_{i=1}^{n} \frac{z_{ij}}{\lambda}\right) = 0,$$

ce qui implique que $\lambda = -\sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij}$.

Selon (3.6.3), l'estimation de w_j est ainsi donnée par :

$$\widehat{w}_j = \frac{-\sum_{i=1}^{n} z_{ij}}{-\sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij}} = \frac{\sum_{i=1}^{n} z_{ij}}{n}.$$

Pour la sous-population G_1, $\widehat{w}_1$ correspond donc au nombre de données appartenant à la sous-population G_1 divisé par le nombre total de données.

3.6.2.2. *Maximisation pour le paramètre k*

L'étape suivante consiste à maximiser par rapport au paramètre k lorsque c et α sont connus. Avec le changement de variable $t = \left(\frac{y}{\alpha}\right)^c$, la fonction de densité (3.3.1) s'écrit :

$$f_t(t \mid k) = k(1 + t)^{-(k+1)}.$$

La fonction de vraisemblance est donc :

$$\mathcal{L}(k) = \prod_{i=1}^{n_1} k\,(1 + t_i)^{-(k+1)}.$$

$$\Longrightarrow \log \mathcal{L}(k) = n_1 \log k - (k + 1)\sum_{i=1}^{n_1} \log(1 + t_i)$$

$$\Longrightarrow \frac{\partial}{\partial k}\log \mathcal{L}(k) = \frac{n_1}{k} - \sum_{i=1}^{n_1}\log(1+t_i) = 0$$
$$\Longrightarrow \widehat{k}_{MV} = \frac{n_1}{\sum_{i=1}^{n_1}\log(1+t_i)}.$$

3.6.2.3. *Maximisation pour le paramètre α*

En se référant à l'équation (3.3.1), le changement de variable $u = y^c$ donne la fonction de densité suivante :

$$f_u(u \mid \alpha, c) = \frac{k}{\alpha^c}\left(1 + \frac{u}{\alpha^c}\right)^{-(k+1)}.$$

La fonction de vraisemblance s'exprime comme :

$$\mathcal{L}(k,\alpha) = \prod_{i=1}^{n_1}\frac{k}{\alpha^c}\left(1+\frac{u_i}{\alpha^c}\right)^{-(k+1)}$$
$$\Longrightarrow \log \mathcal{L}(k,\alpha) = n_1\,\log k - cn_1 \log\alpha - (k+1)\sum_{i=1}^{n_1}\log\left(1+\frac{u_i}{\alpha^c}\right)$$
$$\Longrightarrow \log \mathcal{L}(k,\alpha) = n_1\,\log k + n_1 ck \log\alpha - (k+1)\sum_{i=1}^{n_1}\log(\alpha^c + u_i)$$
$$\Longrightarrow \frac{\partial}{\partial\alpha}\log \mathcal{L}(k,\alpha) = \frac{n_1 ck}{\alpha} - (k+1)c\alpha^{c-1}\sum_{i=1}^{n_1}\frac{1}{\alpha^c + u_i} = 0.$$

Nous devons trouver les racines de α pour l'équation :

$$g(\alpha) = \frac{n_1}{\alpha^c} - \frac{k+1}{k}\sum_{i=1}^{n_1}\frac{1}{\alpha^c + u_i} = 0. \tag{3.6.4}$$

Pour l'estimation du paramètre α, comme pour l'estimation du paramètre c développée ci-après, la résolution de l'équation est implémentée à l'aide de l'algorithme de Newton-Raphson. Supposons que α_0 est la valeur initiale et α la valeur de la racine de l'équation (3.6.4). En vérifiant que α_0 est assez proche de α, avec $g'(\alpha_0) \neq 0$, l'existence d'une solution est assurée, puisque la fonction g est deux fois dérivable et avec dérivée seconde continue. L'initialisation est par ailleurs importante pour assurer la convergence (voir Burden et Faires, 2005).

3.6.2.4. *Maximisation pour le paramètre c*

Toujours en se référant à l'équation (3.3.1), nous posons le changement de variable $v = \frac{y}{\alpha}$ et la fonction de densité est donnée par :

$$f_v(v \mid k, c) = k\,c\,v^{c-1}(1+v^c)^{-(k+1)}.$$

La fonction de vraisemblance s'exprime comme :

$$\mathcal{L}(k,c) = \prod_{i=1}^{n_1} k\, c\, v_i^{c-1}(1+v_i^c)^{-(k+1)}$$

$$\implies \log \mathcal{L}(k,c) = n_1 \log k + n_1 \log c + (c-1)\sum_{i=1}^{n_1} \log v_i - (k+1)\sum_{i=1}^{n_1} \log(1+v_i^c)$$

$$\implies \frac{\partial}{\partial c} \log \mathcal{L}(k,c) = \frac{n_1}{c} - k\sum_{i=1}^{n_1} \log v_i + (k+1)\sum_{i=1}^{n_1} \frac{\log v_i}{1+v_i^c} = 0.$$

Nous devons trouver les racines de c pour cette dernière équation. Nous utilisons aussi l'algorithme de Newton-Raphson pour estimer c.

3.6.2.5. *Identifiabilité*

Dans l'objectif d'accélérer l'algorithme et pour pallier aux problèmes d'identifiabilité, le mode est calculé à chaque itération pour chaque sous-population avec l'estimation des paramètres mis à jour. Les paramètres sont ensuite réordonnés selon le mode. Le mode d'une loi de Burr à trois paramètres est donné par :

$$\left[\frac{(c-1)\alpha^c}{kc+1}\right]^{\frac{1}{c}}.$$

Les étapes E et M sont ensuite répétées avec les paramètres mis à jour ordonnés.

3.7. Résumé des étapes

En résumé, l'algorithme EM avec méthode de Monte-Carlo est implémenté pour la loi de Burr de la façon suivante :

- Initialisation des paramètres w, c, α, k pour chaque sous-population, avec $\sum_{j=1}^{J} w_j = 1$.
- Pour chaque itération r ($r = 0$ pour les paramètres initiaux),
  - Faire l'étape Espérance :
    - Calculer la log-vraisemblance du mélange avec les paramètres trouvés à l'étape précédente ($\boldsymbol{\Psi_r}$).
    - Calculer les probabilités conditionnelles pour chaque y_i ($w_{ij}^{(r)}$).
    - Générer un z_i pour chaque Y_i selon une multinomiale$(1, w_{i1}^{(r)}, ..., w_{iJ}^{(r)})$.
  - Faire l'étape Maximisation :
    - Isoler chaque sous-population selon les z_i générés à l'étape E

  - Pour chaque sous-population G_j, estimer la probabilité d'appartenance $w_j^{(r)}$, et les 3 paramètres définissant la loi de Burr qui seront utilisés à l'étape suivante : $k^{(r+1)}, c_j^{(r+1)}, \alpha_j^{(r+1)}$, en utilisant l'algorithme de Newton-Raphson pour $c_j^{(r+1)}, \alpha_j^{(r+1)}$.
- Étape Identifiabilité :
  - Pour chaque sous-population, calculer le mode de la distribution
  - Trier les paramètres selon le mode correspondant pour pallier aux problèmes d'identifiabilité.

• Reprendre les étapes E et M avec les paramètres mis à jour

L'algorithme s'arrête selon un certain critère de convergence. Tel que mentionné dans McLachlan et Krishnan (1997, chap.6), l'algorithme avec méthodes de Monte Carlo peut poser problème en terme de convergence. En effet, la propriété de monotonicité de l'algorithme est perdue. Le choix des valeurs initiales est d'ailleurs primordial et il peut être judicieux d'appliquer l'algorithme EM en faisant varier le choix des valeurs initiales pour augmenter les chances de converger vers un maximum. Le critère de convergence choisi ici dans le cas du mélange de lois de Burr s'appuie sur la log-vraisemblance, étant donné le grand nombre de paramètres et donc la possible instabilité de ceux-ci. Un graphique de chaque étape est tracé en fonction de la log-vraisemblance et lorsque celle-ci est stable pendant 50 itérations, l'algorithme s'arrête. Les valeurs des paramètres retenues sont celles de la dernière itération. Pour vérifier la stabilité de la log-vraisemblance, à chaque étape, une régression est effectuée entre les 50 valeurs précédentes de log-vraisemblance et les itérations. Si la valeur-p de la pente est non significative (>5 %) pour 50 étapes d'affilée, l'algorithme s'arrête.

Les paramètres finaux sont alors utilisés pour estimer la moyenne de la cavitation dans chacune des grappes. Cette moyenne servira de première estimation pour prédire la cavitation. Les valeurs de ces moyennes par grappe sont données dans la section 3.8.

3.8. Résultats des mélanges appliqués aux cinq grappes

3.8.1. Exemple d'application de l'algorithme EM pour la grappe 1

Nous présentons ici un exemple d'application de l'algorithme EM avec deux sous-populations pour la grappe 1. Les valeurs initiales sont données au tableau 3.2.

La figure 3.2 montre l'évolution de la log-vraisemblance pour chaque étape de l'algorithme EM. Visuellement, nous pouvons voir que la log-vraisemblance se

TABLEAU 3.2. Valeurs des paramètres pour la grappe 1 avant et après EM pour un modèle de mélanges de lois de Burr à deux sous-populations

	Ss-pop.	w	α	c	k
Paramètres	1	0,5	6	12	2
initiaux	2	0,5	9	3	2
Paramètres	1	0,331	6,020	12,769	0,444
après EM	2	0,669	14,725	5,530	5,460

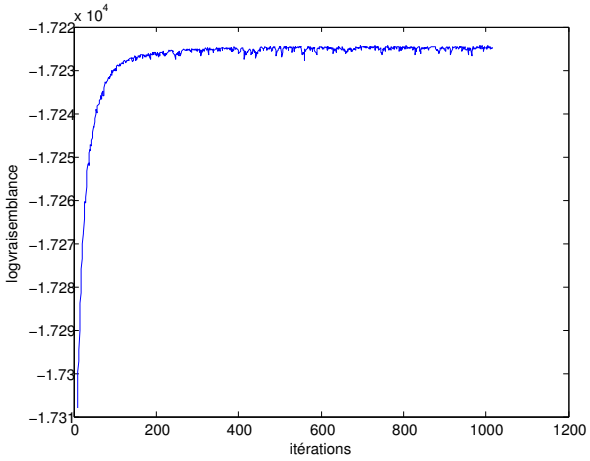


FIGURE 3.2. Évolution de la log-vraisemblance à chaque étape de l'algorithme EM (Grappe 1, mélanges à 2 sous-populations)

stabilise assez rapidement. Selon notre critère basé sur la log-vraisemblance, la convergence est obtenue à la 1018^e itération.

Concernant l'évolution des paramètres dans le cas de la grappe 1, le critère de convergence permet de stabiliser l'ensemble de ceux-ci, quoique certains soient plus volatiles. Ceci est le reflet de la difficulté à obtenir des paramètres stables pour un mélange de lois avec trois paramètres, le nombre de paramètres étant élevé, d'où l'importance de tester différents paramètres initiaux et de les choisir judicieusement. En effet, pour une log-vraisemblance et un ajustement graphique similaires, les valeurs de paramètres de la loi Burr peuvent être assez différents. Comme l'intérêt est ici de bien ajuster la loi et d'estimer la moyenne par grappe, la convergence ci-après convient aux besoins de l'étude et les valeurs finales des paramètres sont présentées au tableau 3.2. La figure 3.3 montre l'ajustement pour un mélange à deux sous-populations pour la grappe 1, avec les valeurs finales des paramètres telles que présentées dans le tableau 3.2. Nous pouvons voir que dans

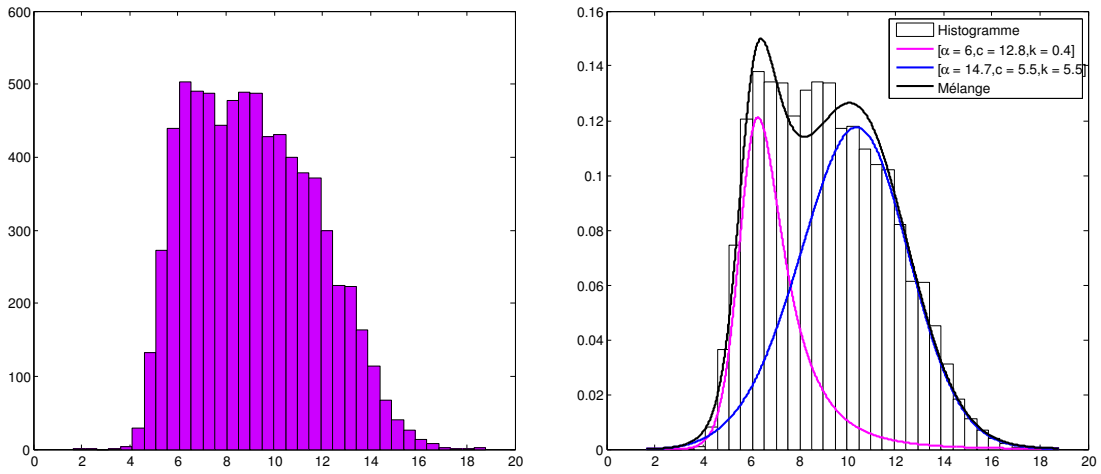


FIGURE 3.3. Ajustement du modèle de mélange à 2 sous-populations pour la grappe 1

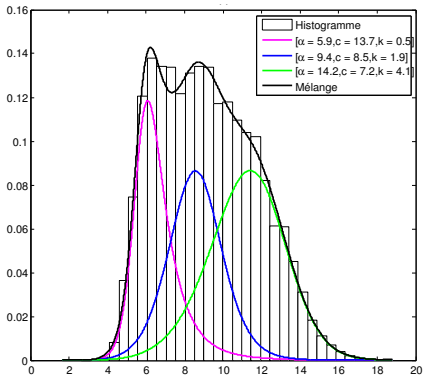
ce cas, l'ajustement à deux sous-populations n'est pas optimal. Nous verrons plus loin que pour la grappe 1, le meilleur ajustement a en effet trois sous-populations.

3.8.2. Comparaison des modèles

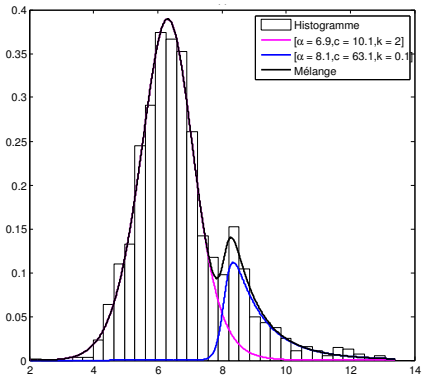
Chacune des grappes a été testée avec une, deux ou trois sous-populations et le meilleur ajustement a été choisi à l'aide du critère BIC. Le tableau 3.3 présente les résultats selon ce critère. Nous choisissons le modèle qui minimise le BIC (en rouge dans le tableau). Le meilleur modèle pour la grappe 1 a trois sous-populations, alors que pour les grappes 2, 4 et 5, le meilleur modèle a deux sous-populations. La grappe 4, quant à elle, performe mieux avec le modèle à une sous-population. Nous pouvons voir à la figure 3.4 la représentation graphique des meilleurs ajustements pour les cinq grappes. Nous voyons que pour l'ensemble des grappes, le meilleur modèle choisi s'ajuste bien à la distribution de chacune des grappes.

TABLEAU 3.3. BIC pour chaque grappe pour un modèle de lois de Burr avec 1, 2 et 3 sous-populations

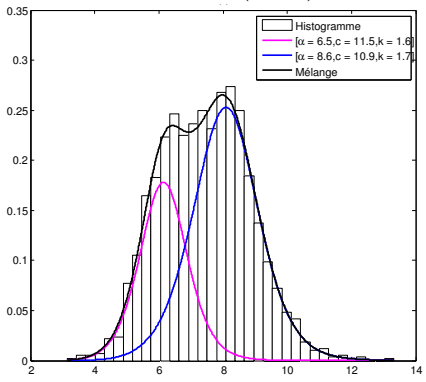
Grappe	$J = 1$	$J = 2$	$J = 3$
1	35169,03	34511,95	34469,82
2	5934,71	5919,41	pas de convergence
3	7199,58	7165,80	7196,27
4	839,49	845,19	pas de convergence
5	71744,25	71717,82	71759,13



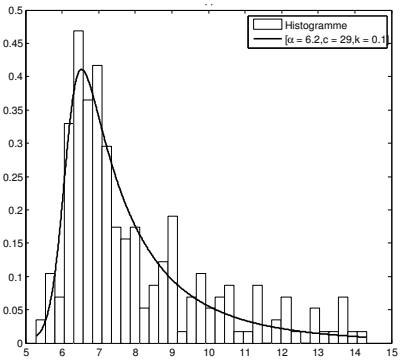
Grappe 1 (J=3)



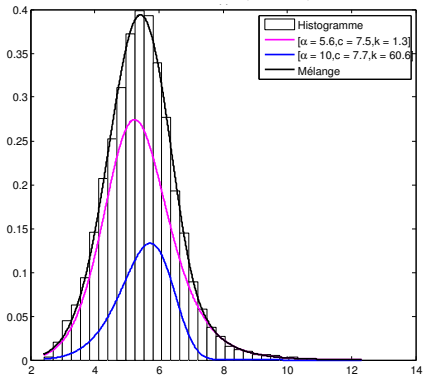
Grappe 2 (J=2)



Grappe 3 (J=2)



Grappe 4 (J=1)



Grappe 5 (J=2)

FIGURE 3.4. Meilleurs ajustements pour les cinq grappes (plus petit BIC)

Comme nous cherchons à estimer la valeur de cavitation par grappe en utilisant la moyenne, nous comparons la moyenne théorique et la moyenne empirique. L'espérance d'un modèle de mélange est donnée par :

$$E(Y) = \sum_{j=1}^{J} w_j * E_j(Y), \tag{3.8.1}$$

avec

$$E_j(Y) = k_j\,\alpha_j \mathrm{B}\left[k_j - \frac{1}{c_j}; \frac{1}{c_j} + 1\right],$$

où B est la fonction bêta donnée par :

$$\mathrm{B(a,b)} = \frac{\Gamma(\mathrm{a})\Gamma(\mathrm{b})}{\Gamma(\mathrm{a}+\mathrm{b})}.$$

Le calcul de la variance est donné par :

$$V(Y) = E(Y^2) - E(Y)^2 = \sum_{j=1}^{J} w_j E_j(Y^2) - E(Y)^2,$$

où

$$E_j(Y^2) = k_j\,\alpha_j^2 \mathrm{B}\left[[k_j - \frac{2}{c_j}; \frac{2}{c_j} + 1\right].$$

Nous obtenons la moyenne théorique en remplaçant les estimations des paramètres dans l'équation (3.8.1). Les résultats sont donnés au tableau 3.4 et nous pouvons voir que les modèles trouvés permettent de bien évaluer la moyenne. En effet, si nous comparons la moyenne théorique et la moyenne empirique, la différence la plus grande se trouve à la grappe 4 ($+0,073$ kg/10 000 h par rapport à la moyenne empirique). Pour les autres grappes, l'écart entre la moyenne empirique et théorique est plus petite que 0,002 kg/10 000 h. Par ailleurs, la différence entre les écarts-types théoriques et empiriques est plus petite que 0,02 kg/10 000 h en valeur absolue, sauf pour la grappe 4. L'écart-type théorique de celle-ci présente une différence de +0,735 kg/10 000 h par rapport à l'écart-type empirique. Nous concluons que l'estimation par la moyenne pour les modèles trouvés avec l'algorithme EM est adéquate, quoique le modèle semble avoir une moins bonne précision sur la grappe 4, ce qui n'est pas surprenant. La grappe 4 contient en effet beaucoup moins de valeurs et regroupe des modes opératoires peu utilisés qui, de prime abord, ont un comportement disparate. De plus, la modélisation avec une seule sous-population ne permet pas de capturer l'hétérogénéité des données aussi bien qu'un autre modèle. Le chapitre 4 s'attardera à comparer ces estimations par la moyenne avec des modèles de régression et les validations sur l'année 2012 seront étudiées au chapitre 5.

À titre informatif, les paramètres des mélanges à 1, 2 ou 3 sous-populations après EM sont présentés, pour chaque grappe, dans les tableaux 3.5 à 3.9.

En définitive, nous avons développé dans ce chapitre un exemple de l'application des modèles de mélange de lois aux données de cavitation. Plus spécifiquement, le choix d'un modèle de loi de Burr a été guidé par la forme asymétrique

TABLEAU 3.4. Comparaison des moyennes et écart-types théoriques et empiriques pour le meilleur ajustement de chaque grappe

Grappe	J	Moyenne théorique	Moyenne empirique	Écart-type théorique	Écart-type empirique
1	3	9,159	9,161	2,563	2,561
2	2	6,773	6,773	1,465	1,446
3	2	7,437	7,438	1,387	1,389
4	1	8,074	8,001	2,751	2,015
5	2	5,454	5,454	1,121	1,111

TABLEAU 3.5. Grappe 1 : paramètres après convergence de l'algorithme EM pour un mélange de lois de Burr

	Ss-pop.	w	α	c	k	Mode	LogVrais.	BIC
$J = 1$	1	1,000	13,263	4,451	4,046	9,040	-17571,142	35169,034
$J = 2$	1	0,331	6,020	12,769	0,444	6,294	-17224,768	34511,953
	2	0,669	14,725	5,530	5,460	10,387		
$J = 3$	1	0,273	5,888	13,676	0,511	6,091	-17185,869	34469,822
	2	0,302	9,446	8,461	1,896	8,568		
	3	0,425	14,239	7,181	4,068	11,417		

TABLEAU 3.6. Grappe 2 : paramètres après convergence de l'algorithme EM pour un mélange de lois de Burr

	Ss-pop.	w	α	c	k	Mode	LogVrais.	BIC
$J = 1$	1	1,000	6,124	10,795	0,631	6,254	-2956,166	5934,713
$J = 2$	1	0,832	6,861	10,070	1,985	6,313	-2933,595	5919,413
	2	0,168	8,104	63,110	0,132	8,351		
$J = 3$	Pas de convergence							

TABLEAU 3.7. Grappe 3 : paramètres après convergence de l'algorithme EM pour un mélange de lois de Burr

	Ss-pop.	w	α	c	k	Mode	LogVrais.	BIC
$J = 1$	1	1,000	8,803	7,238	2,671	7,477	-3588,354	7199,584
$J = 2$	1	0,340	6,467	11,531	1,556	6,146	-3556,214	7165,804
	2	0,660	8,622	10,861	1,714	8,093		
$J = 3$	1	0,249	6,661	11,879	2,503	6,104	-3556,197	7196,271
	2	0,272	8,387	8,084	1,816	7,601		
	3	0,479	8,651	10,962	1,869	8,065		

des données et pour sa grande flexibilité. Pour l'estimation des paramètres, l'algorithme EM a été utilisé. Pour chacune des grappes, un modèle à une, deux ou trois sous-populations a été ajusté et le meilleur modèle a été sélectionné selon le

TABLEAU 3.8. Grappe 4 : paramètres après convergence de l'algorithme EM pour un mélange de lois de Burr

	Ss-pop.	w	α	c	k	Mode	LogVrais.	BIC
$J = 1$	1	1,000	6,163	28,972	0,143	6,534	-411,634	839,490
$J = 2$	1	0,767	6,287	27,390	0,263	6,560	-403,668	845,185
	2	0,233	19,643	6,632	38,465	11,047		
$J = 3$				Pas de convergence				

TABLEAU 3.9. Grappe 5 : paramètres après convergence de l'algorithme EM pour un mélange de lois de Burr

	Ss-pop.	w	α	c	k	Mode	LogVrais.	BIC
$J = 1$	1	1,000	5,844	7,673	1,583	5,350	-35857,003	71744,251
$J = 2$	1	0,725	5,619	7,462	1,317	5,243	-35823,622	71717,816
	2	0,275	9,960	7,691	60,634	5,734		
$J = 3$	1	0,671	5,592	7,512	1,272	5,244	-35824,115	71759,128
	2	0,229	10,990	7,383	125,038	5,602		
	3	0,100	8,469	7,628	13,622	5,897		

critère BIC. Somme toute, cette méthode permet de bien ajuster la distribution pour chacune des grappes.

Dans l'objectif de prédire la cavitation, le premier modèle d'estimation envisagé dans ce mémoire est ainsi basé sur la caractérisation de la distribution de la cavitation par grappe à l'aide des modèles de mélanges de lois. La moyenne théorique obtenue, une fois les paramètres de lois de Burr estimés et l'adéquation du modèle vérifiée, sert ainsi de première estimation de la cavitation par grappe. L'idée étant qu'une fois le mode opératoire identifié, nous pouvons rechercher son appartenance à la grappe correspondante et obtenir un estimé de la cavitation tel que donné par le tableau 3.4.

Une alternative plus précise serait d'identifier la sous-population associée à chaque donnée et d'estimer la cavitation par la moyenne de la sous-population correspondante. Cependant, un des buts éventuels était d'implanter le modèle statistique sur une plateforme d'Hydro-Québec où l'érosion de cavitation instantanée serait prédite en temps réel selon les paramètres d'opération de la turbine. Dans cette optique, il est intéressant de développer un outil qui permet la caractérisation simple et conviviale de la situation opératoire. Dans le contexte d'opération de la centrale, l'identification du mode opératoire est directe, et l'appartenance de celui-ci à une des grappes aussi. Les 9 variables opératoires utilisées dans le modèle de régression au chapitre 4 sont aussi facilement accessibles. C'est pourquoi nous avons privilégié pour la suite du mémoire un modèle plus simple à

mettre en place et qui permettait une bonne estimation tout en conservant des propriétés conviviales d'implémentation. Le chapitre suivant s'attardera à affiner cette estimation en utilisant des modèles de régression par grappe.

Chapitre 4

MODÈLES DE RÉGRESSION

Au chapitre précédent, nous avons développé un outil afin d'ajuster des modèles de mélange de lois de Burr, très flexibles, et qui permettaient de capturer la spécificité de la distribution de la cavitation dans chaque grappe. La première étape consistait donc à utiliser la moyenne théorique de ces distributions pour estimer la cavitation dans chacune des grappes. Dans ce chapitre, nous souhaitons affiner l'estimation de la cavitation en utilisant des modèles de régression à l'intérieur de chacune des grappes qui, nous le rappelons, regroupent les modes opératoires qui présentent des similarités du point de vue des dix variables explicatives. L'hypothèse est que la cavitation dans chacune des grappes n'est pas nécessairement régie par les mêmes conditions opératoires, et donc que les modèles de régression, et particulièrement les variables explicatives, pourraient varier d'une grappe à l'autre.

4.1. Méthodologie

Sur les dix variables explicatives telles que décrites à la section 1.2.1, seules neuf variables sont retenues pour tester les modèles de régression linéaire multiple. En effet, la variable *hauteur de chute* est retirée, puisqu'elle est linéairement dépendante des autres variables (*hauteur de chute* = *amont* - *aval*). Pour chacune des grappes, quatre modèles de régression sont considérés. La variable réponse *cavitation* est utilisée avec et sans transformation log et les variables explicatives sont intégrées au modèle en considérant seulement les effets principaux (ci-après nommé modèle additif) ou en incluant les interactions de premier ordre (ci-après nommé modèle avec interactions). La sélection de variables est réalisée en utilisant la méthode descendante et les variables sont conservées dans le modèle si la valeur-p est plus petite que 5 %. Notons que pour le modèle avec interactions, les effets principaux sont conservés même s'ils ne sont pas statistiquement significatifs, dès lors qu'ils apparaissent dans l'une des interactions retenues. La validation et le

choix du meilleur modèle pour chaque grappe seront présentés au chapitre suivant (chapitre 5).

4.1.1. Transformation

Pour choisir quelles transformations sont retenues dans la méthodologie, nous analysons six transformations Box-Cox (Box et Cox, 1964) de la variable cavitation (y), en particulier au niveau de la forme des résidus. Ces transformations sont obtenues en utilisant six valeurs de λ (-2, -1, 0, 0,5, 1, 2) dans l'équation de la transformation Box-Cox, donnée par l'équation (4.1.1).

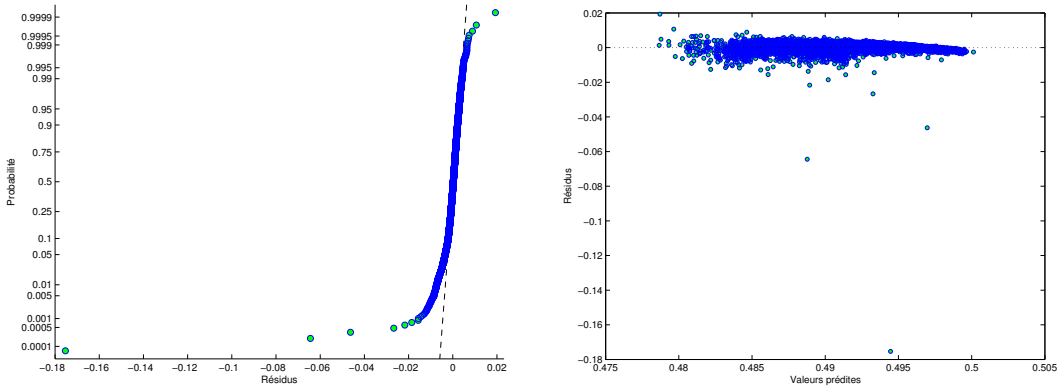
$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(y_i) & \text{si } \lambda = 0. \end{cases} \tag{4.1.1}$$

Pour chacune des grappes, la méthode descendante est appliquée sur le modèle additif et permet de sélectionner les variables significatives. Différents critères sont ensuite étudiés, particulièrement en ce qui a trait aux résidus : la répartition aléatoire autour de zéro, l'hétéroscédasticité et l'apparence de normalité des résidus. Par ailleurs, le R^2 ajusté (R_a^2) est utilisé pour donner une mesure de l'ajustement du modèle en terme de variance expliquée. Trois transformations possibles sont retenues : la transformation log, la transformation racine et le modèle non transformé (NT), puisque pour les cinq grappes, l'une ou l'autre de ces transformations maximise le R^2 ajusté (voir tableau 4.1) et améliore l'allure des résidus. Notons cependant que la transformation log est généralement supérieure à la transformation racine et que de plus, le cas racine n'est jamais seul lorsqu'il est optimal. Ainsi, nous retenons deux cas pour la méthodologie finale : le modèle non transformé et la transformation log.

TABLEAU 4.1. R^2 ajusté pour le modèle additif avec transformation Box-Cox sur la variable cavitation (y), après sélection des variables

	Valeur de λ					
Grappe	-2	-1	0	0,5	1	2
1	0,635	0,802	0,832	0,823	0,802	0,738
2	0,485	0,638	0,713	0,732	0,742	0,733
3	0,631	0,687	0,711	0,711	0,705	0,673
4	0,721	0,763	0,787	0,792	0,792	0,777
5	0,398	0,418	0,419	0,413	0,403	0,375

Un exemple pour la grappe 1, qui présente notamment une courbure dans le graphique des résidus contre les valeurs prédites pour le modèle non transformé,



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

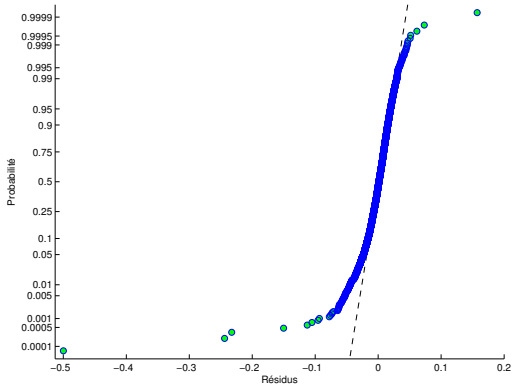
FIGURE 4.1. Analyse des résidus pour la régression linéaire de la grappe 1 avec transformation Box-Cox $\lambda = -2$

est développé pour les six transformations. Les graphiques des résidus sont présentés aux figures 4.1 à 4.6.

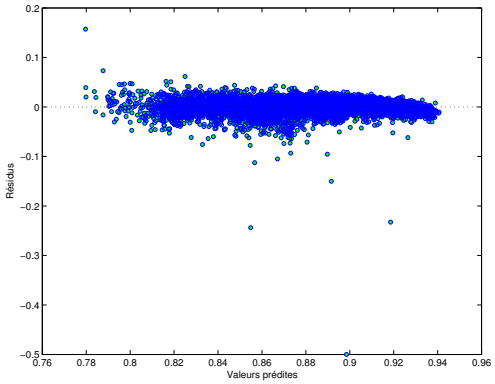
Nous pouvons ainsi voir aux figures 4.1(b) à 4.6(b) que les résidus présentent une courbure à partir de la transformation Box-Cox avec $\lambda = 0{,}5$ qui s'accentue si nous augmentons la valeur de λ. Ces figures mettent aussi en relief une légère augmentation de l'hétéroscédasticité. D'autre part, lorsque nous observons les graphiques de probabilités aux figures 4.1(a) et 4.6(a), nous pouvons voir que les résidus présentent une distribution aux queues plus concentrées que la distribution normale. Par ailleurs, pour les transformations $\lambda = -2$ ou -1 , le R^2 ajusté est moins élevé que celui de la transformation log ($R^2_a = 0,635$ pour $\lambda = -2$ et $R^2_a = 0,802$ pour $\lambda = -1$ en comparaison avec $R^2_a = 0,832$ pour log). Nous voyons donc que pour la grappe 1, la transformation log semble être la meilleure du point de vue du comportement des résidus et du R^2 ajusté. Les autres grappes présentant un comportement assez similaire, nous avons retenu le modèle log-normal additif (LinLog) et avec interactions (InterLog) pour la modélisation, en plus des modèles non transformés (LinNT et InterNT). Rappelons que la loi de Burr couvre les caractéristiques la forme de courbe de la loi log-normale.

4.2. Modèles de régression par grappe

Dans la section suivante, nous présentons les quatre modèles étudiés pour la grappe 1, avec leurs caractéristiques et l'analyse des résidus. Les modèles pour les autres grappes sont disponibles à l'annexe C. Nous rappelons que les quatre modèles prennent la forme suivante :

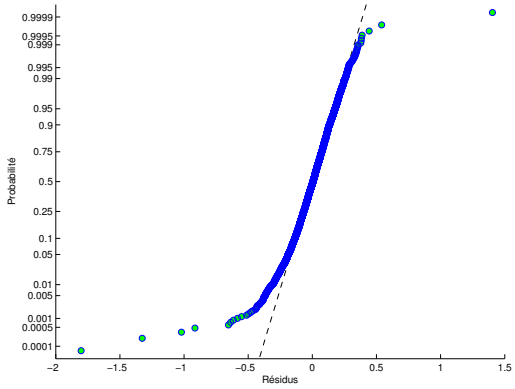


(a) Graphique de probabilités normales

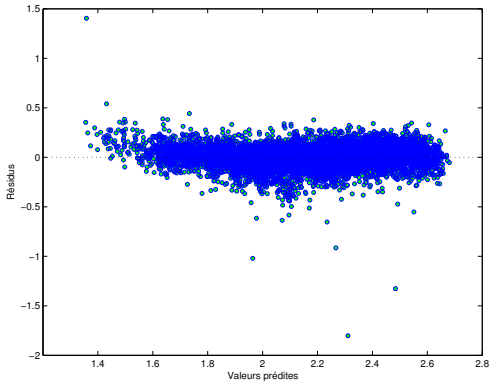


(b) Résidus vs valeurs prédites

FIGURE 4.2. Analyse des résidus pour la régression linéaire de la grappe 1 avec transformation Box-Cox $\lambda = -1$

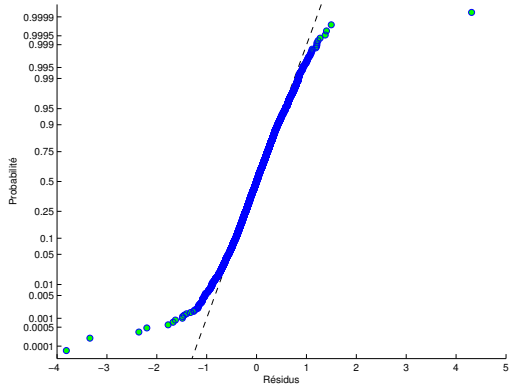


(a) Graphique de probabilités normales

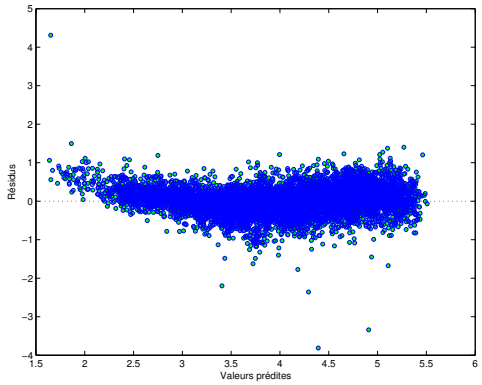


(b) Résidus vs valeurs prédites

FIGURE 4.3. Analyse des résidus pour la régression linéaire de la grappe 1 avec transformation log ($\lambda = 0$)

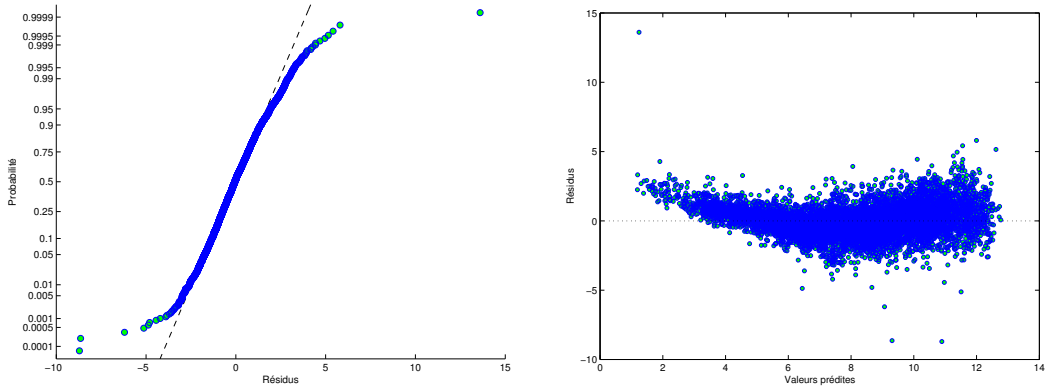


(a) Graphique de probabilités normales



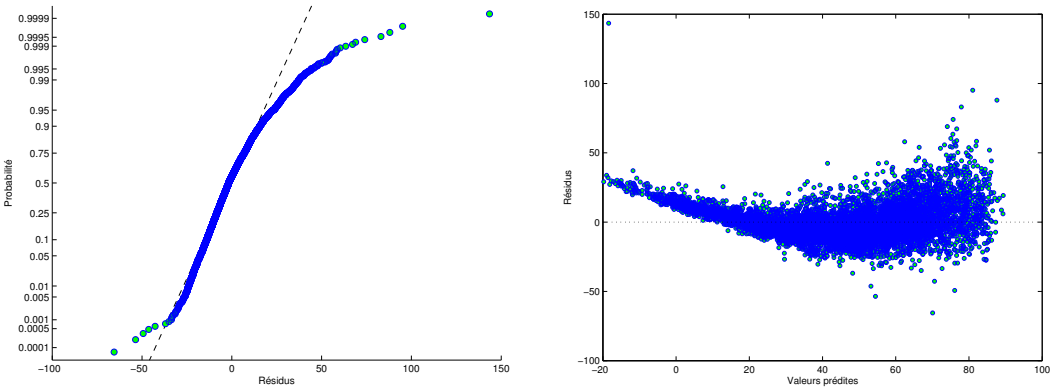
(b) Résidus vs valeurs prédites

FIGURE 4.4. Analyse des résidus pour la régression linéaire de la grappe 1 avec transformation Box-Cox $\lambda = 0,5$



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE 4.5. Analyse des résidus pour la régression linéaire de la grappe 1 transformation Box-Cox $\lambda = 1$



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE 4.6. Analyse des résidus pour la régression linéaire de la grappe 1 avec transformation Box-Cox $\lambda = 2$

- Cavitation ~ Ordonnée à l'origine + 9 prédicteurs (LinNT)
- Log Cavitation ~ Ordonnée à l'origine + 9 prédicteurs (LinLog)
- Cavitation ~ Ordonnée à l'origine + 9 prédicteurs + Interactions de premier ordre (InterNT)
- Log Cavitation ~ Ordonnée à l'origine + 9 prédicteurs + Interactions de premier ordre (InterLog).

La grappe 1 regroupe 13 modes opératoires, dont le détail est disponible à l'annexe A. Sous ces 13 modes opératoires, 7456 données ont été relevées en 2011 pour le groupe 7 et servent à établir le modèle de régression. Notons qu'en général, cette grappe regroupe des modes opératoires avec peu de groupes turbine-alternateur en marche. Le nombre maximal de groupes en marche observé est en effet de quatre (sur une possibilité de huit). La valeur moyenne de cavitation dans cette grappe est de 9,161 kg/10 000 h avec un écart-type de 2,561 kg/10 000 h.

4.2.1. Modèle additif non transformé (LinNT)

Pour le modèle additif non transformé, la méthode descendante retient six variables significatives en plus de l'ordonnée à l'origine.

L'équation pour la cavitation est donnée par :

$$\begin{aligned}\text{Cavitation} = &-79,946 - 0,429 \times \text{Débit} - 0,560 \times \text{Aval} \\ &- 0,009 \times \text{Mvar} + 0,323 \times \text{MW} + 0,267 \times \text{PuissEff} \\ &+ 1,040 \times \text{Tension}\end{aligned}$$

et les détails du modèle sont donnés au tableau 4.2.

TABLEAU 4.2. Détails du modèle additif de régression linéaire sans transformation (LinNT) pour la grappe 1

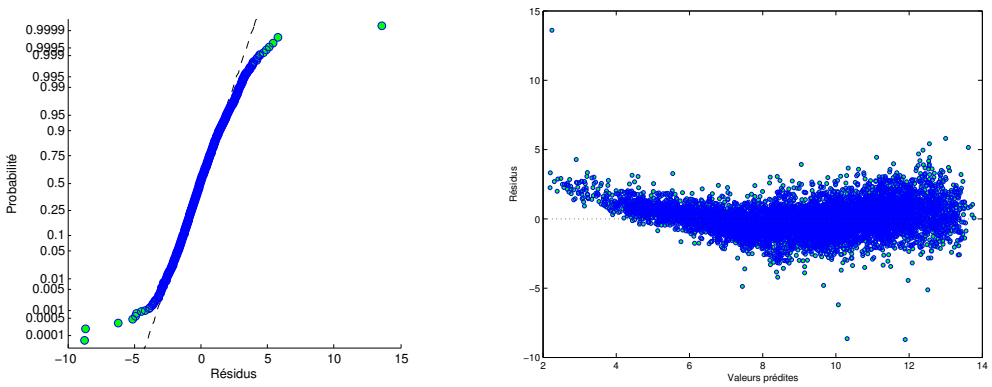
	Estimations	Err. standard	Valeur-p
Ordonnée à l'origine	-79,946	4,489	1,678E-69
Débit	-0,429	0,019	4,522E-114
Aval	-0,560	0,023	6,200E-123
Mvar	-0,009	0,001	5,885E-20
MW	0,323	0,014	3,424E-117
PuissEff	0,267	0,010	6,732E-163
Tension	1,040	0,140	1,358E-13
$R_a^2 = 0,802$			

L'analyse des résidus montre une certaine courbure dans le graphique des valeurs prédites contre les résidus (voir figure 4.7(b)), telle que discutée dans la section 4.1.1. Malgré cette courbure, l'ajustement est assez adéquat et la valeur du coefficient de détermination ajusté R_a^2 est de $0,803$. De plus, la figure 4.7(a) présente le graphique de probabilités normales. Certains points s'éloignent de la droite, mais la majorité semble être cohérente avec une loi normale, avec des queues un peu plus concentrées.

4.2.2. Modèle additif avec transformation log (LinLog)

Pour le modèle additif avec transformation log, la méthode descendante retient sept variables significatives en plus de l'ordonnée à l'origine. Les détails du modèle sont donnés au tableau 4.3.

La transformation log ne semble pas avoir d'effet sur l'amélioration de la normalité des résidus (voir figure 4.7(a) et figure 4.8(a)). Cependant, nous pouvons voir à la figure 4.8(b) que le comportement des résidus est grandement amélioré. La courbure et l'hétéroscédasticité des résidus produits par le modèle LinNT



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE 4.7. Analyse des résidus pour la régression linéaire de la grappe 1 après sélection des variables (modèle LinNT)

TABLEAU 4.3. Détails du modèle de régression linéaire avec transformation log (LinLog) pour la grappe 1

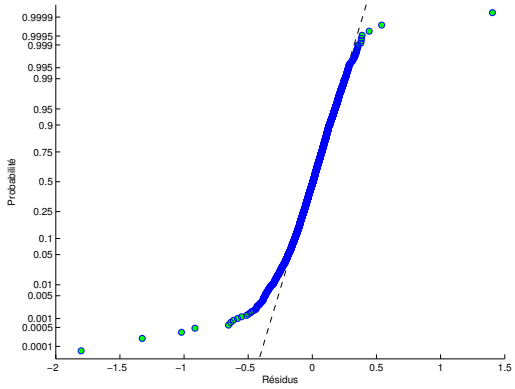
	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-6,621	0,590	5,089E-29
Débit	-0,053	0,002	2,243E-130
Amont	-0,016	0,008	3,544E-02
Aval	-0,039	0,008	7,785E-07
Mvar	-0,001	1E-4	9,899E-07
MW	0,039	0,002	5,810E-131
PuissEff	0,035	0,002	8,928E-53
Tension	0,059	0,014	4,329E-05
$R_a^2 = 0,832$			

semblent s'être beaucoup atténuées (voir figure 4.7(b)). Le modèle avec transformation log paraît mieux respecter les hypothèses de répartition aléatoire autour de zéro et de variance constante pour la grappe 1.

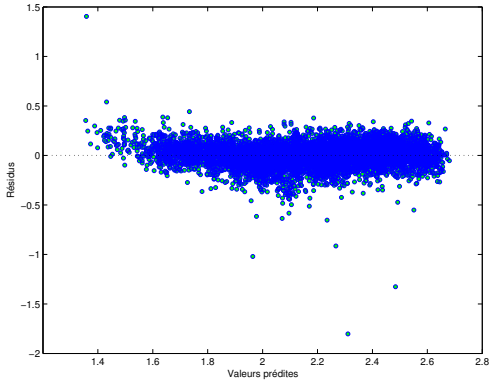
4.2.3. Modèle avec interactions sans transformation (InterNT)

Pour le modèle avec interactions sans transformation sur la variable cavitation, la méthode descendante retient les neuf effets principaux et 21 interactions en plus de l'ordonnée à l'origine (31 termes). Les détails du modèle sont donnés au tableau 4.4.

Le modèle avec interactions de la cavitation non transformée (InterNT) ne semble pas avoir d'effet sur la normalité des résidus lorsque nous le comparons au modèle additif (LinNT) : la figure 4.9(a) ressemble à la figure 4.7(a). Cependant, la courbure des résidus que nous observions à la figure 4.7(b) semble avoir disparu,

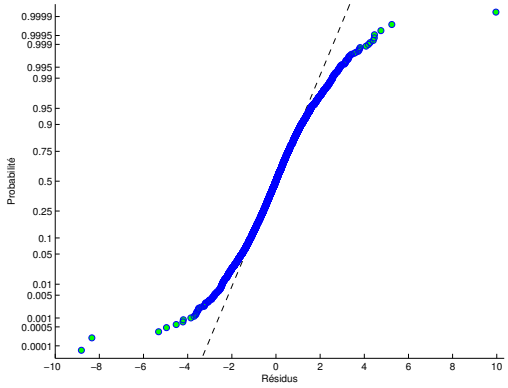


(a) Graphique de probabilités normales

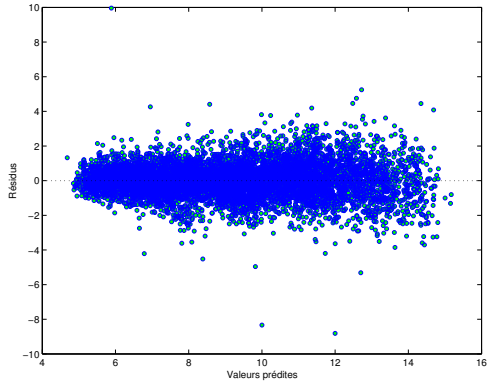


(b) Résidus vs valeurs prédites

FIGURE 4.8. Analyse des résidus pour la régression linéaire de la grappe 1 après sélection des variables (modèle LinLog)



(a) Graphique de probabilités normales



(b) Résidus vs valeurs prédites

FIGURE 4.9. Analyse des résidus pour la régression linéaire de la grappe 1 après sélection des variables (modèle InterNT)

quoiqu'une légère hétéroscédasticité soit toujours décelable (voir figure 4.9(b)). Le modèle avec interactions (InterNT) paraît mieux respecter les hypothèses de répartition aléatoire autour de zéro et de variance constante pour la grappe 1 que le modèle additif (LinNT).

4.2.4. Modèle avec interactions avec transformation log (InterLog)

Pour le modèle avec interactions avec transformation log sur la variable cavitation, la méthode descendante retient les neuf effets principaux et 25 interactions en plus de l'ordonnée à l'origine (35 termes). Les détails du modèle sont donnés au tableau 4.5.

L'analyse des résidus du modèle avec interactions et transformation log (InterLog) est disponible à la figure 4.10. Lorsque nous les comparons avec les résidus

TABLEAU 4.4. Détails du modèle de régression avec interactions de premier ordre sans transformation (InterNT) pour la grappe 1

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-1750,400	415,240	2,524E-05
Courant	39,243	10,022	9,094E-05
Debit	27,907	2,847	1,506E-22
Amont	1,966	2,660	4,598E-01
Aval	19,907	4,814	3,589E-05
Ouverture	-27,046	6,927	9,527E-05
Mvar	-0,806	0,166	1,302E-06
MW	-18,251	1,296	1,892E-44
PuissEff	3,527	0,566	4,986E-10
Tension	101,150	25,879	9,362E-05
Courant :Debit	-0,229	0,059	9,645E-05
Courant :MW	0,201	0,042	2,375E-06
Courant :PuissEff	-0,125	0,025	5,969E-07
Debit :Amont	-0,146	0,014	1,110E-25
Debit :Ouverture	0,075	0,010	4,153E-15
Debit :Mvar	0,002	0,001	6,202E-03
Debit :Tension	-0,365	0,123	3,100E-03
Amont :Aval	0,074	0,031	1,790E-02
Amont :Ouverture	0,275	0,045	6,842E-10
Amont :Mvar	0,005	0,001	3,314E-07
Amont :MW	0,068	0,009	6,335E-14
Amont :Tension	-0,680	0,141	1,532E-06
Aval :Ouverture	-0,144	0,022	8,650E-11
Aval :PuissEff	-0,065	0,010	1,588E-10
Ouverture :Mvar	-0,006	0,002	2,636E-03
Ouverture :MW	-0,066	0,007	7,909E-19
Ouverture :PuissEff	-0,062	0,012	1,800E-07
Ouverture :Tension	0,616	0,291	3,402E-02
Mvar :Tension	0,007	0,002	1,301E-03
MW :PuissEff	0,016	0,002	1,361E-13
MW :Tension	0,195	0,057	6,375E-04
$R_a^2 = 0,852$			

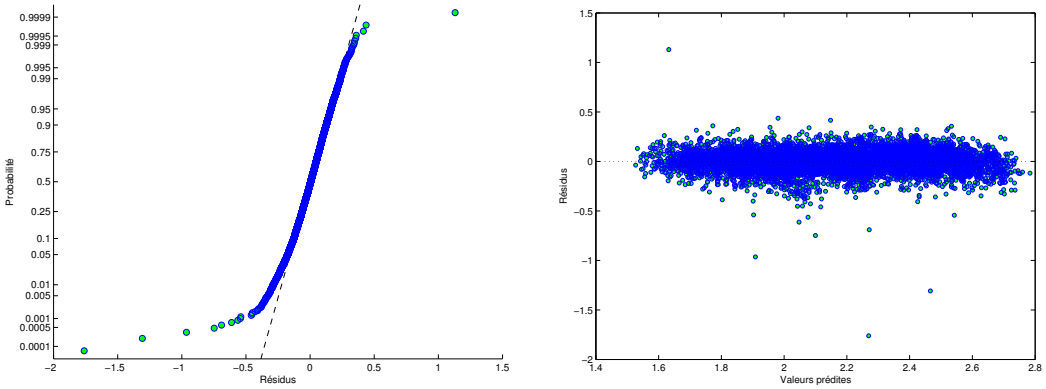
du modèle additif sans transformation log (LinLog) (voir figure 4.8), nous pouvons voir que le modèle avec interactions présente encore moins de courbure et d'hétéroscédasticité que le modèle additif. La normalité des résidus semble quant à elle inchangée.

Ainsi, nous avons pu voir pour la grappe 1 que le modèle additif et sans transformation (LinNT) présentait certaines lacunes au niveau des résidus (courbure et hétéroscédasticité). Le modèle avec interactions (InterNT) permet de bien atténuer la courbure, mais les résidus conservent une certaine hétéroscédasticité. La transformation log sur le modèle additif (LinLog) semble aussi diminuer la courbure des résidus. Le modèle avec interactions et transformation log (InterLog) présente les résidus qui semblent le plus adéquats. Cependant, étant donné le

TABLEAU 4.5. Détails du modèle de régression avec interactions de premier ordre avec transformation log (InterLog) pour la grappe 1

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-324,030	66,405	1,085E-06
Courant	10,786	1,983	5,510E-08
Débit	3,894	0,442	1,432E-18
Amont	1,242	0,412	2,547E-03
Aval	-0,108	0,802	8,930E-01
Ouverture	-5,339	0,936	1,220E-08
Mvar	-0,101	0,019	9,587E-08
MW	-2,335	0,226	7,201E-25
PuissEff	0,360	0,074	1,207E-06
Tension	22,840	4,157	4,039E-08
Courant :Débit	-0,062	0,010	6,037E-10
Courant :MW	0,049	0,007	1,084E-11
Courant :PuissEff	-0,030	0,005	1,559E-09
Débit :Amont	-0,025	0,003	4,348E-14
Débit :Ouverture	0,014	0,001	1,977E-20
Débit :Mvar	0,0002	0,0001	1,625E-03
Débit :PuissEff	0,004	0,001	3,586E-06
Débit :Tension	-0,098	0,020	5,656E-07
Amont :Aval	0,013	0,004	3,739E-04
Amont :Ouverture	0,053	0,011	6,494E-07
Amont :Mvar	0,001	0,0002	1,746E-05
Amont :MW	0,009	0,001	1,672E-28
Amont :Tension	-0,161	0,026	7,024E-10
Aval :Ouverture	-0,035	0,008	3,857E-05
Aval :Mvar	-0,001	0,0003	1,494E-03
Aval :MW	0,004	0,002	2,342E-02
Aval :PuissEff	-0,007	0,001	2,933E-07
Aval :Tension	0,124	0,031	7,405E-05
Ouverture :Mvar	-0,001	0,0002	9,168E-04
Ouverture :MW	-0,012	0,001	3,536E-24
Ouverture :PuissEff	-0,010	0,003	7,156E-04
Ouverture :Tension	0,082	0,034	1,544E-02
Mvar :PuissEff	-0,0002	0,0001	4,613E-02
Mvar :Tension	0,001	0,0003	9,479E-03
MW :Tension	0,061	0,011	5,489E-08
$R_a^2 = 0,855$			

faible gain au niveau du R^2 ajusté (+2,3 % lorsque les interactions sont ajoutées au modèle log) et l'allure très acceptable des résidus du modèle additif (LinLog), il est possible que le choix final favorise un modèle plus simple. La validation du chapitre suivant sur la prédiction des données 2012 servira à choisir lequel de ces quatre modèles sera le plus adéquat en termes de prédiction.



(a)Graphique de probabilités normales (b)Résidus vs valeurs prédites

FIGURE 4.10. Analyse des résidus pour la régression linéaire de la grappe 1 après sélection des variables (modèle InterLog)

Quoique cette section s'attardait seulement sur l'analyse des résidus de la grappe 1, nous avons considéré toutes les grappes et les modèles avec l'analyse des résidus des autres grappes est présenté à l'annexe C.

4.3. Résumé des résultats

Dans cette section, nous présentons un résumé des quatre modèles par grappe développés avec la régression linéaire qui seront validés dans le chapitre suivant. Nous pouvons voir au tableau 4.6 que le nombre de paramètres retenus par la méthode descendante est assez élevé. Pour le modèle de départ avec effets principaux seulement (modèle additif), le plus petit modèle non transformé retient six variables explicatives, en plus de l'ordonnée à l'origine, alors que le plus petit modèle avec transformation log retient quatre variables. Les variables retenues pour les modèles additifs sont disponibles au tableau 4.7. Le même phénomène se produit pour les modèles avec interactions. Sur un maximum de 45 coefficients, en plus de l'ordonnée à l'origine, le modèle en retient au minimum 24 et au maximum 37, selon les grappes. Notons que pour chaque grappe, tous les effets principaux sont retenus, puisque nous retrouvons au moins une interaction contenant l'effet principal qui est significative. Les tableaux 4.8 et 4.9 présentent les variables retenues pour les modèles avec interactions pour la grappe 1. Un graphique présentant les nuages de points de toutes les interactions est aussi présenté à la figure 4.11 pour la grappe 1 (voir annexe D pour les autres grappes.)

La question qui se pose par la suite est de déterminer si les modèles avec interactions apportent de l'information supplémentaire, étant donné leur complexité. Nous présentons au tableau 4.10 les coefficients de détermination ajustés R_a^2 pour comparer les différents modèles. Pour chaque grappe, le modèle qui maximise le

TABLEAU 4.6. Nombres de paramètres retenus par la méthode descendante, excluant l'ordonnée à l'origine

Grappe	LinNT	LinLog	InterNT	InterLog
1	6	7	30	34
2	7	6	30	31
3	6	7	24	25
4	6	4	27	24
5	8	8	37	37
Modèle complet	9	9	45	45

TABLEAU 4.7. Variables conservées dans les modèles de régression linéaires après sélection des variables

Grappe	Modèle	Courant	Débit	Amont	Aval	Ouv.	Mvar	MW	PuissEff	Tension
1	LinNT		x		x		x	x	x	x
	LinLog		x	x	x		x	x	x	x
2	LinNT	x	x	x			x	x	x	x
	LinLog	x	x	x			x	x	x	
3	LinNT	x	x		x			x	x	x
	LinLog	x	x		x	x		x	x	x
4	LinNT	x	x	x			x		x	x
	LinLog	x	x	x					x	
5	LinNT	x	x	x	x		x	x	x	x
	LinLog	x	x	x	x		x	x	x	x

TABLEAU 4.8. Variables conservées dans le modèle de régression avec interactions sans transformation après sélection des variables (InterNT) pour la grappe 1

	Courant	Débit	Amont	Aval	Ouverture	Mvar	MW	PuissEff	Tension
Courant	x	x					x	x	
Débit		x	x		x	x			x
Amont			x	x	x	x	x		x
Aval				x	x			x	
Ouverture					x	x	x	x	x
MW						x			x
Mvar							x	x	x
PuissEff								x	
Tension									x

R^2 ajusté est mis en évidence en rouge. Le gain pour le R^2 ajusté lorsque nous considérons les modèles avec interactions (avec ou sans transformations) varie entre $+2,3\%$ pour la grappe 1 avec log à $+15,3\%$ pour la grappe 5 non transformée. Notons que nous ne présentons pas ici une comparaison avec le critère BIC, puisque celle-ci est valide seulement si nous comparons différents modèles sur les

TABLEAU 4.9. Variables conservées dans le modèle de régression avec interactions avec transformation log après sélection des variables (InterLog) pour la grappe 1

	Courant	Débit	Amont	Aval	Ouv.	Mvar	MW	PuissEff	Tension
Courant	x	x					x	x	
Débit		x	x		x	x		x	x
Amont			x	x	x	x	x		x
Aval				x	x	x	x	x	x
Ouverture					x	x	x	x	x
Mvar						x		x	x
MW							x		x
PuissEff								x	
Tension									x

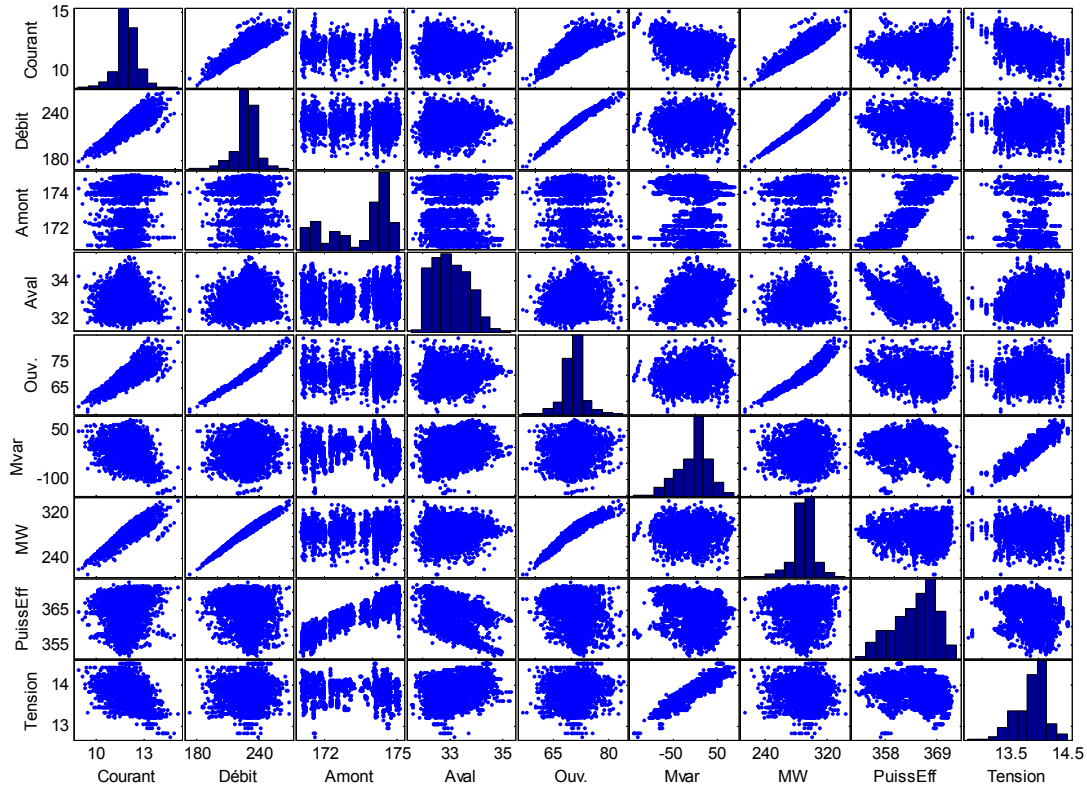


FIGURE 4.11. Nuages de points des interactions entre les neuf variables indépendantes pour la grappe 1

mêmes données. Elle ne permet donc pas, par exemple, de comparer un modèle transformé log avec un modèle non transformé, tel que c'est le cas ici.

Ainsi, ce chapitre a permis de présenter quatre modèles de régression par grappe, en utilisant neuf variables opératoires comme variables explicatives. La

TABLEAU 4.10. R^2 ajusté pour les modèles retenus par la méthode descendante

Grappe	LinNT	LinLog	InterNT	InterLog
1	0,802	0,832	0,852	0,855
2	0,742	0,713	0,813	0,774
3	0,705	0,711	0,758	0,759
4	0,792	0,787	0,909	0,888
5	0,403	0,419	0,556	0,564

variable *hauteur de chute* a été retirée. Les modèles considérés utilisent une transformation log ou aucune transformation sur la variable réponse *cavitation*. De plus, nous avons considéré un modèle avec effets principaux seulement (modèle additif) ou avec interactions de premier ordre incluses, en utilisant la méthode de sélection descendante. Notons que les quatre modèles retenus contiennent un nombre élevé de variables par rapport au modèle complet de départ. Pour comparer les modèles, nous avons utilisé le critère du R^2 ajusté, qui semble favoriser les modèles avec interactions. Cependant, étant donné le nombre de paramètres élevé, ce critère n'est pas garant d'une bonne prédiction et le chapitre suivant s'attardera à développer la validation des modèles sur un jeu de données différent, soient les données pour l'année 2012. Le choix du meilleur modèle par grappe en termes de prédiction sera aussi présenté au chapitre suivant.

Chapitre 5

VALIDATION DU MODÈLE

Suite à l'élaboration des quatre modèles de régression par grappe développés au chapitre précédent, nous présentons ici les différentes étapes qui mènent à la sélection du meilleur modèle pour chacune des grappes. Dans un premier temps, nous évaluons la robustesse du choix des prédicteurs sur les données 2011. Nous validons par la suite les modèles sur les données de l'année 2012, et la sélection du meilleur modèle par grappe est faite en se basant notamment sur l'erreur quadratique moyenne de prédiction (*MSEP*). Finalement, la dernière section est consacrée au calcul de l'érosion cumulative et à la comparaison de l'érosion cumulative basée sur les modèles statistiques avec l'érosion cumulative basée sur les données d'érosion instantanée du Caviciel. Notons que pour la suite de ce chapitre, le vocabulaire « vraies valeurs de cavitation » se réfère aux données d'érosion instantanée du Caviciel, c'est-à-dire aux données qui ont servi à construire les modèles statistiques.

5.1. Robustesse du choix des prédicteurs

Dans cette section, la robustesse du choix des prédicteurs est évaluée en utilisant la méthode de validation croisée avec omission de m observations (*Leave-m-out*)(Kohavi, 1995). Un échantillon d'apprentissage est formé en sélectionnant aléatoirement 2/3 des données pour l'année 2011 . Dans cet échantillon, nous identifions les données qui correspondent à la grappe g, ($g = 1, ..., 5$), et les valeurs des coefficients sont calculées en appliquant la régression linéaire multiple avec le choix des prédicteurs tel que trouvé dans les modèles présentés au chapitre 4 (section 4.2) et à l'annexe C. Puis, le modèle est validé avec l'échantillon de test, qui équivaut au tiers restant des données. Pour ce faire, les données de l'échantillon test correspondant à chacune des grappes sont identifiées, et le modèle de la grappe adéquate est utilisé pour calculer les prédictions de cavitation. Pour chaque grappe, et pour les cinq modèles (quatre modèles de régression + le

modèle de mélange de lois du chapitre 5), le processus est répété 50 fois et nous présentons les diagrammes en boîte du critère BIC de prédiction pour comparer la robustesse des cinq modèles (voir figure 5.1). L'équation du BIC de prédiction est :

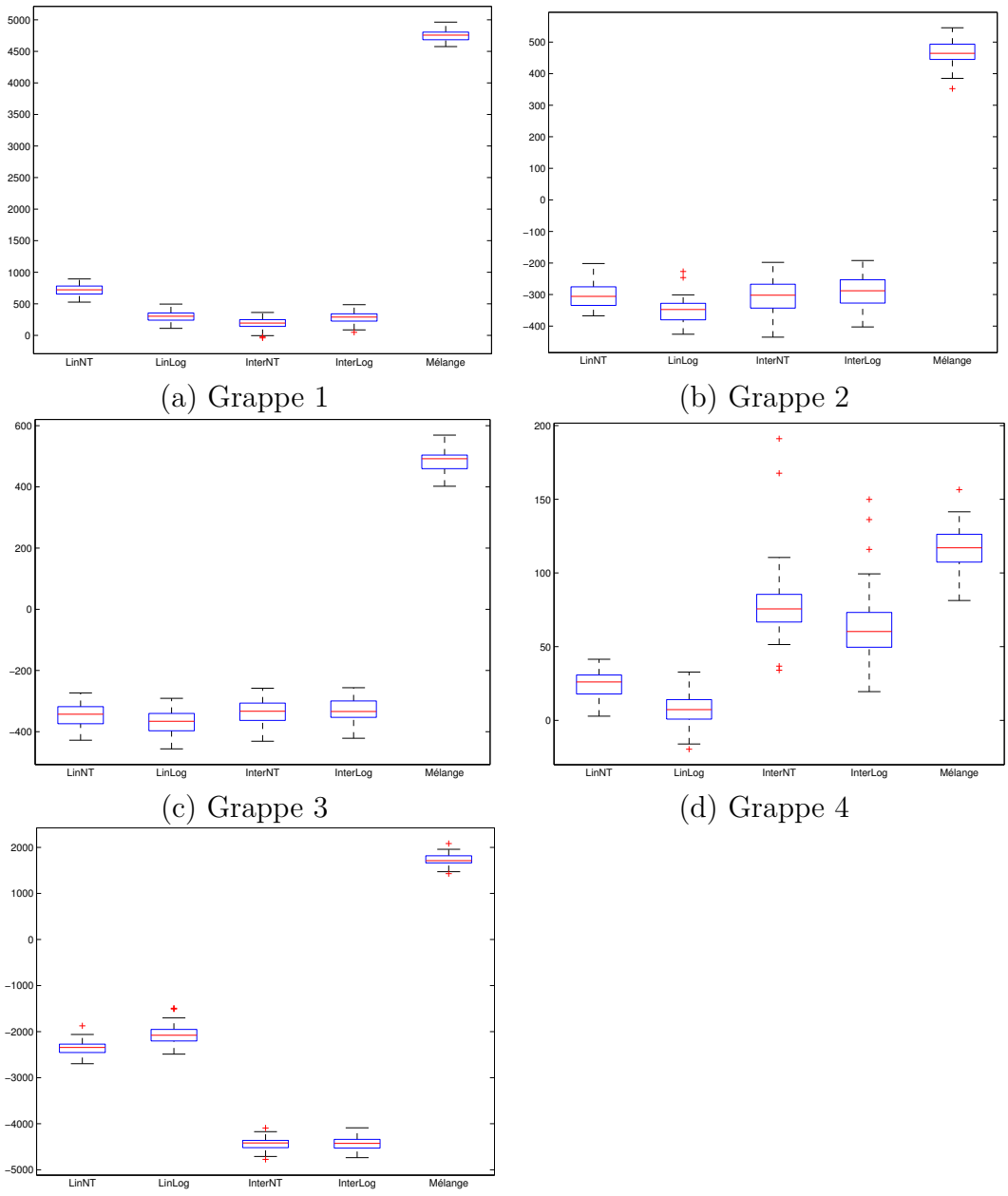
$$BIC_{préd} = n\log(SSE) - n\log(n) + p\log(n), \tag{5.1.1}$$

où p est le nombre de paramètres du modèle construit avec l'échantillon d'apprentissage et n le nombre d'observations de l'échantillon-test.

Sans surprise, nous pouvons voir à la figure 5.1 que l'estimation de la cavitation par la moyenne théorique du mélange de lois tel que donnée au chapitre 5 présente une médiane du BIC de prédiction beaucoup plus élevée que les quatre autres modèles, basés sur la régression. La suite de la validation se concentrera donc sur les quatre modèles de régression.

Concernant la robustesse des prédicteurs, nous pouvons voir que les quatre modèles des grappes 1 et 2 présentent des BIC avec des mesures de dispersion similaires. Le modèle avec interactions non transformées (InterNT) semble légèrement favorisé pour la grappe 1 alors que le modèle sans interactions avec transformation log (LinLog) semble le meilleur pour la grappe 2 (voir figure 5.1 (a) et (b)). Nous rappelons qu'un bon modèle est caractérisé par un BIC le plus petit possible. Les diagrammes en boîte de la grappe 3 (voir figure 5.1(c)), quant à eux, ne présentent pas de différences notables au niveau des quatre modèles de régression, tant du point de vue de la médiane que de la dispersion. La grappe 4 présente un meilleur BIC pour les modèles sans interactions (voir figure 5.1 (d)) et le modèle de mélange de lois performe assez bien relativement aux modèles de régression. À l'opposé, pour la grappe 5, les modèles avec interactions sont meilleurs du point de vue du BIC que les modèles sans interactions (voir figure 5.1(e)), la médiane étant beaucoup plus basse avec des valeurs de dispersion similaire.

Somme toute, concernant la robustesse du choix des prédicteurs, les diagrammes en boîte nous permettent de voir que les modèles avec interactions semblent définitivement meilleurs pour la grappe 5 et peu performants pour la grappe 4. Pour les trois autres grappes, les quatre modèles de régression paraissent relativement équivalents. Les modèles de mélange de lois, pour leur part, sont définitivement écartés de l'analyse à partir de ce point. Rappelons en outre qu'au chapitre 4, nous avons vu que le modèle InterLog maximisait le R^2 ajusté pour les grappes 1, 3 et 5, alors que le modèle InterNT maximisait le R^2 ajusté pour les grappes 2 et 4 (voir tableau 4.10).



(a) Grappe 1

(b) Grappe 2

(c) Grappe 3

(d) Grappe 4

(e) Grappe 5

FIGURE 5.1. Diagrammes en boîte des BIC de prédiction des cinq modèles construits par validation croisée (Annéee 2011)

Ainsi, en regard de ces critères, aucun modèle ne semble être uniformément performant dans toutes les grappes. Le R_a^2 semble pencher vers les modèles avec interactions (voir tableau 4.10), alors que les diagrammes en boîte ne favorisent pas le même modèle d'une grappe à l'autre (voir figure 5.1). Cependant, étant donné que l'intérêt de ces modèles réside en premier lieu dans leur capacité à faire

des bonnes prédictions, le critère de sélection favorisé au final sera plutôt basé sur la validation sur le jeu de données 2012, présentée dans la section suivante.

5.2. Validation sur les données de l'année 2012

Pour valider le modèle sur les observations de l'année 2012, la stratégie est d'identifier le mode opératoire correspondant à chaque observation et d'associer la valeur de la grappe telle que donnée par le dendrogramme (voir figure 2.3). Une fois identifiée la grappe d'appartenance de l'observation, nous ajustons tous les modèles pour prédire cette observation. Par la suite, nous déterminons le meilleur modèle par grappe g en minimisant l'erreur quadratique moyenne de prédiction (MSE_g), donnée par l'équation :

$$MSE_g = \frac{1}{n_g}\sum_{i=1}^{n_g}(u_i - \widehat{u}_i)^2,$$

où n_g est le nombre d'observations de l'année 2012 dans la grappe g, $\widehat{u}_i$ est la valeur prédite de cavitation et u_i est la vraie valeur de cavitation. Notons que $\widehat{u}_i = \widehat{y}_i$ si la cavitation est non transformée et $\widehat{u}_i = \exp(\widehat{\log(y_i)})$ si la cavitation est log-transformée.

5.2.1. Nouveaux modes opératoires

Un des défis que présentent les données 2012 est la présence de modes opératoires non observés en 2011, et par conséquent, non attribués à une grappe. Ces modes opératoires et leur fréquence sont donnés au tableau 5.1. Ils sont au nombre de 16 (sur 72 modes opératoires observés au total en 2012) et représentent 1,32 % des données de la zone opératoire principale. Nous les associons à l'une des cinq grappes en choisissant celle qui minimise la variance intra-grappe (VIG), donnée par l'équation :

$$\sum_{g=1}^{G}\sum_{k=1}^{m_g}(\mathbf{x_{gk}} - \overline{\mathbf{x}}_\mathbf{g})(\mathbf{x_{gk}} - \overline{\mathbf{x}}_\mathbf{g})', \tag{5.2.1}$$

où $G = 5$ est le nombre de grappes, m_g est le nombre de modes opératoires dans la grappe g, x_{gk} le vecteur de 31 caractéristiques standardisées associées au mode opératoire k et $\overline{x}_g$, le vecteur des moyennes des 31 caractéristiques des modes opératoires dans la grappe g.

Pour déterminer quelle grappe est attribuée à ces modes opératoires, la première étape est de calculer les caractéristiques d'intérêt liées à chaque mode opératoire, tels que définies à la section 2.2.1 du chapitre 2, c'est-à-dire que pour

Tableau 5.1. Modes opératoires 2012 non observés en 2011

ModeOp	Fréq.	Fréq. (%)	Grappe assignée	Moyenne cavitation	Écart-type cavitation	Médiane cavitation
« 3678 »	29	0,001	2	9,934	0,963	9,734
« 2467 »	1	<0,001	3	4,356	<0,001	4,356
« 2567 »	36	0,001	3	5,663	0,462	5,625
« 12567 »	120	0,003	3	6,776	0,621	6,772
« 12347 »	3	<0,001	3	5,268	0,160	5,239
« 2347 »	2	<0,001	3	10,146	0,419	10,146
« 1378 »	27	0,001	3	10,304	0,718	10,142
« 13678 »	26	0,001	3	8,739	0,794	8,648
« 234678 »	61	0,002	3	7,258	0,872	7,342
« 124567 »	1	<0,001	4	4,618	<0,001	4,618
« 247 »	8	<0,001	4	10,884	1,181	10,376
« 245678 »	15	<0,001	5	5,508	0,461	5,446
« 12467 »	4	<0,001	5	4,332	0,204	4,342
« 124678 »	3	<0,001	5	5,262	0,185	5,323
« 1467 »	72	0,002	5	5,247	0,462	5,310
« 125678 »	60	0,002	5	5,398	0,414	5,407

Note : La fréquence relative est donnée sur le nombre d'observations total en 2012 dans la zone d'opération principale (zones 3 à 5) : n = 35404

chacun des modes opératoires, nous utilisons la moyenne, l'écart-type et la médiane des 10 variables explicatives, ainsi que la fréquence du mode opératoire. Ainsi, chacun des modes opératoires est représenté par un vecteur de 31 caractéristiques, exactement selon la même stratégie qu'en 2011. Le vecteur est ensuite standardisé en utilisant les mêmes moyennes et écart-types que pour l'année 2011.

Pour chaque nouveau mode opératoire de 2012, une des cinq grappes lui est assignée et la variance intra-grappe est calculée (voir équation (5.2.1)). Le choix de la grappe finale est l'assignation qui minimise la variance intra-grappe. Les résultats sont donnés au tableau 5.1. Parmi les nouveaux modes opératoires observés en 2012, nous pouvons voir qu'aucun mode opératoire n'est assigné à la grappe 1, qu'un seul est assigné à la grappe 2, 8 à la grappe 3, 2 à la grappe 4 et 5 à la grappe 5.

5.2.2. Statistiques descriptives pour l'année 2012

Tous les modes opératoires observés en 2012 ont maintenant une grappe correspondante, qui les groupe selon leurs caractéristiques opératoires similaires. Le tableau 5.2 nous présente les statistiques descriptives par grappe des valeurs de cavitation de l'année 2012 pour la zone d'opération principale. Comme en 2011 (voir tableau 3.4), la grappe 5 présente la plus petite moyenne de cavitation et

est peu dispersée. C'est aussi la grappe qui contient de loin le plus de données. À l'opposé, la grappe 1 montre une moyenne plus élevée de cavitation avec de grandes valeurs de dispersion, tout comme c'était le cas en 2011. Notons que la grappe 4, caractérisée par des modes opératoires rares, contient très peu de valeurs en 2012 ($n = 46$).

TABLEAU 5.2. Statistiques de cavitation par grappe pour les données de l'année 2012

Grappe	1	2	3	4	5
Moyenne	8,795	7,237	6,284	7,402	5,981
Écart-type	2,309	1,840	1,164	2,124	1,213
Minimum	1,456	3,703	3,105	4,618	2,631
Maximum	18,342	13,490	11,992	13,643	14,336
Médiane	8,374	7,095	6,124	6,688	5,924
Fréquence	6129	1031	1907	46	26125

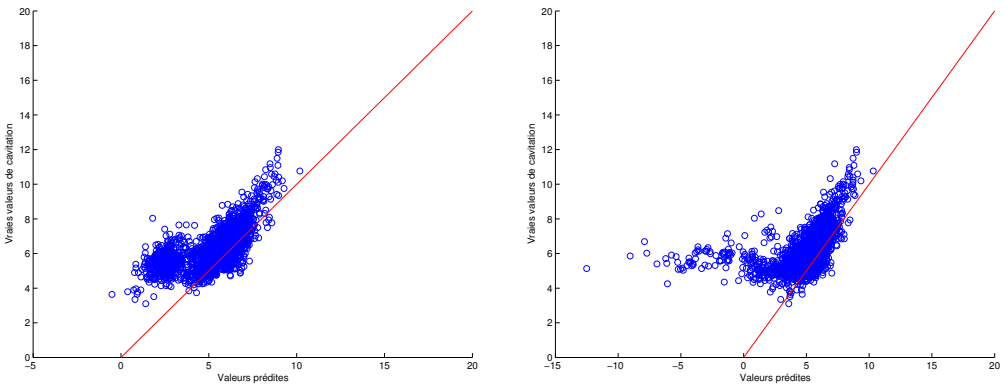
5.3. Sélection du modèle

Pour chaque grappe, les valeurs de cavitation de 2012 sont prédites avec les quatre modèles de régression correspondants. Avant de pouvoir procéder à la sélection du modèle, nous remarquons que certains modèles induisent des valeurs de prédiction incorrectes : certaines valeurs d'érosion sont négatives, et d'autres sont très éloignées de l'étendue dans laquelle nous observons généralement les valeurs de cavitation. La section suivante présente les différentes décisions liées à ces valeurs de prédiction incorrectes.

5.3.1. Traitement des valeurs de prédiction incorrectes

Les prédictions à valeurs négatives sont observées dans les modèles sans transformations, et seulement dans les grappes 3 à 5. La grappe 3 présente une seule valeur négative pour le modèle LinNT et 70 pour le modèle InterNT. La grappe 4 présente quant à elle 29 et 35 valeurs négatives pour les modèles LinNT et InterNT respectivement, alors que la grappe 5 en a une seule pour le modèle LinNT. Les modèles log quant à eux évitent ce problème et nous en tiendrons compte au moment de la sélection finale du modèle. Les figures 5.2 et 5.3 montrent la relation entre les vraies valeurs et les valeurs prédites en 2012 pour les grappes 3 et 4. Nous voyons en particulier que les modèles avec interactions produisent certaines valeurs négatives beaucoup plus éloignées de zéro.

Par ailleurs, une seule donnée produit des valeurs de prédictions positives très éloignées de l'étendue de cavitation habituelle. Ces valeurs s'observent dans

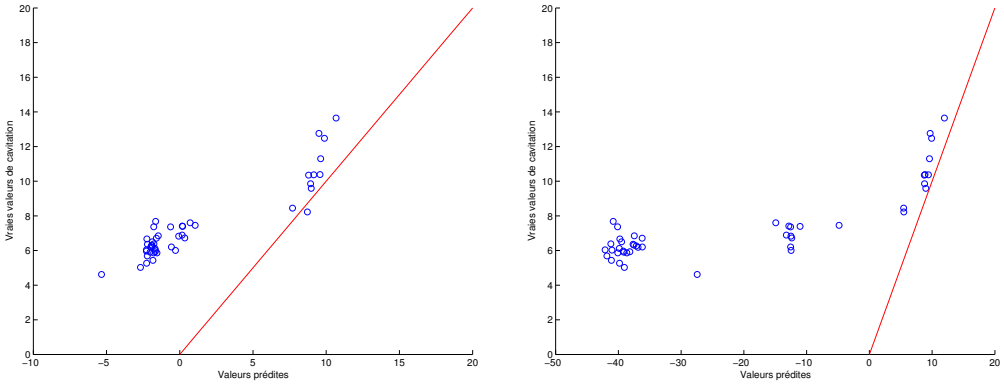


(a) LinNT (b) InterNT

FIGURE 5.2. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation pour la Grappe 3

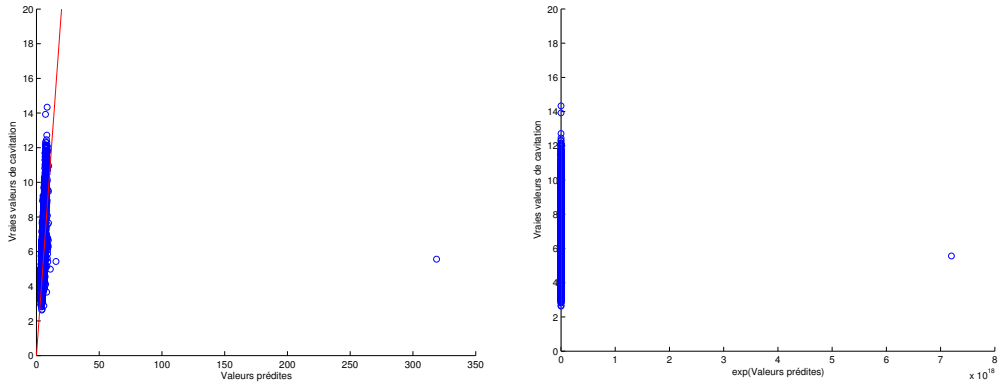
la grappe 5 et seulement dans les modèles avec interactions (voir figure 5.4). Le détail des paramètres de la donnée est donné à la table 5.3. Nous pouvons voir que la vraie valeur de cavitation est de 5,562 kg/10 000 h, mais que les prédictions avec les modèles avec interactions sont respectivement de 318,783 et de $7,200E+18$ pour le modèle sans et avec transformation log. Ceci peut être expliqué par le fait que le courant et la tension sont à 0, alors que la puissance active est à 289,035 MW, ce qui est improbable. Cette erreur dans la donnée pourrait être due à la distance de temps maximale de 2 minutes 30 entre les données opératoires et les données de cavitation. Dans le cas qui nous concerne, les données opératoires sont en effet observées 2 minutes 14 plus tôt que la donnée de cavitation et ceci pourrait expliquer la difficulté à obtenir une prédiction.

Pour prendre une décision éclairée pour la sélection de modèle, nous avons donc consulté les experts pour discuter de la correction des valeurs négatives de prédiction et de la valeur aberrante de prédiction. Les valeurs négatives sont ainsi remplacées par 0. Pour pallier à toute valeur aberrante de prédiction dans le futur, nous établissons un seuil à 40 kg/10 000 h pour le groupe turbine alternateur 7 au-dessus duquel toutes les valeurs prédites de cavitation sont remplacées par la moyenne théorique de la grappe telle que donnée par les modèles de mélanges du chapitre 3 (voir tableau 3.4). Ce choix de correction s'explique par les contraintes d'implémenter en temps réel une méthode simple. Dans le cas présent, nous remplaçons donc la prédiction par la moyenne théorique de la grappe 5, c'est-à-dire 5,454 kg/10 000 h.



(a) LinNT (b) InterNT

FIGURE 5.3. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation pour la Grappe 4



(c) InterNT (d) InterLog

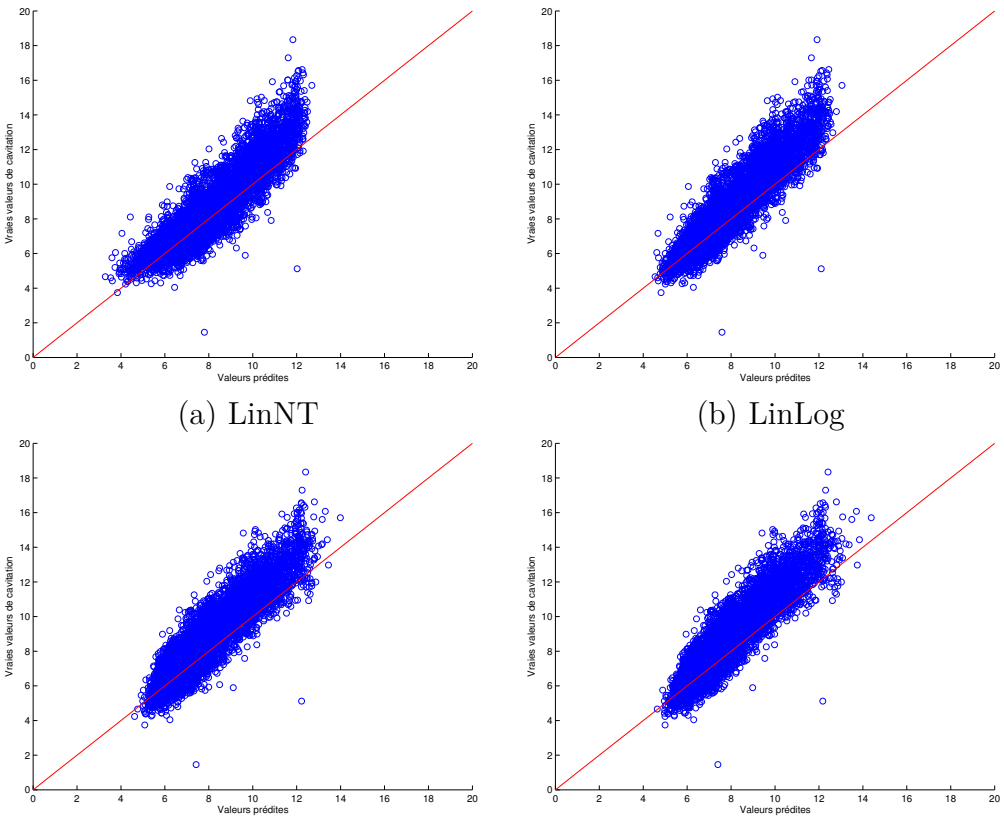
FIGURE 5.4. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation pour la Grappe 5

TABLEAU 5.3. Valeur aberrante de prédiction pour la grappe 5 avant correction

Date	Cavitation	Courant	Débit	Amont	Aval	Ouv.	Mvar
2012-02-18 17 :02 :14	5,562	0,000	233,177	173,152	36,630	71,938	1,350
MW	PuissEff	Tension	Pred LinNT	Pred LinLog	Pred InterNT	Pred InterLog	
289,035	353,214	0,000	-24,567	0,075	318,783	7,200E+18	

5.3.2. Comparaison et choix des modèles finaux

Les figures 5.5 à 5.9 montrent la relation entre les vraies valeurs et les valeurs prédites après corrections des valeurs incorrectes. Pour la grappe 1, nous voyons que les quatre modèles sont à peu près équivalents et que le modèle prédit bien les vraies valeurs. Pour la grappe 2, les modèles avec interactions linéarisent mieux la relation. Par ailleurs, pour les grappes 3 et 4, les modèles LinNT, InterNT



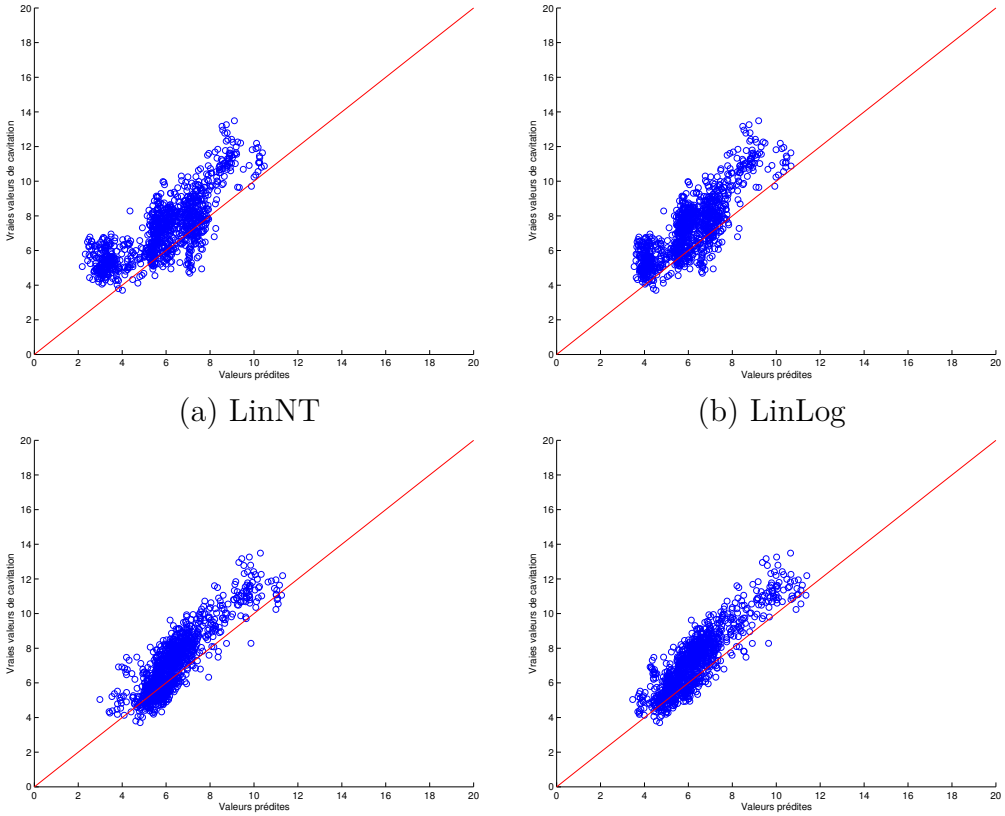
(a) LinNT (b) LinLog

(c) InterNT (d) InterLog

FIGURE 5.5. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation après correction pour la Grappe 1

et InterLog paraissent moins bien fonctionner pour l'aspect prédictif. En effet, une certaine sous-estimation est observable, surtout pour les petites valeurs, et la relation est courbée plutôt que linéaire. Le modèle LinLog est celui qui paraît le plus adapté. En ce qui concerne la grappe 5 (voir figure 5.9), les modèles sans interactions produisent des prédictions plutôt dispersées, contrairement aux modèles avec interactions. D'un autre côté, ces derniers semblent plus biaisés, et sous-estiment les vraies valeurs, surtout les valeurs élevées.

Pour choisir les meilleurs modèles après avoir corrigé les valeurs de prédiction incorrectes, nous présentons le $MSEP$ au tableau 5.4. Nous pouvons voir que selon ce critère, le modèle le plus adéquat pour la grappe 1 est le modèle LinNT, pour la grappe 2 le modèle InterNT, pour les grappes 3 et 4, le modèle LinLog et pour la grappe 5 le modèle InterNT. Une décision liée strictement au $MSEP$ mène ainsi à une grande variété de choix de modèles parmi les grappes. Le choix de la transformation log, pour sa part, permet d'éviter les prédictions négatives, peu souhaitables dans un contexte d'érosion. Notons que pour la grappe 1, le deuxième meilleur modèle est le modèle LinLog et celui pour la grappe 2 est InterLog. Tel



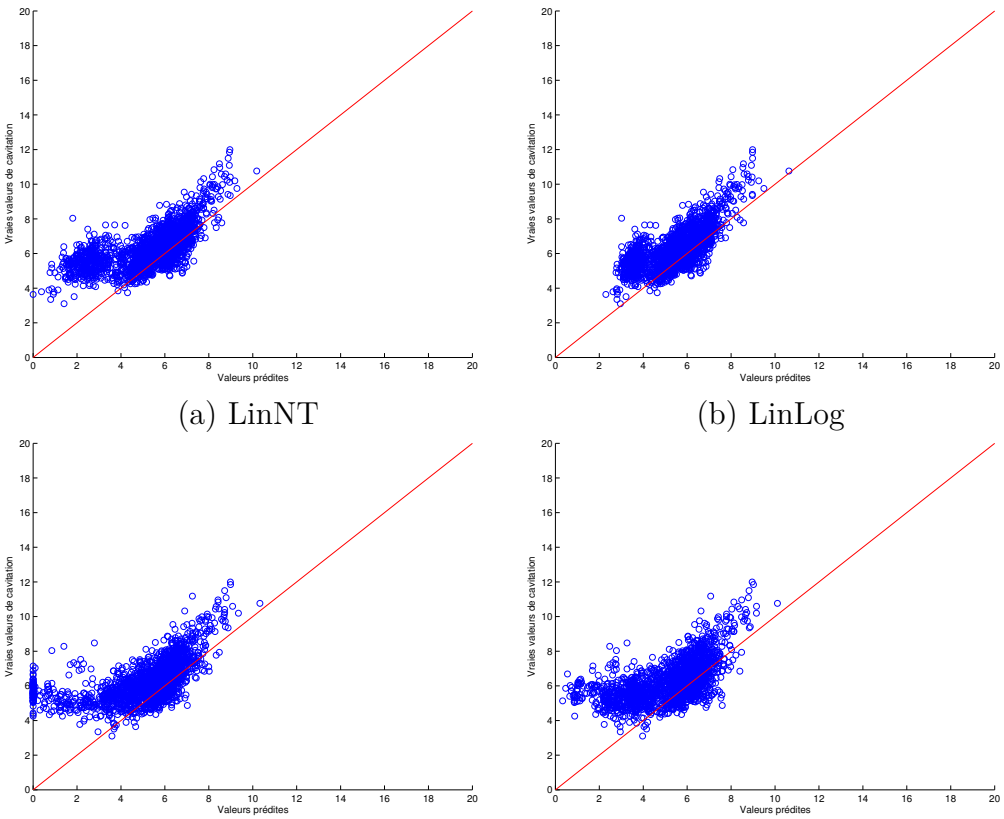
(a) LinNT (b) LinLog

(c) InterNT (d) InterLog

FIGURE 5.6. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation après correction pour la Grappe 2

que mentionné précédemment, pour la grappe 5, les modèles sans transformations produisent des prédictions plutôt dispersées (voir figure 5.9), contrairement aux modèles avec interactions. Le modèle InterLog est le troisième meilleur modèle selon le MSEP et semble raisonnable dans un souci d'uniformité. En effet, étant donné que la majorité des données de cavitation de cette grappe est contenue dans un intervalle étroit (94,91 % des données entre 3 et 8 kg/10 000 h), moins de précision pour la grappe 5 ne pose pas problème et nous considérons les prédictions satisfaisantes.

Somme toute, le tableau 5.4 montre que globalement et considérant les cinq grappes, les quatre modèles sont assez équivalents. Pour éviter les valeurs négatives et en se basant sur la minimisation du MSEP pour les cinq grappes, les modèles finaux choisis sont donc le modèle LinLog pour les grappes 1, 3 et 4 et le modèle InterLog pour les grappes 2 et 5. Ces résultats sont aussi cohérents avec la discussion sur la robustesse des prédicteurs à la section 5.1. Tous les modèles choisis présentent une relation assez linéaire entre les valeurs prédites et les vraies



(a) LinNT (b) LinLog

(c) InterNT (d) InterLog

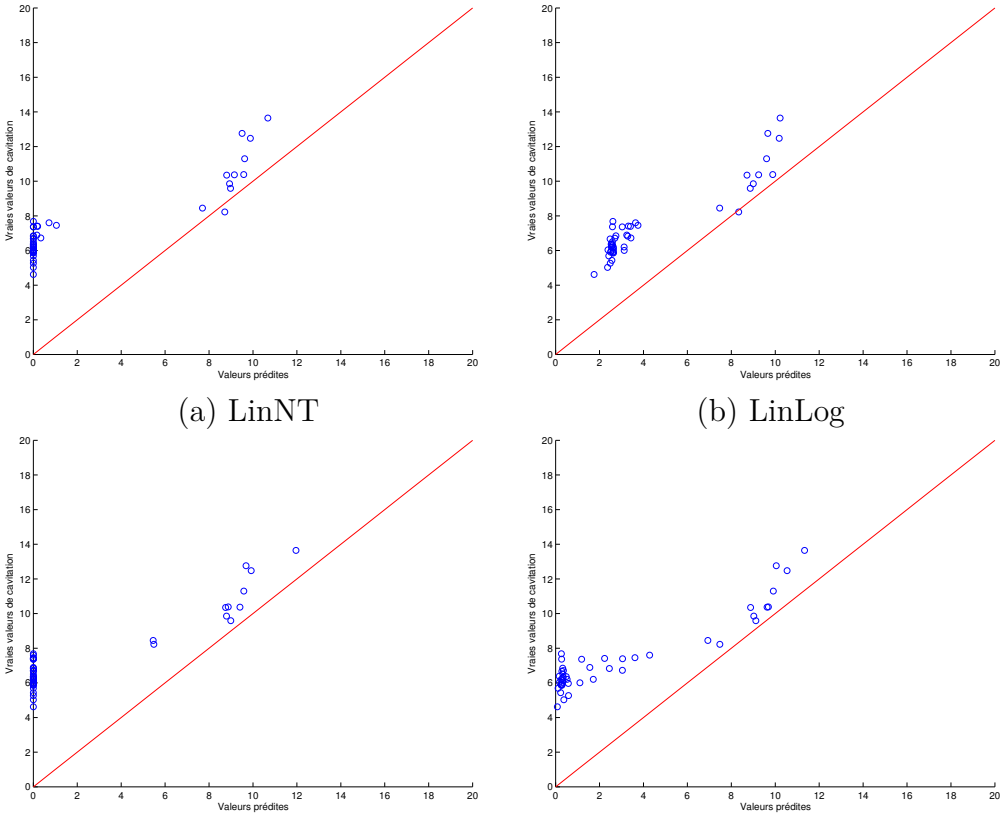
FIGURE 5.7. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation après correction pour la Grappe 3

valeurs de cavitation. Nous observons cependant un certain biais et il sera important de garder en tête que pour chaque grappe, le modèle statistique sous-estime les vraies valeurs. Étant donné la complexité du phénomène, nous considérons le modèle satisfaisant pour les objectifs actuels.

TABLEAU 5.4. *MSEP* par grappe pour l'année 2012 après correction

Grappe	LinNT	LinLog	InterNT	InterLog
1	1,403	1,576	1,593	1,719
2	3,202	2,485	1,585	1,617
3	2,644	1,446	3,460	2,838
4	31,283	10,931	32,304	23,848
5	0,897	0,982	0,886	0,979

La figure 5.10 montre la relation entre toutes les grappes pour le modèle final choisi. Nous pouvons voir en (a) que le modèle prédit bien les valeurs pour l'année 2011. Pour ce qui est de 2012, une certaine sous-estimation est observable pour l'ensemble des grappes. En particulier, les grappes 3 et 4 paraissent un peu



(a) LinNT (b) LinLog

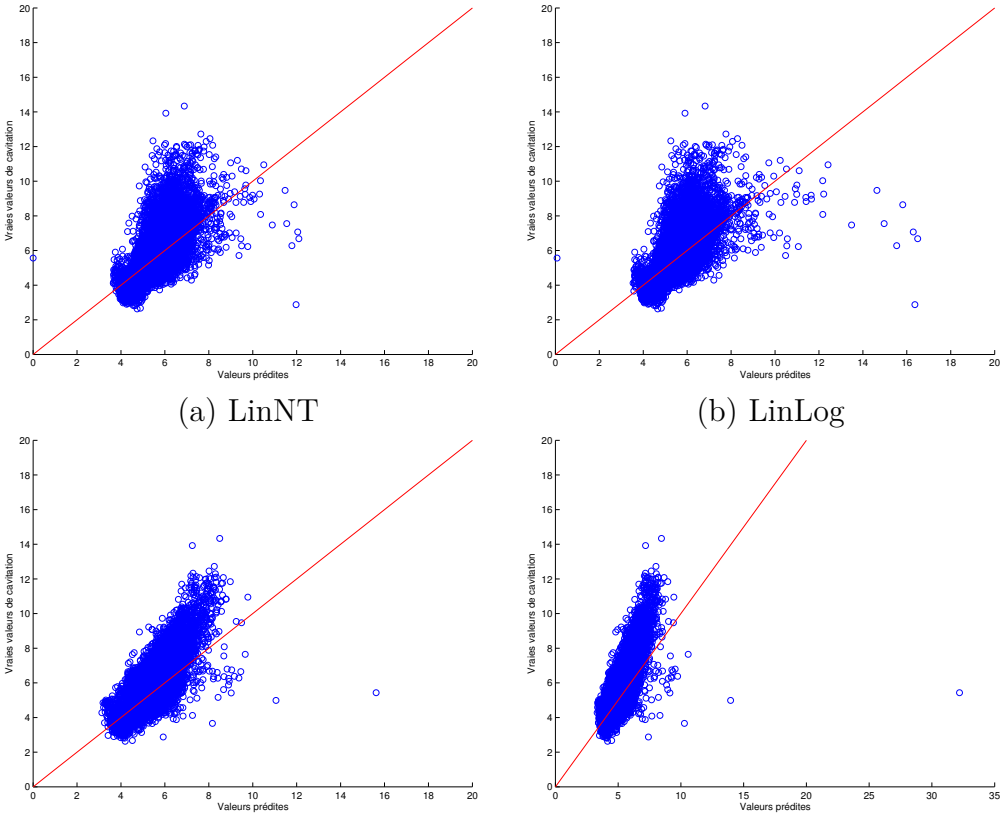
(c) InterNT (d) InterLog

FIGURE 5.8. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation après correction pour la Grappe 4

moins bien performer au niveau des prédictions. Ceci pourrait être dû au choix du nombre de grappes. Comme on peut le voir à la figure 5.11, la décision de conserver cinq grappes, mue principalement par un critère de nombre de données minimal par grappe, allait de pair avec une plus grande hétérogénéité de la grappe 3. La grappe 4, quant à elle, regroupe des modes opératoires en général peu fréquents et ne respectant pas l'ordre de priorité de démarrage, et il est peu surprenant que les prédictions pour cette grappe soient moins précises. Cependant, la capacité prédictive du modèle répond aux besoins actuels d'Hydro-Québec. Nous recommandons d'affiner le modèle dans le futur, en particulier au niveau de la sous-estimation : une fonction de perte autre que la fonction de perte quadratique pourrait être appropriée, et l'étude des données sur l'année 2013 pourra aussi donner une information plus pointue.

5.4. Érosion cumulative

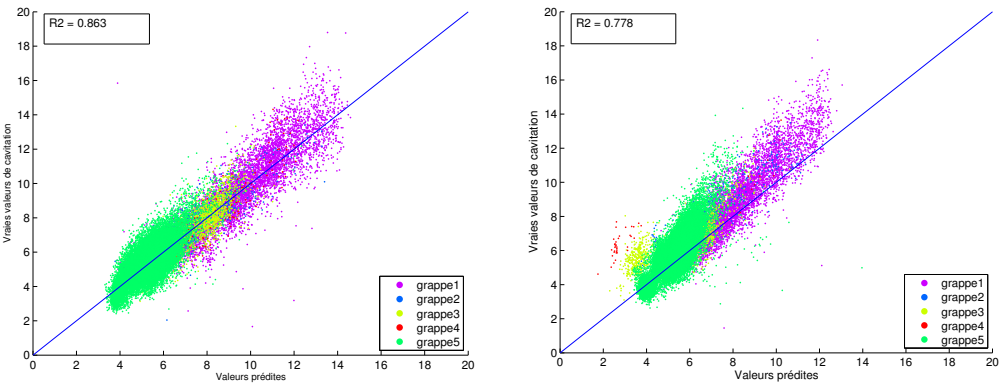
Une des informations importantes qu'Hydro-Québec souhaite retirer des modèles de prédiction est la valeur d'érosion cumulative. Cette dernière permet de



(a) LinNT (b) LinLog

(c) InterNT (d) InterLog

FIGURE 5.9. Comparaison entre les vraies valeurs et les valeurs prédites de cavitation après correction pour la Grappe 5



Année 2011 Année 2012

FIGURE 5.10. Modèle de prédiction général

mieux planifier les périodes de maintien et de réparation des turbines hydrauliques et de ce fait, de diminuer les pertes de revenus en arrêtant la turbine au moment opportun. Cette section s'attarde donc à comparer l'érosion cumulative calculée avec les valeurs d'érosion instantanée du Caviciel (vraies valeurs) pour le groupe

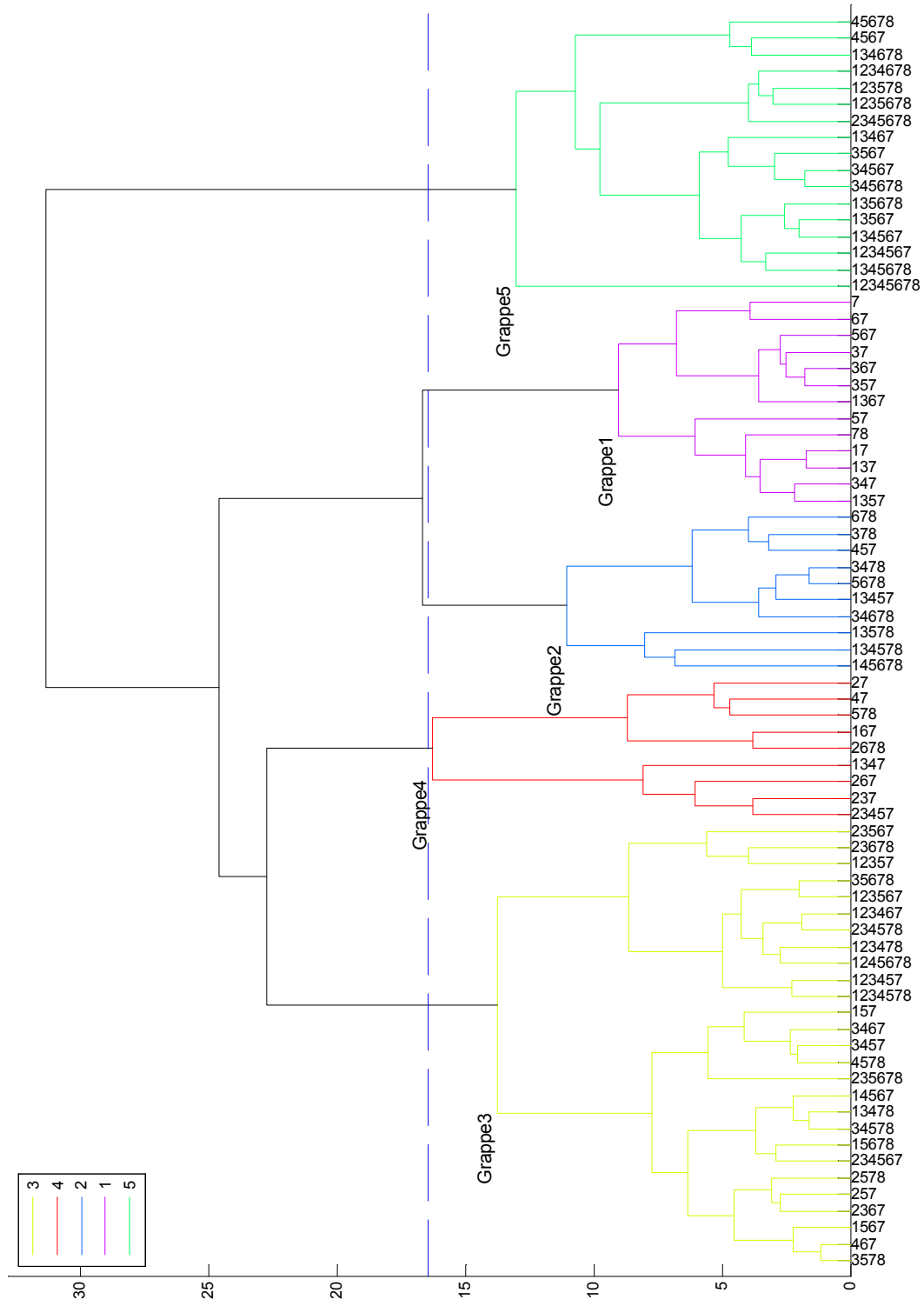


FIGURE 5.11. Dendrogramme

turbine-alternateur 7 et l'érosion cumulative telle que calculée avec les valeurs prédites du modèle statistique final basé sur le groupe 7. L'érosion cumulative quotidienne fournie par le Caviciel n'est pas utilisée dans la section suivante.

Comme les modèles statistiques ont été construits seulement sur la zone d'opération principale (zones de puissance 3 à 5), aucune valeur de prédiction n'est disponible pour les zones hors d'opération principale (zones 1, 2 et 6)(voir figure 1.7). Pour s'assurer que l'érosion cumulative calculée est bien représentative des années 2011 et 2012, nous utilisons les valeurs d'érosion instantanée du Caviciel de ces zones dans le vecteur de prédiction. Notons que ces zones d'opération seront étudiées plus en détail dans un autre volet du projet d'Hydro-Québec, mais ne font pas l'objet de ce mémoire. À titre de référence, nous présentons les statistiques descriptives par zone pour les années 2011 et 2012 aux tableaux 5.5 et 5.6. Notons que le groupe 7 semble avoir moins cavité en moyenne en 2012 qu'en 2011, et de façon moins variable, exception faite de la zone 4.

Tableau 5.5. Statistiques descriptives par zone pour l'année 2011

Zone	1	2	3	4	5	6
Moyenne	2,819	3,243	9,385	6,501	4,650	5,772
Écart-type	2,006	1,946	2,160	2,030	1,799	2,055
Minimum	0,000	0,541	1,663	2,045	2,417	3,819
Maximum	11,213	6,529	15,709	18,797	15,847	12,093
Médiane	2,743	3,205	9,255	5,957	4,170	4,841
Fréquence	158	10	1493	30218	3656	47

Tableau 5.6. Statistiques descriptives par zone pour l'année 2012

Zone	1	2	3	4	5	6
Moyenne	2,717	2,937	8,440	6,633	4,237	4,747
Écart-type	1,193	1,920	1,627	1,728	1,241	1,185
Minimum	0,097	1,579	2,878	1,456	2,831	3,433
Maximum	7,547	4,295	16,617	18,342	14,743	7,105
Médiane	2,691	2,937	8,180	6,214	3,919	4,515
Fréquence	154	2	943	32005	2290	10

Pour le calcul de l'érosion cumulative, un projet parallèle a été réalisé par l'équipe de Systèmes informationnels scientifiques de l'IREQ, qui consistait à étudier de façon approfondie la fonction de l'érosion cumulative quotidienne fournie par le Caviciel et à comparer différentes méthodes de calcul.

L'objectif étant ici de comparer l'érosion cumulative du groupe 7 calculée avec les vraies valeurs et celle calculée avec les valeurs prédites par le modèle

statistique, nous utilisons une méthode de calcul simple, qui permet de donner un ordre de grandeur suffisamment proche et qui est plus appropriée dans le contexte de nos données, soit la somme de Riemann. Cette méthode de calcul est appliquée sur l'érosion instantanée réelle et prédite. Cependant, la base de données de cavitation avec laquelle nous travaillons contient seulement des données si la turbine est en opération. La vitesse de rotation qui sert à déterminer si la turbine fonctionne n'est pas conservée par le logiciel. Dans le calcul de l'érosion, nous avons considéré que si deux mesures d'érosion instantanée consécutives étaient séparées par un intervalle de temps de plus de 40 minutes, la turbine avait subi un arrêt entre ces deux données et la valeur de cavitation instantanée correspondante était mise à 0.

Nous présentons l'érosion cumulative de juillet 2011 à décembre 2012, puisqu'une mesure par remplissage avait été prise en juillet 2011 et le système a été calibré à 5,8 kg d'érosion. Le calcul de l'érosion cumulative est donné par l'expression suivante :

$$\text{Érosion cumulative} = \sum_t \text{Érosion instantanée}_t \times \Delta t \times \mathbb{1}(\Delta t < 40)$$

où Δt est l'intervalle de temps entre t et $t-1$ et $\mathbb{1}(\Delta t < 40)$ est l'indicatrice qui permet d'identifier si la turbine a été arrêtée entre deux mesures d'érosion instantanée. Nous appelons cette méthode la somme de Riemann et la figure 5.12 montre les valeurs d'érosion cumulative par mois calculées avec les prédictions et avec les vraies valeurs d'érosion instantanée, suivant cette méthode. Ces deux calculs correspondent respectivement à la ligne rouge pour le modèle statistique et la ligne verte pour les vraies valeurs d'érosion instantanée. Nous pouvons voir au tableau 5.7 que le modèle statistique sous-estime en général les valeurs d'érosion cumulative calculée avec les érosions instantanées du Caviciel. Les mois dont les estimations du modèle statistique sont plus grandes que celles avec les valeurs du Caviciel sont en rouge. Somme toute, l'évaluation de l'érosion cumulative par le modèle statistique donne une estimation adéquate, quoiqu'en général un peu inférieure, de la valeur d'érosion cumulative calculée avec les vraies valeurs.

En définitive, ce chapitre a permis de sélectionner le meilleur modèle par grappe parmi les quatre modèles de régression. D'abord, la robustesse du choix des prédicteurs a été évaluée en utilisant la validation croisée sur les données de 2011. La sélection du modèle final par grappe est ensuite effectuée en se basant sur la validation sur un nouveau jeu de données, soient les données 2012. Les modèles choisis sont le modèle LinLog pour les grappes 1, 3 et 4 et le modèle InterLog pour les grappes 2 et 5. La dernière section, consacrée à la comparaison de l'érosion

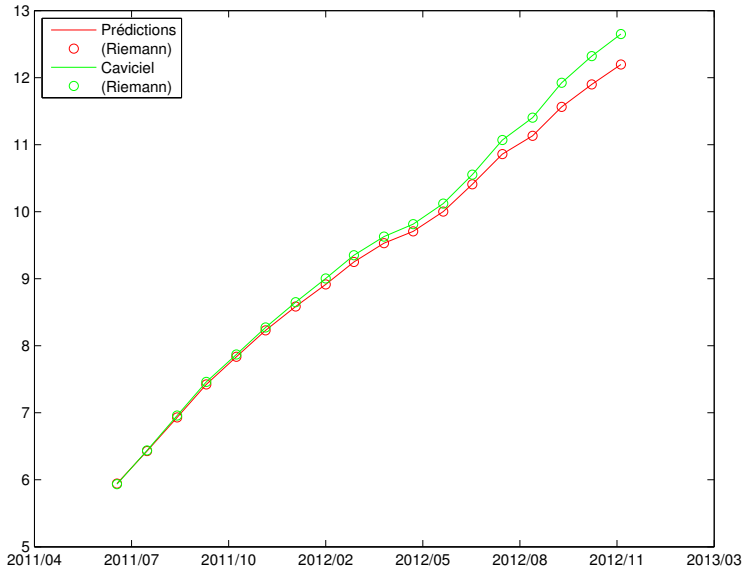


FIGURE 5.12. Comparaison de l'érosion cumulative avec les valeurs prédites par le modèle statistique et les vraies valeurs

TABLEAU 5.7. Érosion cumulative en kg par mois à partir de la mise à niveau de 5,8 kg en juillet 2011

	Prédiction Riemann	Vraies valeurs Riemann
19 au 31 juillet 2011	0,143	0,135
août-11	0,485	0,503
sept-11	0,502	0,522
oct-11	0,492	0,501
nov-11	0,413	0,409
déc-11	0,393	0,405
janv-12	0,356	0,376
févr-12	0,329	0,353
mars-12	0,335	0,347
avr-12	0,280	0,278
mai-12	0,177	0,186
juin-12	0,295	0,308
juil-12	0,406	0,430
août-12	0,451	0,518
sept-12	0,273	0,333
oct-12	0,433	0,520
nov-12	0,333	0,398
déc-12	0,299	0,330

cumulative calculée avec les valeurs instantanées prédites et celles du Caviciel, a mis en lumière que le modèle statistique présente une faiblesse en particulier au

niveau de la sous-estimation quasi systématique des valeurs d'érosion cumulative par mois.

CONCLUSION ET TRAVAUX FUTURS

La cavitation, phénomène hydraulique qui résulte de l'implosion de cavités de vapeur près des turbines, est un phénomène complexe et difficile à étudier sur le terrain. L'érosion de cavitation sur les matériaux est un des modes principaux de dégradation, d'où l'importance de l'étudier pour mieux planifier les périodes de maintenance sur les turbines hydrauliques. Les travaux de ce mémoire, réalisés avec le soutien de l'institut de recherche d'Hydro-Québec, portaient plus spécifiquement sur l'étude de l'érosion de cavitation du groupe turbine-alternateur 7 de la centrale W pendant la période de l'année 2011, et son lien avec les conditions opératoires des sept groupes partageant la même chambre d'équilibre.

L'objectif principal était de construire un modèle probabiliste pour prédire l'érosion de cavitation sur une turbine hydraulique de type Francis, en fonction des conditions d'opération. Quoique plusieurs études soient consacrées à l'érosion de cavitation par le biais de simulations numériques, peu d'auteurs avaient abordé ce sujet sous l'angle statistique. En particulier, nous avons souhaité améliorer la compréhension du phénomène qui, par sa complexité, pose de nombreux défis et ce, en développant un outil supplémentaire pour permettre de mieux planifier les périodes de maintenance et les coûts qui y sont liés.

Dans le premier chapitre, nous avons décrit la mise en oeuvre de la base de données et les résultats de l'analyse exploratoire de la cavitation du groupe 7. Plusieurs aspects de l'hétérogénéité du phénomène ont en effet été mis en évidence, qui semblent liés à une certaine saisonnalité ou aux modes opératoires de la centrale. En particulier, le contraste entre la période estivale et la période hivernale a été souligné. La première est plutôt caractérisée par des pics et des creux de cavitation du groupe 7 correspondant aux arrêts et départs simultanés des autres turbines, par une large étendue de cavitation et par l'influence de la hauteur de chute sur la cavitation. La période hivernale quant à elle, met en relief le lien entre la cavitation et la puissance active, peu d'arrêts-départs simultanés des autres groupes et une étendue de cavitation beaucoup plus étroite. La plurimodalité de la distribution de cavitation a aussi été mise en évidence.

Dans le chapitre 2, la nécessité de s'éloigner de la contrainte temporelle nous a amené à aborder le problème de l'hétérogénéité de la cavitation sous l'angle des modes opératoires, c'est-à-dire de la configuration marche-arrêt des huit groupes turbine-alternateur partageant la chambre d'équilibre. Les 76 modes opératoires observés ont été regroupés en cinq grappes selon des caractéristiques homogènes liées aux variables opératoires en utilisant le regroupement hiérarchique.

L'étude de la distribution de la cavitation dans chacune des grappes a été effectuée au chapitre 3. Plus précisément, l'algorithme Espérance-Maximisation a été utilisé pour estimer les paramètres de modèles de mélanges de lois de Burr de type XII, adéquats dans le cas de distribution plurimodale et permettant une grande flexibilité, en intégrant différents types d'asymétrie. Cette stratégie a permis de caractériser précisément la distribution de la cavitation par grappe de modes opératoires, et une première estimation de la cavitation par grappe est obtenue à l'aide de la moyenne théorique du mélange de lois de Burr sélectionné.

Dans le chapitre 4, nous affinons l'estimation de la cavitation. Pour chacune des grappes, quatre modèles de régression sont utilisés sur les données de l'année 2011, en utilisant la cavitation comme variable dépendante (log-transformée ou non) et les variables opératoires comme variables indépendantes (avec interactions ou pas). L'ajustement de ces modèles est adéquat, et en se basant sur le critère du R^2 ajusté, les meilleurs modèles par grappe sont les modèles avec interactions (cavitation non transformée pour les grappes 2 et 5 et avec transformation log pour les grappes 1, 3 et 4).

Le chapitre 5 présente la validation des quatre modèles de régression par grappe qui mène à la sélection des modèles finaux. La robustesse des prédicteurs est évaluée sur les données de l'année 2011, et la capacité prédictive des modèles sur les données de l'année 2012, à l'aide de l'erreur quadratique moyenne de prédiction. Le choix des modèles finaux se porte sur le modèle sans interactions avec transformation log pour les grappes 1, 3 et 4 et pour le modèle avec interactions et transformation log pour les grappes 2 et 5, en tenant compte de l'erreur quadratique moyenne, de la robustesse des prédicteurs et d'un souci de cohérence des modèles choisis pour chaque grappe.

En se basant sur des critères tels le MSEP, le R^2 et la comparaison avec l'érosion cumulative réelle, nous concluons que le modèle final permet de bien estimer l'érosion instantanée du groupe 7. Cependant, un certain biais est présent et le modèle a tendance à sous-estimer les vraies valeurs de cavitation, ce qui se réflète aussi dans l'estimation de l'érosion cumulative.

Cette étude comporte certaines limites, en particulier dans l'application du modèle statistique pour la prédiction. Les conditions opératoires étudiées sont

spécifiques au groupe ciblé, soit le groupe 7 de la centrale W. La généralisation du modèle à d'autres turbines ou d'autres conditions opératoires n'est pas recommandée. D'autre part, pour les besoins de cette étude, les observations ont été considérées indépendantes, alors qu'une dépendance temporelle est très probable étant donné la nature des données.

Les travaux développés dans ce mémoire ont été très bien reçus par l'équipe de recherche d'Hydro-Québec, qui a démontré un intérêt pour la poursuite des recherches. Un programme Matlab détaillant la méthodologie a été fourni et un second étudiant du Département de mathématiques et statistique de l'Université de Montréal est présentement impliqué dans la continuation du projet.

Dans ce mémoire, l'objectif était d'améliorer la compréhension du phénomène et, si possible, d'utiliser le modèle statistique pour prédire l'érosion de ce groupe turbine-alternateur spécifique sans utiliser le système vibratoire. Les travaux futurs prévoient notamment de reproduire la méthodologie sur les trois autres groupes turbine-alternateur munis du système Caviciel et de développer des modèles statistiques spécifiques à d'autres turbines. Les données sont aussi présentement collectées à une autre centrale dans des conditions opératoires différentes. Un autre volet des projets inclut l'étude de l'impact de l'opération hors conditions optimales sur l'érosion de cavitation. La validation du modèle actuel sur les données de 2013 et 2014 est aussi un objectif à court terme. Nous recommandons des travaux sur son amélioration, en particulier en ce qui a trait à la sous-estimation systématique des vraies valeurs d'érosion, par exemple, par le biais d'une fonction de perte plus adaptée. Il pourrait être intéressant par ailleurs d'utiliser un modèle de régression avec des erreurs Burr ou d'intégrer la notion des mélanges de régression par grappe. Une autre avenue pourrait être d'intégrer un modèle physique liant les conditions opératoires à la cavitation, travaux présentement en développement par le constructeur de turbine hydraulique.

Au final, les modèles statistiques développés dans ce mémoire sont un premier outil de prévision de l'érosion de cavitation, applicable sur un groupe turbine-alternateur spécifique avec des conditions opératoires propres à la centrale W. Il est intéressant de voir que l'angle statistique, bien que peu utilisé dans ce domaine, présente des résultats prometteurs afin de mieux caractériser la durée de vie résiduelle des turbines hydrauliques du parc hydroélectrique.

Bibliographie

Bodson-Clermont, P.-M. (2013). Étude de la distribution de l'érosion de cavitation sur un groupe turbine-alternateur et une période cible à l'aide de l'algorithme EM. Rapport technique, Université de Montréal. STT4000.

Bokhari, M. U. et N. Ahmad (2014). Incorporating burr type xii testing-efforts into software reliability growth modeling and actual data analysis with applications. *Journal of Software* **9**(6), 1389–1400.

Bourdon, P. (2000). *Détection vibratoire de l'érosion de cavitation des turbines Francis*. Thèse, École Polytechnique de Montréal. Canada.

Box, G. E. P. et D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B.* **26**, 211–252.

Burden, R. L. et J. D. Faires (2005). *Numerical analysis*. Prindle, Weber & Schmidt, Boston, Mass.

Burr, I. W. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics* **13**, 215–232.

Celeux, G. et J. Diebolt (1986). L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Revue de statistique appliquée* **34**(2), 35–52.

Chan, K. et J. Ledolter (1995). Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association* **90**(429), 242–252.

Dempster, A. P., N. M. Laird et D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological* **39**(1), 1–38.

Evin, G., J. Merleau et L. Perreault (2011). Two-component mixtures of normal, gamma, and gumbel distributions for hydrological applications. *Water Resources Research* **47**(8).

Ferreira, L. et D. B. Hitchcock (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics. Simulation and Computation* **38**(8-10), 1925–1949.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models : Modeling and Applications to Random Processes*. Springer.

Giroux, A. M., A. Merkhouf, L. Marcouiller et S. Cupillard (2011). Analyse numérique du comportement dynamique des groupes turbines-alternateurs - Étude préliminaire. Rapport technique IREQ-2011-0014, Institut de recherche d'Hydro-Québec.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, pp. 1137–1143. Morgan Kaufmann Publishers Inc.

Kuiper, F. K. et L. Fisher (1975). 391 : A Monte Carlo comparison of six clustering procedures. *Biometrics* **31**(3), 777–783.

Lafleur, F. (2011). Travaux volet détection de la cavitation : projet Préddit. Rapport technique No IREQ-2011-0067, Institut de recherche d'Hydro-Québec (IREQ).

Lafleur, F. (2012). Vibratory detection system of cavitation erosion : historic and algorithm validation. In *Proceedings of the 8th International Symposium on Cavitation, CAV2012*. Singapore : Research Publishing Services.

Liu, C. et D. B. Rubin (1994). The ecme algorithm : a simple extension of em and ecm with faster monotone convergence. *Biometrika* **81**(4), 633–648.

McLachlan, G. et T. Krishnan (1997). *The EM algorithm and extensions, Second Edition*. John Wiley & Sons.

Meng, X.-L. et D. B. Rubin (1993). Maximum likelihood estimation via the ecm algorithm : A general framework. *Biometrika* **80**(2), 267–278.

Mooi, E. et M. Sarstedt (2011). *A concise guide to market research : The process, data, and methods using IBM SPSS statistics*. Springer.

Moreau, D. (2012). Particularités et contraintes d'exploitation, Centrale W. Rapport technique, Hydro-Québec Production.

Morgan, B. J. T. et A. P. G. Ray (1995). Non-uniqueness and inversions in cluster analysis. *Applied statistics* **44**(1), 117–134.

Quian, Z. D., J. D. Yang et W. X. Huai (2007). Numerical simulation and analysis of pressure pulsation in hydraulic turbine with air admission. *Journal of Hydrodynamics* **19**(4), 467–472.

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics* **6**(2), 461–464.

Shao, Q. (2004). Notes on maximum likelihood estimation for the three-parameter Burr XII distribution. *Computational statistics & data analysis* **45**(3), 675–687.

Tadikamalla, P. R. (1980). A look at the Burr and related distributions. *International Statistical Review/Revue Internationale de Statistique* **48**(3), 337–344.

Tan, P.-N., M. Steinbach et V. Kumar (2006). *Introduction to data mining*, Volume 1. Pearson Addison Wesley Boston.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**(301), 236–244.

Annexe A

LISTE DES MODES OPÉRATOIRES

Le tableau ci-après présente la liste des modes opératoires observés conjointement avec une valeur de cavitation du groupe 7 dans les zones de puissance 3 à 5 pour l'année 2011. Nous rappelons que huit groupes turbine-alternateur partagent la chambre d'équilibre et que le mode opératoire liste seulement les groupes en fonctionnement, c'est-à-dire quand le débit est différent de 0 (voir section 2.1.2). Le nombre de groupes en marche définit le nombre de groupes turbine-alternateur en fonctionnement dans la chambre d'équilibre. L'identification de la grappe fait référence au regroupement des modes opératoires (nommé grappe) développé dans le chapitre 2 (voir figure 2.3).

Tableau A.1. Liste des modes opératoires observés pour l'année 2011

Grappe	**Mode opératoire**	**Fréquence relative(%)**	**Fréquence**	**Nb groupes Marche**
1	1357	574	1,62	4
1	1367	100	0,28	4
1	357	1749	4,95	3
1	137	401	1,13	3
1	567	169	0,48	3
1	367	134	0,38	3
1	347	61	0,17	3
1	37	1554	4,39	2
1	17	329	0,93	2
1	78	303	0,86	2
1	57	268	0,76	2
1	67	96	0,27	2
1	7	1718	4,86	1
2	145678	56	0,16	6

Suite sur la page suivante

Tableau A.1 – suite de la page précédente

Grappe	Mode opératoire	Fréquence relative(%)	Fréquence	Nb groupes Marche
2	134578	39	0,11	6
2	34678	310	0,88	5
2	13457	124	0,35	5
2	13578	6	0,02	5
2	5678	741	2,10	4
2	3478	277	0,78	4
2	457	104	0,29	3
2	378	78	0,22	3
2	678	3	0,01	3
3	1234578	160	0,45	7
3	1245678	86	0,24	7
3	123478	71	0,20	6
3	234578	63	0,18	6
3	123567	31	0,09	6
3	123467	20	0,06	6
3	235678	8	0,02	6
3	234567	6	0,02	6
3	123457	3	0,01	6
3	34578	378	1,07	5
3	35678	317	0,90	5
3	15678	24	0,07	5
3	13478	22	0,06	5
3	12357	16	0,05	5
3	14567	7	0,02	5
3	23678	3	0,01	5
3	23567	1	$< 0{,}001$	5
3	4578	264	0,75	4
3	3457	263	0,74	4
3	3467	186	0,53	4
3	3578	39	0,11	4
3	1567	27	0,08	4
3	2367	12	0,03	4
3	2578	5	0,01	4
3	257	15	0,04	3
3	157	11	0,03	3
3	467	11	0,03	3

Suite sur la page suivante

Tableau A.1 – suite de la page précédente

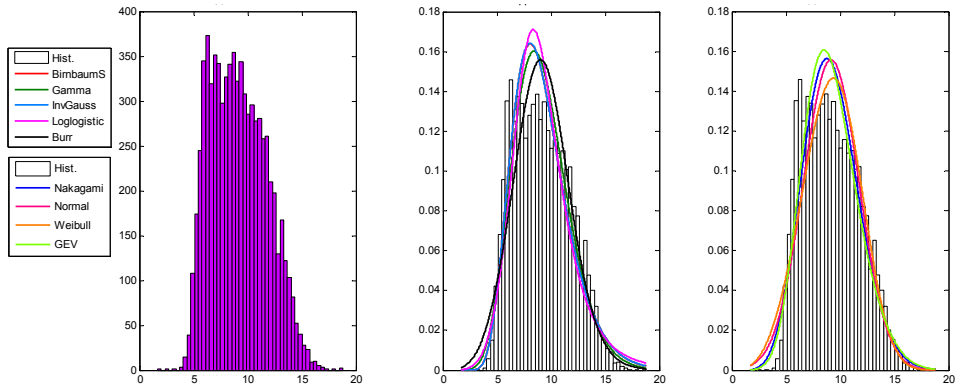
Grappe	Mode opératoire	Fréquence relative(%)	Fréquence	Nb groupes Marche
4	23457	3	0,01	5
4	2678	25	0,07	4
4	1347	1	< 0,001	4
4	578	147	0,42	3
4	167	20	0,06	3
4	267	2	0,01	3
4	237	1	0,00	3
4	27	15	0,04	2
4	47	9	0,03	2
5	12345678	11428	32,31	8
5	1345678	3277	9,27	7
5	2345678	685	1,94	7
5	1234567	411	1,16	7
5	1235678	343	0,97	7
5	1234678	95	0,27	7
5	345678	1875	5,30	6
5	134567	1455	4,11	6
5	135678	300	0,85	6
5	134678	78	0,22	6
5	123578	65	0,18	6
5	13567	1132	3,20	5
5	34567	963	2,72	5
5	13467	35	0,10	5
5	45678	32	0,09	5
5	3567	1675	4,74	4
5	4567	52	0,15	4
Total		35367	100	

Annexe B

LISTE DES DISTRIBUTIONS ENVISAGÉES POUR LA CAVITATION PAR GRAPPE

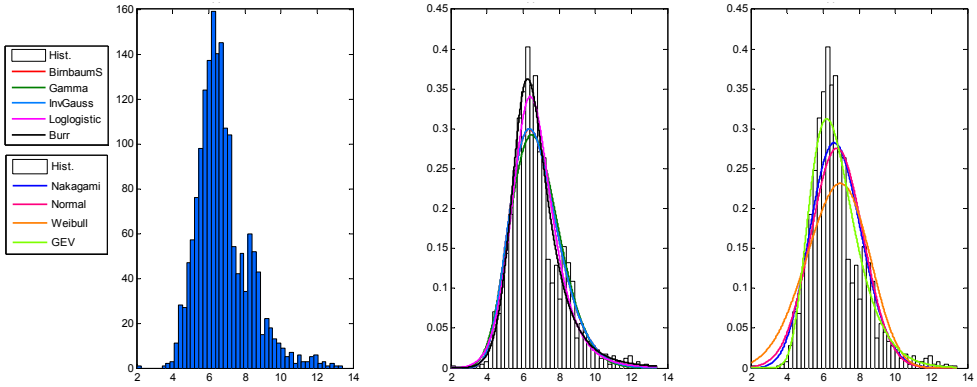
Au chapitre 3, plusieurs distributions ont été ajustées initialement aux données de cavitation par grappe (voir section 3.3). Nous nous sommes limités aux distributions disponibles dans Matlab 2013b, avec un support adéquat pour nos données. Les neuf distributions ajustées sont donc : Birnbaum-Sauders, gamma, inverse gaussienne, log-logisitique, Burr type XII à 3 paramètres, Nakagami, normale, Weibull et valeurs extrêmes généralisée (GEV). Le détail de ces distributions, de leur support et de leurs premiers moments est décrit dans le tableau suivant. De plus, les figures B.1 à B.5 présentent graphiquement les ajustements des neufs lois retenues pour les cinq grappes. Le BIC pour chacun des ajustements est présenté au tableau 3.1. Selon ce critère, la loi de Burr est celle qui s'adapte le mieux aux données avec un ajustement unimodal.

Nom	Param.	Support	Densité	Espérance	Variance
Birnbaum- Saunders	$\beta > 0$ $\gamma \geq 0$	$x > 0$	$f(x \mid \beta, \gamma) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\sqrt{\frac{x}{\beta}} - \sqrt{\frac{\beta}{x}}\right)^2}{2\gamma^2}\right)\left(\frac{\sqrt{\frac{x}{\beta}} - \sqrt{\frac{\beta}{x}}}{2\gamma x}\right)$	$\beta(1 + \frac{\gamma^2)}{2}$	$(\gamma\beta)^2(1 + \frac{5\gamma^2}{4})$
Gamma	$a > 0$ $b > 0$	$x > 0$	$f(x\|a, b) = \frac{a}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}}$	ab	ab^2
Inverse- Gaussienne	$\mu > 0$ $\lambda > 0$	$x > 0$	$f(x\|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(\frac{-\lambda}{2\mu^2 x}(x - \mu)^2\right)$	μ	$\frac{\mu^3}{\lambda}$
Log- Logistique	$\mu > 0$ $\sigma > 0$	$x \geq 0$	$f(x\|\mu, \sigma) = \frac{1}{\sigma x} \frac{e^z}{(1+e^z)^2}$ avec $z = \frac{log(x) - \mu}{\sigma}$	$\exp(\mu + \log\Gamma(1 + \sigma)$ $+ \log\Gamma(1 - \sigma))$ si $\sigma < 1$	$\exp(2\mu + \log\Gamma(1 + 2\sigma)$ $+ \log\Gamma(1 - 2\sigma)) - E(X)^2$ si $\sigma < \frac{1}{2}$
Burr	$\alpha > 0$ $c > 0$ $k > 0$	$x > 0$	$f(x\|\alpha, c, k) = \frac{kc}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}\left(1 + \left(\frac{x}{\alpha}\right)^c\right)^{-(k+1)}$	$k\alpha\ \mathrm{B}(\mathrm{k} - \frac{1}{\mathrm{c}}; \frac{1}{\mathrm{c}} + 1)$	$E(X^2) - E(X)^2$ où $E(X^2) = k\alpha^2\ \mathrm{B}(\mathrm{k} - \frac{2}{\mathrm{c}}; \frac{2}{\mathrm{c}} + 1)$
Nakagami	$\mu > 0$ $\omega > 0$	$x > 0$	$f(x\|\mu, \omega) = 2\left(\frac{\mu}{\omega}\right)^\mu \frac{1}{\Gamma(\mu)} x^{(2\mu-1)} \exp\left(-\frac{\mu}{\omega}x^2\right)$	$\frac{\Gamma(\mu+\frac{1}{2})}{\Gamma(\mu)}\left(\frac{\omega}{\mu}\right)^{1/2}$	$\omega\left(1 - \frac{1}{\mu}\left(\frac{\Gamma(\mu+\frac{1}{2})}{\Gamma(\mu)}\right)^2\right)$
Normale	μ $\sigma \geq 0$	$\mathbb{R}$	$f(x\|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$	μ	σ^2
Weibull	$a > 0$ $b > 0$	$x \geq 0$	$f(x\|a, b) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} \exp\left(-\left(\frac{x}{a}\right)^b\right)$	$a\Gamma\left(1 + \frac{1}{b}\right)$	$a^2\Gamma\left(1 + \frac{2}{b}\right) - E(X)^2$
GEV (Generalized Exteme Value)	$k \in \mathbb{R}$ $\sigma \geq 0$ $\mu \in \mathbb{R}$	Si $k = 0$, $x \in \mathbb{R}$ Si $k > 0$, $x \geq \mu - \frac{\sigma}{k}$ Si $k < 0$, $x < \mu - \frac{\sigma}{k}$	Si $k = 0$, $f(x\|0, \mu, \sigma) = \frac{1}{\sigma} \exp\left[-\exp\left(-\frac{(x-\mu)}{\sigma}\right) - \frac{(x-\mu)}{\sigma}\right]$ Si $k \neq 0$, $f(x\|k, \mu, \sigma) = \frac{1}{\sigma} \exp\left[-\left(1 + k\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{k}}\right] *$ $\left[1 + k\frac{(x-\mu)}{\sigma}\right]^{-1-\frac{1}{k}}$	Si $k = 0$, $\mu + \sigma\gamma$ Si $k \neq 0$ et $k < 1$, $\mu + \sigma\frac{\Gamma(1-k)-1}{k}$ Si $k \geq 1$, ∞ avec γ la constante d'Euler	Si $k = 0$, $\sigma^2\frac{\pi^2}{6}$ Si $k \neq 0$ et $k < 1/2$, $\sigma^2(g_2 - g_1^2)/k^2$ Si $k \geq 1/2$, ∞ avec $g_p = \Gamma(1 - pk)$



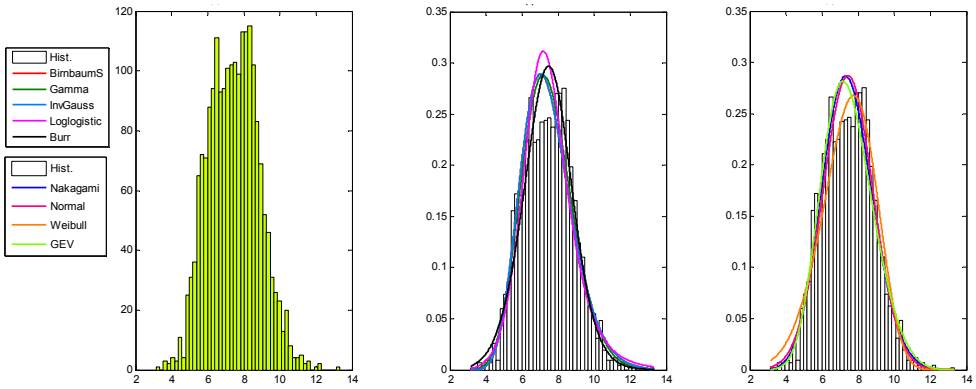
(a) Histogramme sans ajustement (b) Ajustement de 5 distr. (c) Ajustement de 4 distr.

FIGURE B.1. Ajustement de 9 distributions unimodales aux données de cavitation de la grappe 1



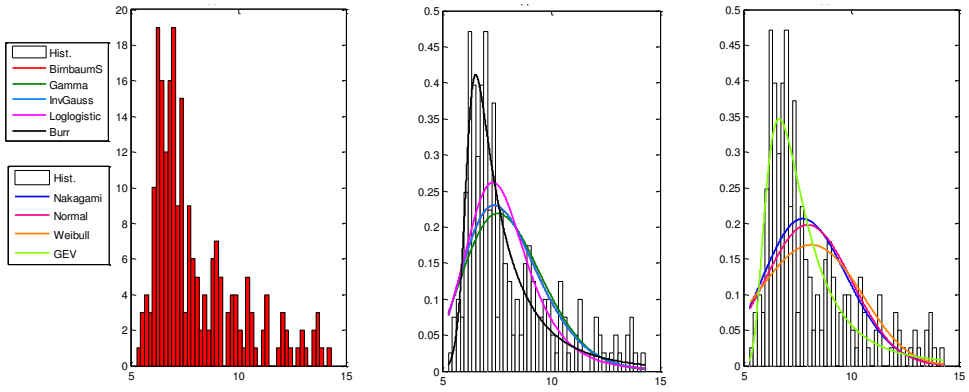
(a) Histogramme sans ajustement (b) Ajustement pour 5 distr. (c) Ajustement pour 4 distr.

FIGURE B.2. Ajustement de 9 distributions unimodales aux données de cavitation de la grappe 2



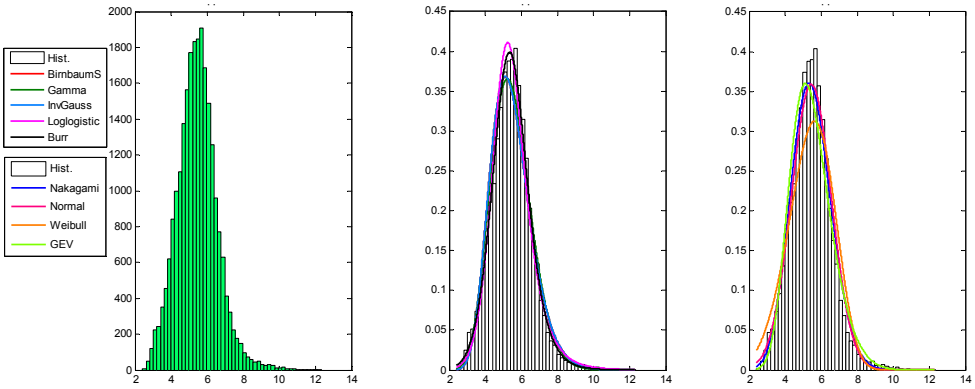
(a) Histogramme sans ajustement (b) Ajustement pour 5 distr. (c) Ajustement pour 4 distr.

FIGURE B.3. Ajustement de 9 distributions unimodales aux données de cavitation de la grappe 3



(a) Histogramme sans ajustement (b) Ajustement pour 5 distr. (c) Ajustement pour 4 distr.

FIGURE B.4. Ajustement de 9 distributions unimodales aux données de cavitation de la grappe 4



(a) Histogramme sans ajustement (b) Ajustement pour 5 distr. (c) Ajustement pour 4 distr.

FIGURE B.5. Ajustement de 9 distributions unimodales aux données de cavitation de la grappe 5

Annexe C

MODÈLES DE RÉGRESSION ET ANALYSE DES RÉSIDUS POUR LES GRAPPES 2 À 5

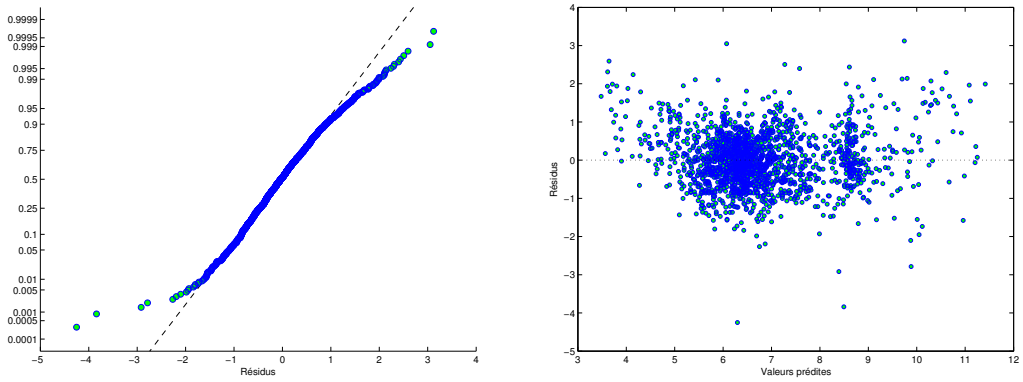
La section 4.2 du chapitre 4 présentait les quatre modèles pour la grappe 1 soient les modèles avec et sans interactions, avec une transformation log ou non de la variable dépendante cavitation (LinNT, LinLog, InterNT, InterLog). Dans cet annexe, nous présentons les quatre modèles pour les grappes 2 à 5, avec l'analyse des résidus. Pour connaître l'appartenance d'un mode opératoire à une des cinq grappes, le lecteur se référera à l'annexe A.

C.1. Grappe 2

C.1.1. Modèle linéaire non transformé (LinNT)

Tableau C.1. Détails du modèle de régression linéaire sans transformation (LinNT) pour la grappe 2

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-77,581	9,033	1,936E-17
Courant	-0,232	0,100	2,081E-02
Debit	-0,204	0,020	7,620E-23
Amont	-0,289	0,049	3,440E-09
Mvar	-0,010	0,001	1,608E-11
MW	0,144	0,017	1,409E-17
PuissEff	0,378	0,013	1,360E-155
Tension	0,458	0,197	2,015E-02



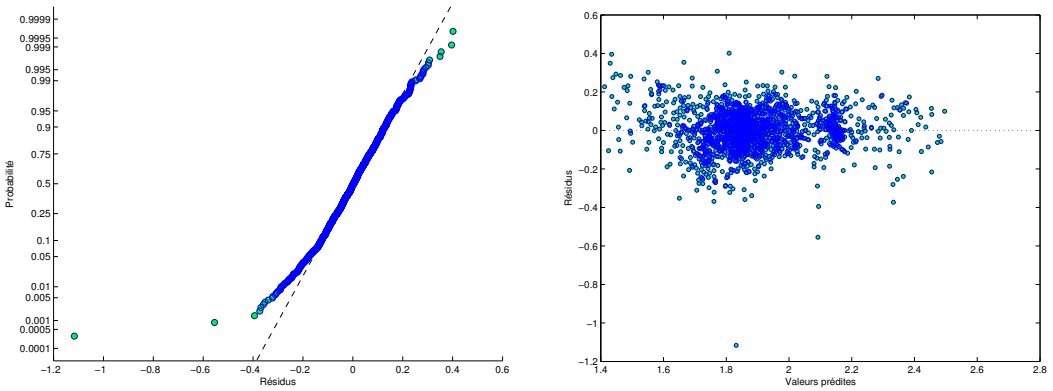
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.1. Analyse des résidus pour la régression linéaire de la grappe 2 après sélection des variables (modèle LinNT)

C.1.2. Modèle linéaire avec transformation log (LinLog)

TABLEAU C.2. Détails du modèle de régression linéaire avec transformation log (LinLog) pour la grappe 2

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-8,073	1,294	5,535E-10
Courant	-0,047	0,014	5,862E-04
Debit	-0,035	0,003	1,398E-29
Amont	-0,032	0,007	1,181E-05
Mvar	-0,001	0,000	3,652E-09
MW	0,025	0,002	2,827E-23
PuissEff	0,047	0,002	2,239E-116



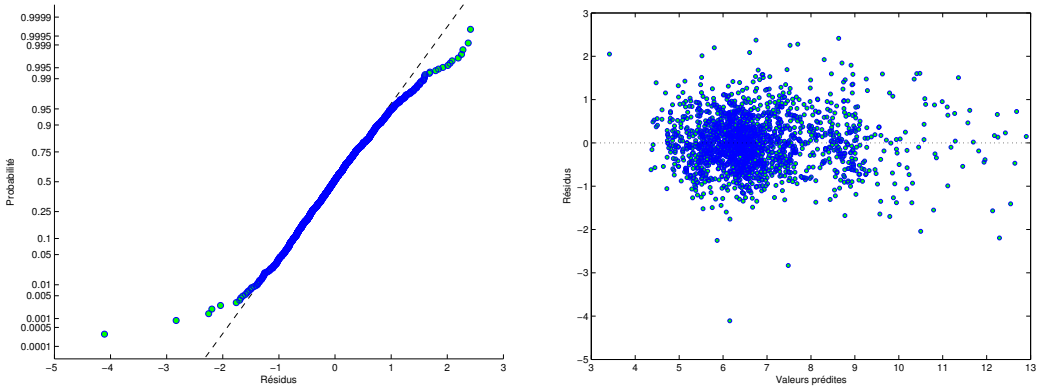
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.2. Analyse des résidus pour la régression linéaire de la grappe 2 après sélection des variables (modèle LinLog)

C.1.3. Modèle avec interactions sans transformation (InterNT)

TABLEAU C.3. Détails du modèle de régression avec interactions sans transformation (InterNT) pour la grappe 2

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-9334,000	1596,000	5,930E-09
Courant	-52,379	14,258	2,466E-04
Debit	34,930	10,761	1,193E-03
Amont	51,514	9,238	2,855E-08
Aval	78,291	16,914	3,955E-06
Ouverture	13,652	14,255	3,384E-01
Mvar	-0,951	0,433	2,822E-02
MW	-31,655	6,805	3,544E-06
PuissEff	22,367	4,402	4,162E-07
Tension	-24,032	8,754	6,110E-03
Courant :PuissEff	0,147	0,040	2,168E-04
Debit :Amont	-0,150	0,060	1,258E-02
Debit :Aval	0,142	0,042	6,686E-04
Debit :Ouverture	0,058	0,009	7,193E-10
Debit :Mvar	0,005	0,001	2,893E-06
Debit :Tension	-1,305	0,168	1,445E-14
Amont :Aval	-0,436	0,092	2,463E-06
Amont :Ouverture	0,224	0,103	2,983E-02
Amont :Mvar	0,007	0,003	8,480E-03
Amont :MW	0,089	0,034	9,266E-03
Amont :PuissEff	-0,121	0,025	1,179E-06
Aval :Ouverture	-0,362	0,088	4,354E-05
Aval :MW	-0,059	0,025	1,577E-02
Aval :Tension	0,580	0,252	2,164E-02
Ouverture :MW	-0,041	0,006	7,646E-11
Ouverture :PuissEff	-0,131	0,019	1,231E-11
Ouverture :Tension	0,411	0,114	3,264E-04
Mvar :MW	-0,004	0,001	4,167E-06
Mvar :Tension	-0,015	0,003	2,019E-07
MW :PuissEff	0,022	0,006	8,469E-05
MW :Tension	0,949	0,126	6,687E-14



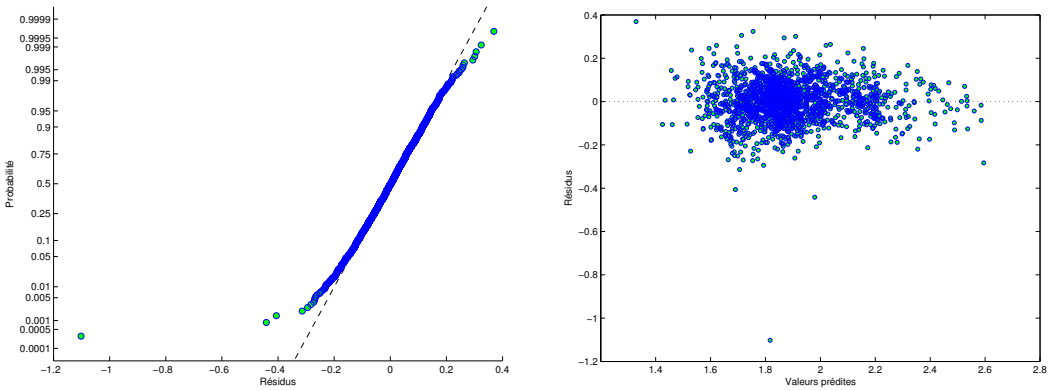
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.3. Analyse des résidus pour la régression linéaire de la grappe 2 après sélection des variables (modèle InterNT)

C.1.4. Modèle avec interactions avec transformation log (InterLog)

TABLEAU C.4. Détails du modèle de régression avec interactions avec transformation log (InterLog) pour la grappe 2

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-1600,000	219,650	4,916E-13
Courant	-17,248	4,145	3,329E-05
Debit	6,384	1,653	1,163E-04
Amont	9,011	1,239	5,407E-13
Aval	11,481	2,657	1,642E-05
Ouverture	0,105	2,212	9,620E-01
Mvar	-0,096	0,038	1,176E-02
MW	-4,861	1,005	1,446E-06
PuissEff	4,124	0,656	3,992E-10
Tension	-8,174	1,512	7,283E-08
Courant :Aval	0,117	0,044	7,550E-03
Courant :PuissEff	0,037	0,008	8,235E-06
Debit :Amont	-0,040	0,010	5,710E-05
Debit :Aval	0,025	0,008	1,076E-03
Debit :Ouverture	0,010	0,001	1,827E-12
Debit :Mvar	0,001	0,000	2,191E-04
Debit :PuissEff	0,005	0,002	1,200E-03
Debit :Tension	-0,215	0,028	4,051E-14
Amont :Aval	-0,072	0,014	4,305E-07
Amont :Ouverture	0,063	0,018	4,998E-04
Amont :MW	0,021	0,005	6,995E-05
Amont :PuissEff	-0,022	0,004	6,978E-10
Aval :Ouverture	-0,075	0,016	5,158E-06
Aval :MW	-0,013	0,003	6,171E-05
Aval :Tension	0,231	0,042	4,110E-08
Ouverture :MW	-0,007	0,001	3,594E-13
Ouverture :PuissEff	-0,026	0,005	8,178E-09
Ouverture :Tension	0,053	0,018	3,888E-03
Mvar :MW	-0,001	0,000	2,150E-04
Mvar :PuissEff	0,000	0,000	9,863E-04
Mvar :Tension	-0,002	0,000	3,937E-05
MW :Tension	0,158	0,021	2,148E-13



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

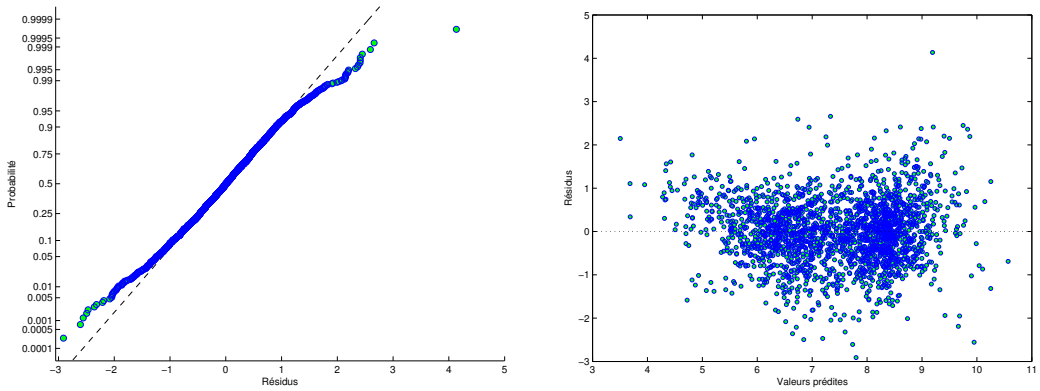
FIGURE C.4. Analyse des résidus pour la régression linéaire de la grappe 2 après sélection des variables (modèle InterLog)

C.2. GRAPPE 3

C.2.1. Modèle linéaire non transformé (LinNT)

TABLEAU C.5. Détails du modèle de régression linéaire sans transformation (LinNT) pour la grappe 3

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-92,855	10,593	3,830E-18
Courant	1,568	0,218	7,850E-13
Debit	-0,372	0,045	1,314E-16
Aval	0,165	0,056	3,425E-03
MW	0,173	0,035	7,222E-07
PuissEff	0,272	0,025	1,319E-26
Tension	0,900	0,168	1,010E-07



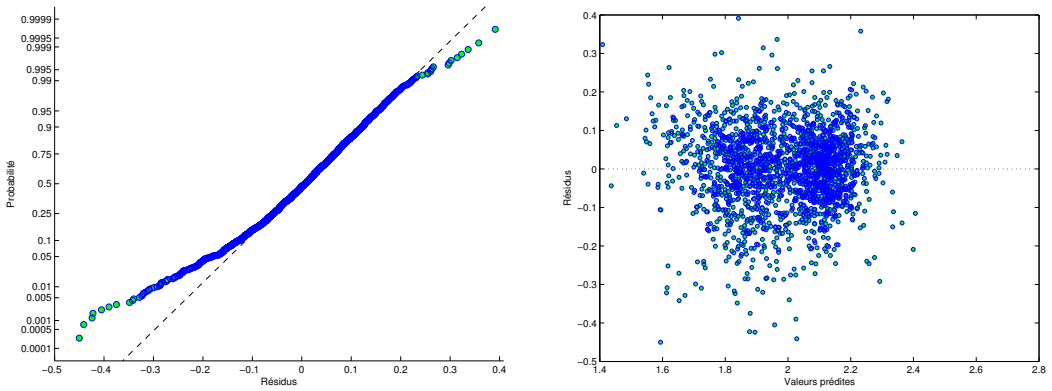
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.5. Analyse des résidus pour la régression linéaire de la grappe 3 après sélection des variables (modèle LinNT)

C.2.2. Modèle linéaire avec transformation log (LinLog)

TABLEAU C.6. Détails du modèle de régression linéaire avec transformation log (LinLog) pour la grappe 3

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-10,673	1,447	2,347E-13
Courant	0,206	0,030	7,517E-12
Debit	-0,075	0,007	3,457E-24
Aval	0,032	0,008	3,074E-05
Ouverture	0,019	0,008	1,194E-02
MW	0,038	0,005	4,487E-14
PuissEff	0,033	0,003	1,407E-21
Tension	0,129	0,023	3,554E-08



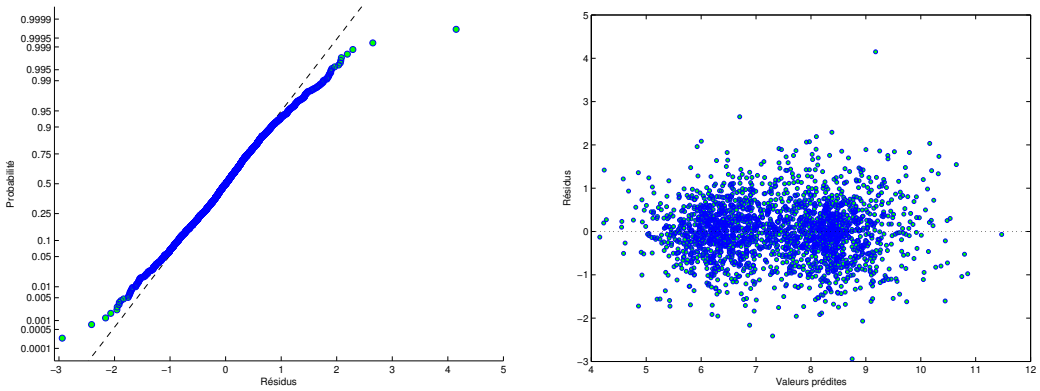
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.6. Analyse des résidus pour la régression linéaire de la grappe 3 après sélection des variables (modèle LinLog)

C.2.3. Modèle avec interactions sans transformation (InterNT)

TABLEAU C.7. Détails du modèle de régression avec interactions sans transformation (InterNT) pour la grappe 3

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-10795,000	2489,000	1,515E-05
Courant	-635,520	118,660	9,480E-08
Debit	18,408	6,789	6,757E-03
Amont	56,262	13,996	6,036E-05
Aval	69,554	12,377	2,177E-08
Ouverture	52,049	10,024	2,283E-07
Mvar	1,508	0,530	4,455E-03
MW	0,168	0,055	2,283E-03
PuissEff	23,964	6,325	1,557E-04
Tension	-9,367	8,335	2,612E-01
Courant :Amont	3,971	0,666	2,962E-09
Courant :Aval	1,338	0,237	2,013E-08
Courant :Mvar	0,047	0,007	3,306E-10
Courant :Tension	-7,438	1,297	1,112E-08
Debit :Amont	-0,222	0,038	4,106E-09
Debit :Mvar	-0,002	0,000	2,057E-05
Debit :MW	-0,003	0,000	2,713E-11
Debit :PuissEff	0,044	0,006	1,081E-13
Debit :Tension	0,422	0,082	2,498E-07
Amont :Aval	-0,155	0,068	2,329E-02
Amont :Mvar	-0,009	0,003	2,302E-03
Amont :PuissEff	-0,124	0,034	2,858E-04
Aval :Ouverture	-0,270	0,060	6,772E-06
Aval :PuissEff	-0,113	0,011	9,288E-23
Ouverture :PuissEff	-0,118	0,023	4,625E-07



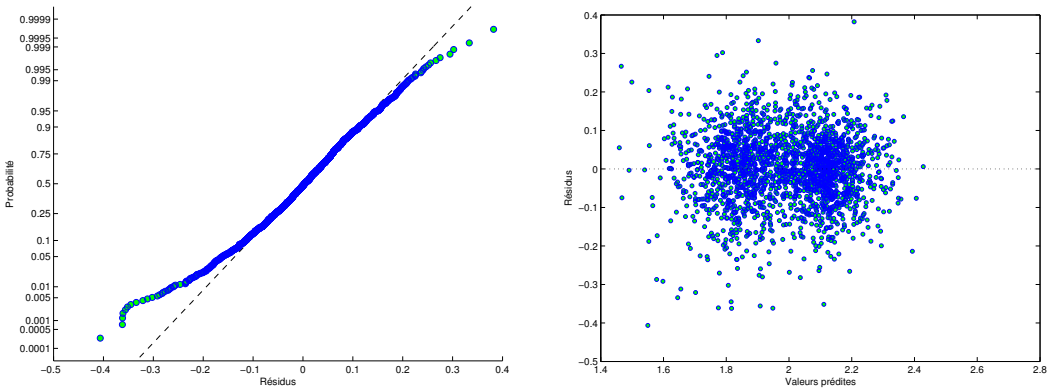
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.7. Analyse des résidus pour la régression linéaire de la grappe 3 après sélection des variables (modèle InterNT)

C.2.4. Modèle avec interactions avec transformation log (InterLog)

TABLEAU C.8. Détails du modèle de régression avec interactions avec transformation log (InterLog) pour la grappe 3

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-2451,500	379,620	1,327E-10
Courant	-83,097	13,887	2,571E-09
Debit	-1,318	0,302	1,367E-05
Amont	11,652	2,130	5,051E-08
Aval	7,318	1,875	9,831E-05
Ouverture	17,984	2,603	6,512E-12
Mvar	0,231	0,091	1,114E-02
MW	0,408	0,074	3,561E-08
PuissEff	5,482	0,918	2,790E-09
Tension	34,211	9,832	5,130E-04
Courant :Amont	0,512	0,078	7,865E-11
Courant :Aval	0,259	0,037	3,369E-12
Courant :Mvar	0,006	0,001	8,891E-16
Courant :MW	-0,006	0,001	5,607E-09
Courant :Tension	-0,969	0,177	5,240E-08
Debit :PuissEff	0,002	0,001	5,033E-03
Debit :Tension	0,048	0,011	9,078E-06
Amont :Aval	-0,020	0,010	4,251E-02
Amont :Ouverture	-0,103	0,015	7,069E-12
Amont :Mvar	-0,001	0,001	8,077E-03
Amont :PuissEff	-0,026	0,005	2,660E-07
Aval :MW	-0,011	0,002	3,619E-07
Aval :PuissEff	-0,006	0,002	2,006E-04
Aval :Tension	-0,133	0,066	4,493E-02
Ouverture :Mvar	-0,001	0,000	5,537E-08
PuissEff :Tension	-0,080	0,021	1,557E-04



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

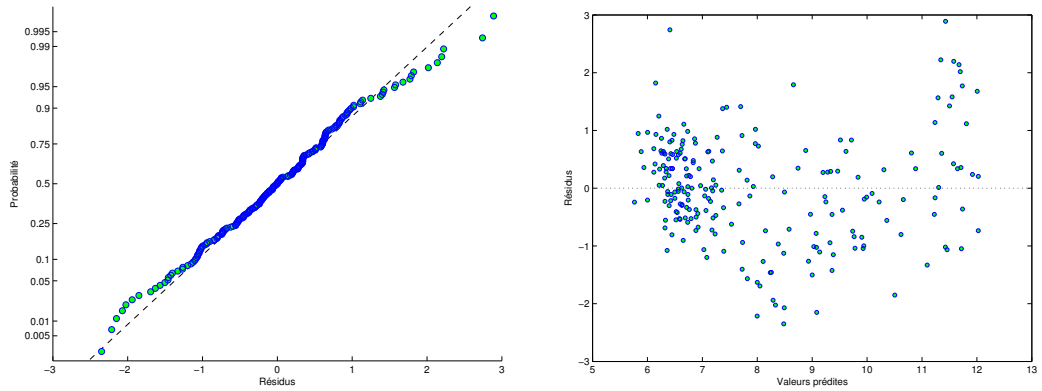
FIGURE C.8. Analyse des résidus pour la régression linéaire de la grappe 3 après sélection des variables (modèle InterLog)

C.3. GRAPPE 4

C.3.1. Modèle linéaire non transformé (LinNT)

TABLEAU C.9. Détails du modèle de régression linéaire sans transformation (LinNT) pour la grappe 4

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-731,280	60,548	5,315E-26
Courant	2,647	0,807	1,216E-03
Debit	-0,200	0,044	1,001E-05
Amont	3,531	0,377	9,527E-18
Mvar	-0,015	0,007	3,198E-02
PuissEff	0,249	0,047	2,810E-07
Tension	3,423	1,013	8,660E-04



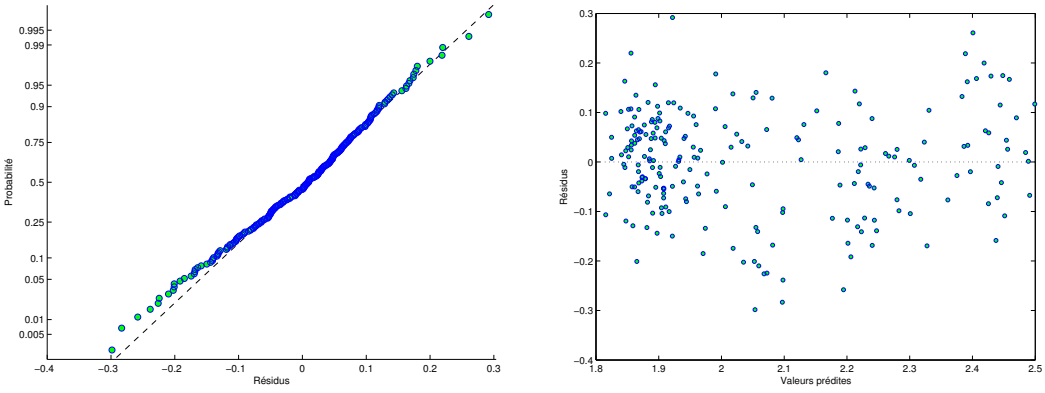
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.9. Analyse des résidus pour la régression linéaire de la grappe 4 après sélection des variables (modèle LinNT)

C.3.2. Modèle linéaire avec transformation log (LinLog)

TABLEAU C.10. Détails du modèle de régression linéaire avec transformation log (LinLog) pour la grappe 4

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-79,945	6,823	6,521E-25
Courant	0,102	0,032	1,416E-03
Debit	-0,013	0,002	2,688E-10
Amont	0,394	0,042	1,593E-17
PuissEff	0,042	0,003	4,847E-34

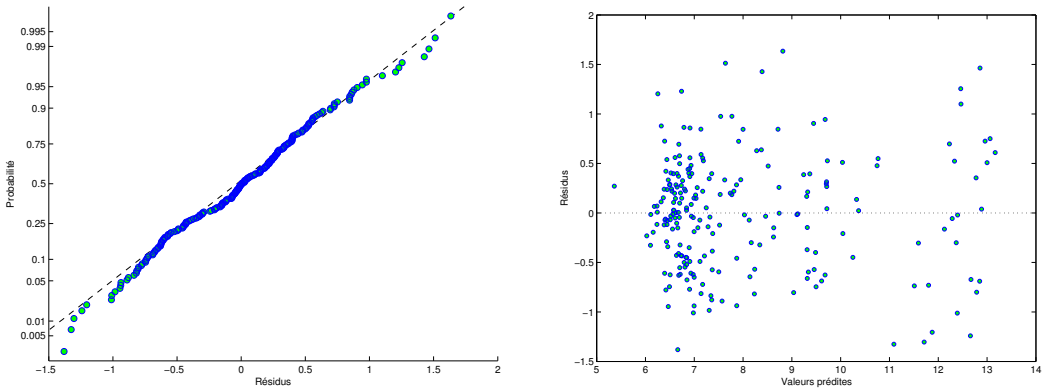


(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.10. Analyse des résidus pour la régression linéaire de la grappe 4 après sélection des variables (modèle LinLog)

C.3.3. Modèle avec interactions sans transformation (InterNT)

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-201080,000	46401,000	2,348E-05
Courant	-26,895	369,560	9,421E-01
Debit	24,992	14,630	8,918E-02
Amont	1099,300	256,980	2,957E-05
Aval	1637,500	318,810	6,772E-07
Ouverture	17,488	7,016	1,351E-02
Mvar	2,104	0,637	1,141E-03
MW	-27,773	23,531	2,393E-01
PuissEff	396,760	98,632	8,227E-05
Tension	817,010	262,160	2,107E-03
Courant :Debit	-0,816	0,365	2,675E-02
Courant :Amont	6,658	2,515	8,773E-03
Courant :Aval	-6,371	2,824	2,520E-02
Courant :MW	0,504	0,251	4,552E-02
Courant :PuissEff	-2,413	0,802	2,970E-03
Debit :Mvar	0,056	0,010	2,166E-07
Debit :PuissEff	0,170	0,042	8,377E-05
Debit :Tension	-5,382	1,364	1,111E-04
Amont :Aval	-9,163	1,750	4,251E-07
Amont :MW	-0,300	0,128	2,074E-02
Amont :PuissEff	-2,130	0,540	1,112E-04
Aval :Ouverture	-0,515	0,207	1,358E-02
Aval :Mvar	-0,062	0,018	9,273E-04
Aval :MW	0,482	0,121	9,192E-05
Aval :PuissEff	-0,203	0,095	3,329E-02
Mvar :MW	-0,044	0,008	9,290E-08
MW :Tension	4,010	1,015	1,088E-04
PuissEff :Tension	-2,078	0,673	2,300E-03



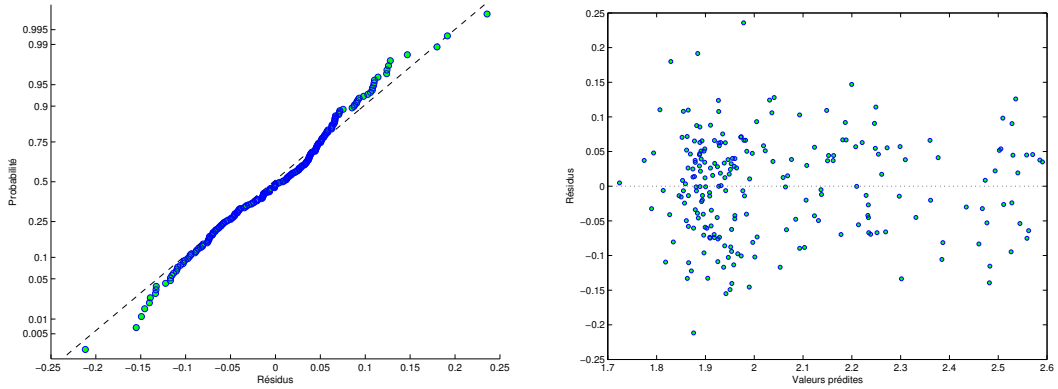
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.11. Analyse des résidus pour la régression linéaire de la grappe 4 après sélection des variables (modèle InterNT)

C.3.4. Modèle avec interactions avec transformation log (InterLog)

TABLEAU C.11. Détails du modèle de régression avec interactions avec transformation log (InterLog) pour la grappe 4

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	2668,000	2210,500	2,289E-01
Courant	-297,830	92,566	1,511E-03
Debit	-46,156	19,545	1,917E-02
Amont	-18,538	13,201	1,618E-01
Aval	106,180	30,319	5,703E-04
Ouverture	1,883	0,877	3,297E-02
Mvar	0,196	0,067	3,883E-03
MW	45,244	15,310	3,505E-03
PuissEff	1,848	0,856	3,213E-02
Tension	-404,930	126,070	1,539E-03
Courant :Amont	1,869	0,570	1,225E-03
Courant :PuissEff	-0,076	0,038	4,656E-02
Debit :Amont	0,292	0,113	1,079E-02
Debit :Mvar	0,005	0,001	2,482E-05
Debit :Tension	-0,337	0,116	3,983E-03
Amont :Aval	-0,624	0,174	4,273E-04
Amont :MW	-0,297	0,089	1,012E-03
Amont :Tension	2,703	0,768	5,379E-04
Aval :Ouverture	-0,056	0,026	3,237E-02
Aval :Mvar	-0,006	0,002	3,453E-03
Aval :MW	0,022	0,008	5,593E-03
Mvar :MW	-0,004	0,001	1,409E-05
MW :PuissEff	0,006	0,002	8,885E-03
MW :Tension	0,268	0,087	2,218E-03
PuissEff :Tension	-0,185	0,065	4,610E-03



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

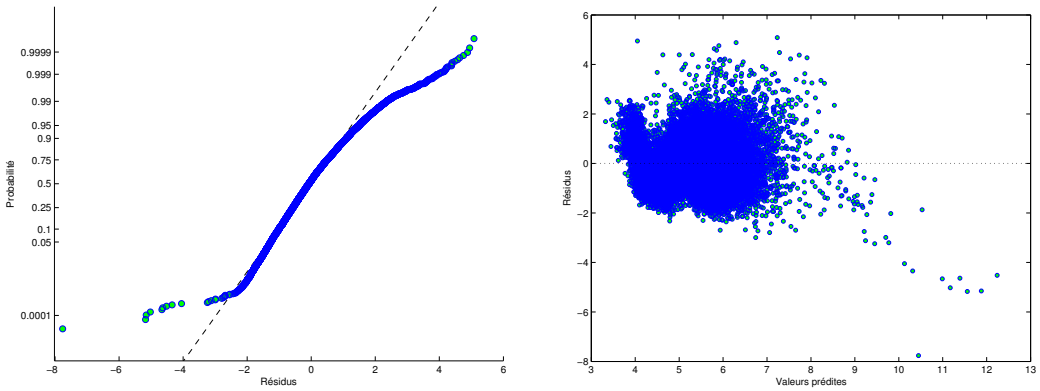
FIGURE C.12. Analyse des résidus pour la régression linéaire de la grappe 4 après sélection des variables (modèle InterLog)

C.4. GRAPPE 5

C.4.1. Modèle linéaire non transformé (LinNT)

TABLEAU C.12. Détails du modèle de régression linéaire sans transformation (LinNT) pour la grappe 5

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-90,681	2,501	2,450E-280
Courant	0,827	0,075	4,989E-28
Debit	0,249	0,009	1,439E-168
Amont	0,757	0,021	1,240E-280
Aval	-0,726	0,022	4,798E-241
Mvar	-0,012	0,001	5,256E-78
MW	-0,306	0,008	1,306E-287
PuissEff	-0,023	0,004	5,591E-08
Tension	1,453	0,094	9,576E-54



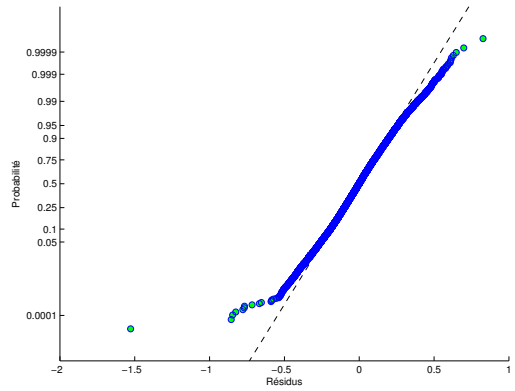
(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.13. Analyse des résidus pour la régression linéaire de la grappe 5 après sélection des variables (modèle LinNT)

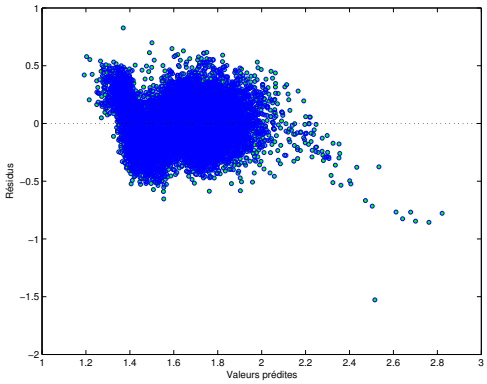
C.4.2. Modèle linéaire avec transformation log (LinLog)

TABLEAU C.13. Détails du modèle de régression linéaire avec transformation log (LinLog) pour la grappe 5

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	-11,842	0,458	1,205E-145
Courant	0,099	0,014	7,315E-13
Debit	0,032	0,002	2,543E-82
Amont	0,116	0,004	1,918E-200
Aval	-0,110	0,004	2,334E-168
Mvar	-0,002	0,000	2,180E-59
MW	-0,042	0,002	2,025E-168
PuissEff	-0,005	0,001	7,738E-13
Tension	0,222	0,017	4,084E-38



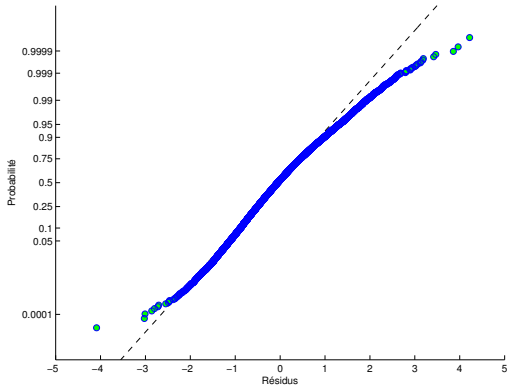
(a) Graphique de probabilités normales



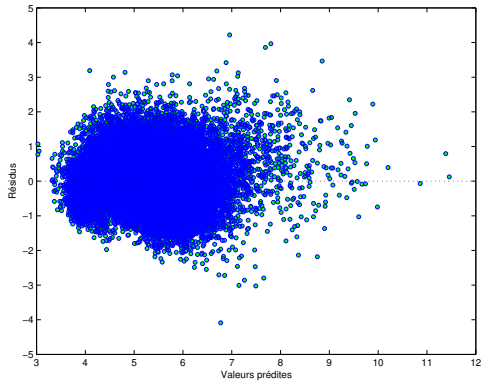
(b) Résidus vs valeurs prédites

FIGURE C.14. Analyse des résidus pour la régression linéaire de la grappe 5 après sélection des variables (modèle LinLog)

C.4.3. Modèle avec interactions sans transformation (InterNT)



(a) Graphique de probabilités normales



(b) Résidus vs valeurs prédites

FIGURE C.15. Analyse des résidus pour la régression linéaire de la grappe 5 après sélection des variables (modèle InterNT)

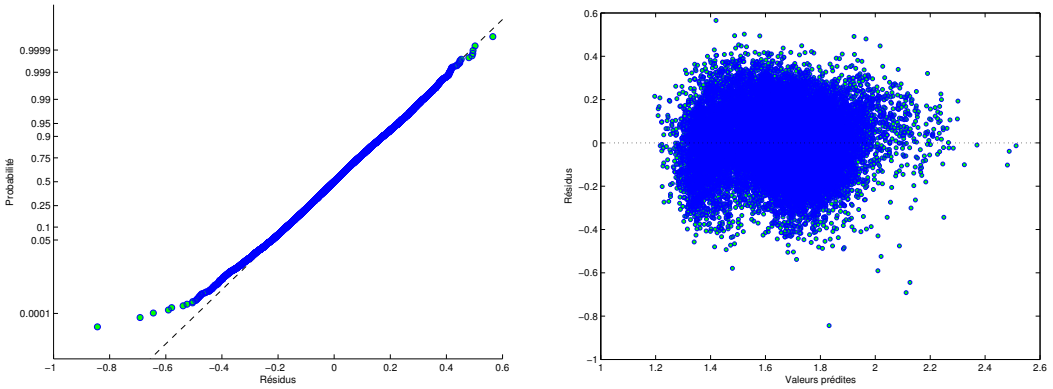
TABLEAU C.14. Détails du modèle de régression avec interactions sans transformation (InterNT) pour la grappe 5

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	5836,200	315,220	5,363E-76
Courant	-12,765	10,287	2,147E-01
Debit	-13,644	1,561	2,406E-18
Amont	-40,205	2,092	1,023E-81
Aval	46,383	2,100	5,232E-107
Ouverture	-26,844	2,095	1,900E-37
Mvar	3,031	0,170	5,355E-71
MW	12,543	1,191	7,203E-26
PuissEff	-8,682	0,730	1,587E-32
Tension	-125,300	14,224	1,348E-18
Courant :Debit	0,100	0,010	2,072E-22
Courant :Aval	-0,139	0,058	1,699E-02
Courant :Ouverture	-0,188	0,040	2,873E-06
Courant :Mvar	-0,023	0,005	5,742E-07
Courant :PuissEff	-0,050	0,017	2,621E-03
Courant :Tension	1,856	0,497	1,895E-04
Debit :Amont	-0,029	0,009	6,595E-04
Debit :Aval	-0,015	0,005	1,137E-03
Debit :Ouverture	0,092	0,003	4,020E-213
Debit :Mvar	-0,006	0,001	4,223E-15
Debit :PuissEff	0,032	0,002	2,237E-47
Amont :Aval	0,014	0,006	1,833E-02
Amont :Ouverture	0,158	0,012	3,132E-38
Amont :Mvar	-0,022	0,001	6,133E-72
Amont :MW	0,036	0,006	4,011E-10
Amont :PuissEff	0,059	0,004	6,768E-41
Amont :Tension	0,303	0,072	2,922E-05
Aval :Ouverture	-0,157	0,011	9,403E-44
Aval :Mvar	0,017	0,001	6,525E-45
Aval :PuissEff	-0,093	0,005	2,569E-89
Ouverture :Mvar	-0,005	0,001	9,030E-08
Ouverture :MW	-0,079	0,003	7,185E-134
Ouverture :Tension	0,623	0,083	7,326E-14
Mvar :MW	0,007	0,001	1,962E-32
Mvar :Tension	0,005	0,002	2,759E-02
MW :PuissEff	-0,028	0,002	1,892E-37
MW :Tension	-0,239	0,034	3,122E-12
PuissEff :Tension	0,212	0,025	4,037E-17

C.4.4. Modèle avec interactions avec transformation log (InterLog)

TABLEAU C.15. Détails du modèle de régression avec interactions avec transformation log (InterLog) pour la grappe 5

	Estimations	Err. Standard	Valeur-p
Ordonnée à l'origine	901,060	56,065	8,114E-58
Courant	4,718	2,296	3,989E-02
Debit	-1,843	0,283	7,062E-11
Amont	-5,234	0,448	2,151E-31
Aval	5,354	0,520	8,780E-25
Ouverture	-5,690	0,452	2,835E-36
Mvar	0,497	0,030	3,061E-61
MW	1,653	0,187	1,111E-18
PuissEff	-1,720	0,150	3,164E-30
Tension	-12,118	2,514	1,444E-06
Courant :Amont	-0,072	0,011	6,260E-11
Courant :Aval	0,051	0,011	3,717E-06
Courant :Ouverture	-0,039	0,007	6,807E-08
Courant :Mvar	-0,004	0,001	3,156E-06
Courant :MW	0,018	0,002	6,202E-33
Courant :Tension	0,255	0,090	4,602E-03
Debit :Aval	-0,005	0,001	3,660E-06
Debit :Ouverture	0,021	0,001	6,865E-158
Debit :Mvar	-0,001	0,000	1,388E-07
Debit :PuissEff	0,003	0,001	2,539E-08
Debit :Tension	-0,042	0,011	1,106E-04
Amont :Aval	0,002	0,001	2,750E-02
Amont :Ouverture	0,029	0,002	7,237E-35
Amont :Mvar	-0,003	0,000	9,928E-53
Amont :MW	0,005	0,001	1,141E-08
Amont :PuissEff	0,010	0,001	3,941E-35
Amont :Tension	-0,067	0,023	3,266E-03
Aval :Ouverture	-0,031	0,003	1,497E-28
Aval :Mvar	0,002	0,000	1,256E-22
Aval :PuissEff	-0,014	0,001	8,282E-57
Aval :Tension	0,145	0,026	1,803E-08
Ouverture :Mvar	-0,001	0,000	8,115E-05
Ouverture :MW	-0,017	0,001	3,358E-113
Ouverture :PuissEff	0,003	0,001	2,874E-04
Ouverture :Tension	0,108	0,028	1,447E-04
Mvar :MW	0,001	0,000	8,962E-25
MW :PuissEff	-0,004	0,000	6,975E-22
PuissEff :Tension	0,049	0,006	2,339E-18



(a) Graphique de probabilités normales (b) Résidus vs valeurs prédites

FIGURE C.16. Analyse des résidus pour la régression linéaire de la grappe 5 après sélection des variables (modèle InterLog)

Annexe D

NUAGES DE POINTS PAR GRAPPE

À l'annexe C, nous avons présenté les différents modèles de régression pour les grappes 2 à 5. Pour mieux comprendre les relations unissant chacune des variables indépendantes, nous présentons ici les nuages de points des interactions de ces neufs variables pour les grappes 2 à 5. L'information pour la grappe 1 est disponible à la section 4.3 du chapitre 4 (voir figure 4.11).

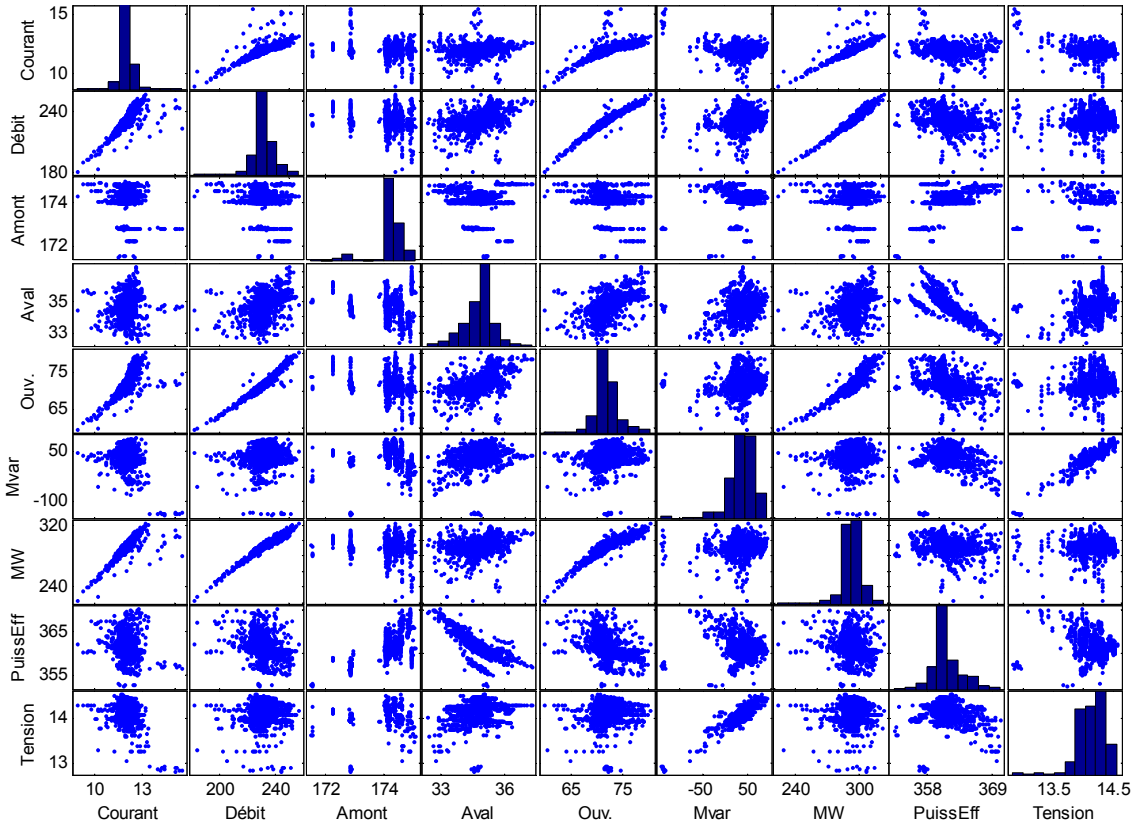


FIGURE D.1. Nuages de points des interactions entre les neuf variables indépendantes pour la grappe 2

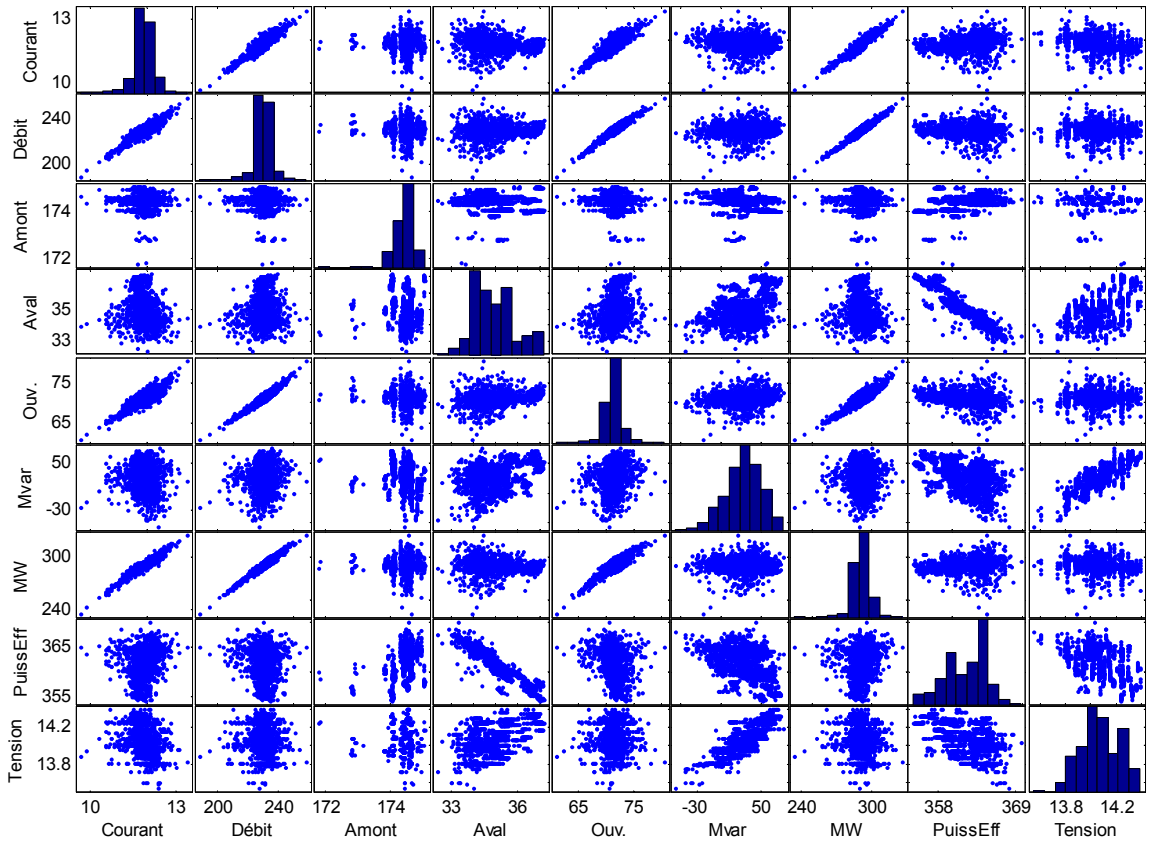


FIGURE D.2. Nuages de points des interactions entre les neuf variables indépendantes pour la grappe 3

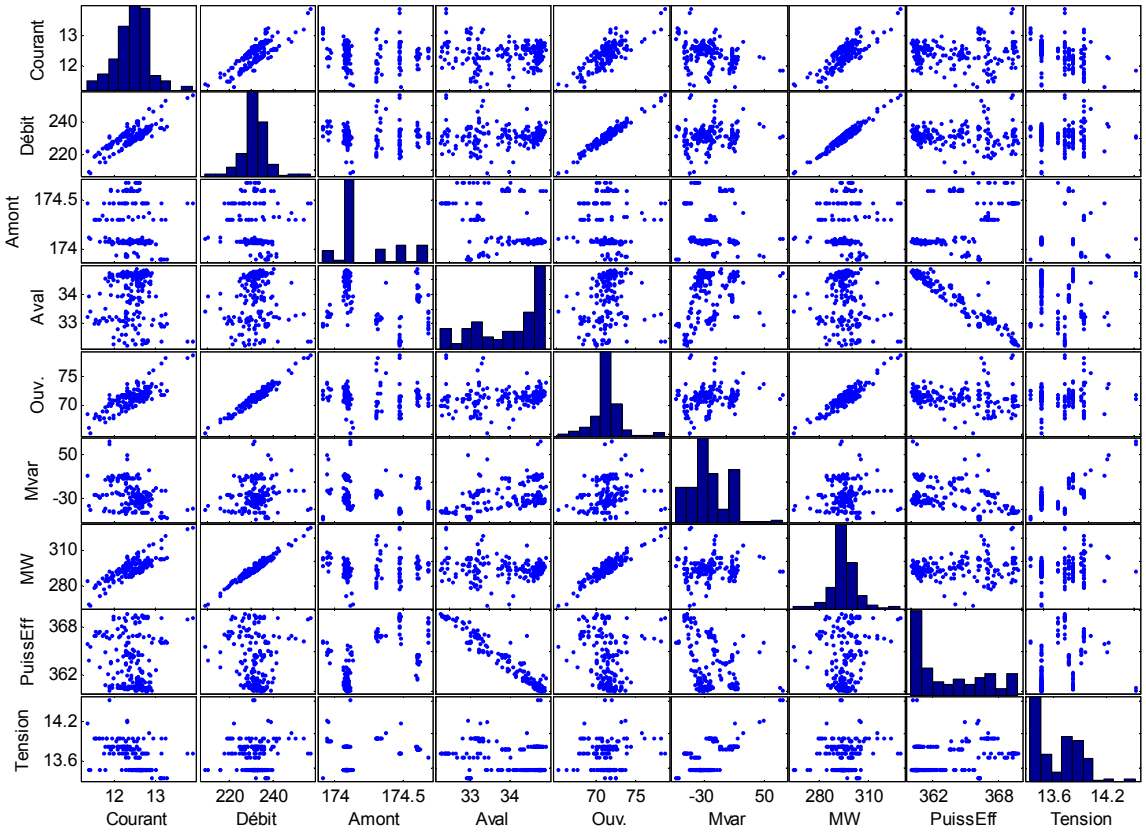


FIGURE D.3. Nuages de points des interactions entre les neuf variables indépendantes pour la grappe 4

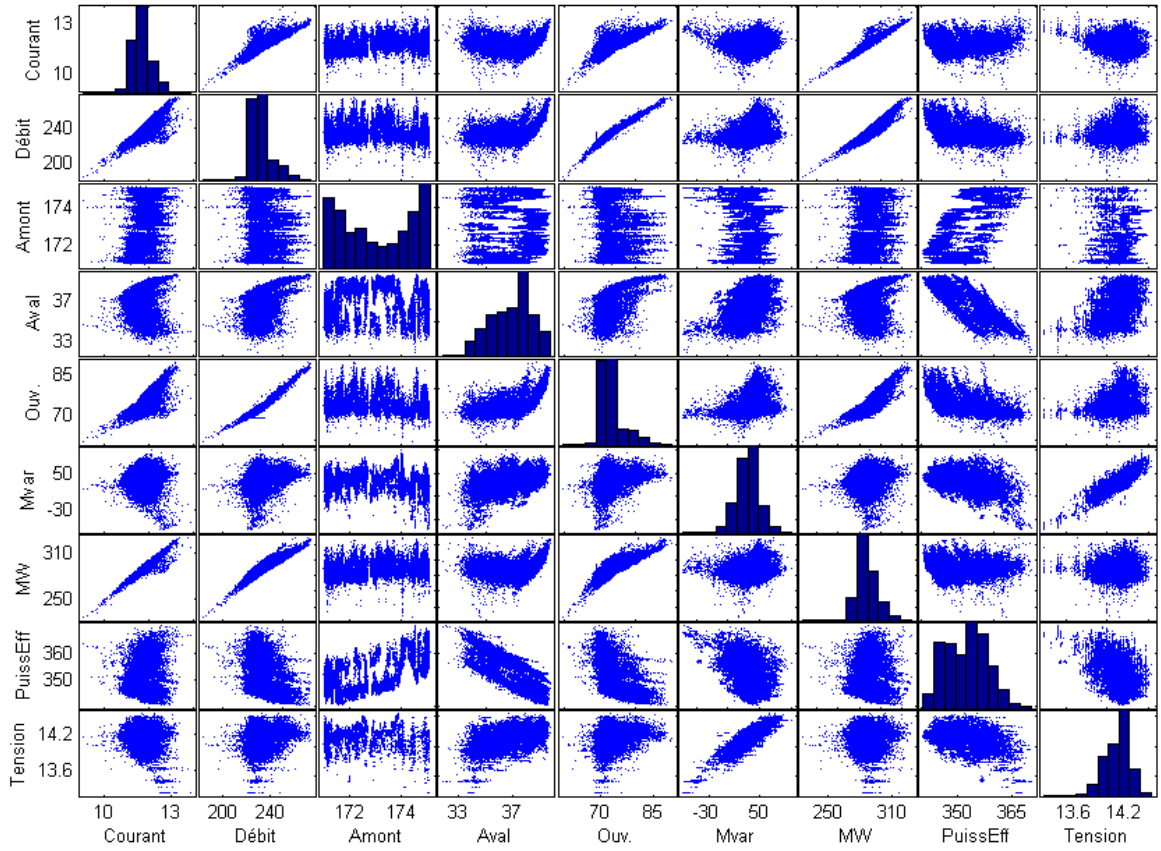


FIGURE D.4. Nuages de points des interactions entre les neuf variables indépendantes pour la grappe 5