Université de Montréal

Évaluation de la modélisation et des prévisions de la vitesse du vent menant à l'estimation de la production d'énergie annuelle d'une turbine éolienne

par

Janie Coulombe

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

avril 2015

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

Évaluation de la modélisation et des prévisions de la vitesse du vent menant à l'estimation de la production d'énergie annuelle d'une turbine éolienne

présenté par

Janie Coulombe

a été évalué par un jury composé des personnes suivantes :

_____
(président-rapporteur)

Christian Léger
_____
(directeur de recherche)

_____
(membre du jury)

Mémoire accepté le:

_____

SOMMAIRE

Suite à un stage avec la compagnie *Hatch*, nous possédons des jeux de données composés de séries chronologiques de vitesses de vent mesurées à divers sites dans le monde, sur plusieurs années. Les ingénieurs éoliens de la compagnie *Hatch* utilisent ces jeux de données conjointement aux banques de données d'*Environnement Canada* pour évaluer le potentiel éolien afin de savoir s'il vaut la peine d'installer des éoliennes à ces endroits. Depuis quelques années, des compagnies offrent des simulations méso-échelle de vitesses de vent, basées sur divers indices environnementaux de l'endroit à évaluer. Les ingénieurs éoliens veulent savoir s'il vaut la peine de payer pour ces données simulées, donc si celles-ci peuvent être utiles lors de l'estimation de la production d'énergie éolienne et si elles pourraient être utilisées lors de la prévision de la vitesse du vent long terme. De plus, comme l'on possède des données mesurées de vitesses de vent, l'on en profitera pour tester à partir de diverses méthodes statistiques différentes étapes de l'estimation de la production d'énergie. L'on verra les méthodes d'extrapolation de la vitesse du vent à la hauteur d'une turbine éolienne et l'on évaluera ces méthodes à l'aide de l'erreur quadratique moyenne. Aussi, on étudiera la modélisation de la vitesse du vent par la distribution *Weibull* et la variation de la distribution de la vitesse dans le temps. Finalement, l'on verra à partir de la validation croisée et du bootstrap si l'utilisation de données méso-échelle est préférable à celle de données des stations de référence, en plus de tester un modèle où les deux types de données sont utilisées pour prédire la vitesse du vent. Nous testerons la méthodologie globale présentement utilisée par les ingénieurs éoliens pour l'estimation de la production d'énergie d'un point de vue statistique, puis tenterons de proposer des changements à cette méthodologie, qui pourraient améliorer l'estimation de la production d'énergie annuelle.

Mots clés : Énergie éolienne, Modélisation de la vitesse du vent, Distribution Weibull, Bootstrap par blocs, Extrapolation de la vitesse du vent, Coefficient de cisaillement du vent.

SUMMARY

Following an internship with the company *Hatch*, we have access to datasets that are composed of wind speed time series measured at different sites accross the world and over several years. The wind speed engineers from *Hatch* are using these datasets jointly with *Environment Canada* databases in order to ascertain the wind energy potential of these sites and to know whether it is worth installing wind turbines there. For a few years, some companies are also offering mesoscale simulations of wind speed based on different environmental characteristics from the site we want to evaluate. We would like to know if it is worth paying for those mesoscale datasets and if they can be used to provide better estimations of the wind energy potential. Among other things, these data could be used to provide a better estimation of the long term mean wind speed. Since we already possess measured datasets, we will also use them to test, with statistical methods, the methodology currently used and the different steps leading to an estimation of the wind energy production. First of all, we will see what are the different methods that could be used to extrapolate wind speed to a wind turbine's height and we will evaluate those methods with the mean squared extrapolation error. Also, we will study wind distribution modelling by the *Weibull* distribution and consider its variability over time. Finally, cross-validation and block bootstrap will be used to see whether we should use mesoscale data instead of wind data from *Environment Canada* or whether it would even be beneficial to use both kind of data to predict wind speed. In summary, the whole methodology used by wind speed engineers to estimate the energy production will be tested from a statistical point of view and we will attempt to propose changes in this methodology that could improve the estimation of the wind speed annual energy production.

Keywords : Wind energy, Wind speed modelling, Weibull distribution, Block bootstrap, Wind speed extrapolation, Wind shear coefficient.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Il s'agit d'un travail qui nécessite beaucoup de persévérance, que la recherche et la rédaction d'un mémoire. Je n'aurais pas pu terminer sans l'aide précieuse de mes amis et de ma famille. À l'Université, mes amis les plus proches, Audrey-Anne, Paule Marjolaine et Alexandre, m'ont supportée durant des mois et m'ont aussi aidée dans les aspects plus techniques de ma recherche. Je tiens aussi à remercier d'autres amis très proches, soient Catherine, Noémie, Audrey, Paméla, Charlène et Sébastien, qui ont été derrière moi durant tout le processus et qui demeurent, tout comme les premiers, des amis incroyables sans qui ce travail n'aurait probablement pas été possible. Je remercie ma famille pour le grand support dont ils ont fait preuve durant les cinq dernières années passées à l'Université. Ils ont eu confiance en moi et m'ont fortement encouragée tout au long de la route. Je remercie aussi mon superviseur, monsieur Christian Léger, avec qui j'ai eu le bonheur de travailler et sans qui plusieurs opportunités ne se seraient pas présentées à moi. Entre autres, mon implication dans la Société statistique du Canada, où j'ai pu participer à différents comités, me permet de garder contact avec plusieurs statisticiens au Canada. Aussi, son aide constante et sa façon d'aborder les différentes questions m'ont permis de développer mon sens critique et de me construire un coffre d'outils tous très utiles pour ma future carrière. Je remercie le Département de mathématiques et de statistique de l'Université de Montréal, qui m'a offert un envionnement où il fut facile de créer des liens forts avec mes collègues, que je côtoierai encore longtemps parce qu'ils sont devenus des amis, et où il fut un plaisir d'apprendre et d'évoluer. Finalement, je tiens à remercier grandement tous les professeurs qui m'ont enseigné les mathématiques et la statistique depuis l'école primaire. Je crois fermement que sans l'intérêt dont ils ont fait preuve durant mon cheminement, mes choix de carrière auraient été différents, à mon plus grand regret. Merci à vous tous d'avoir changé pour le mieux les dernières années et d'avoir ainsi participé avec moi au commencement d'une carrière qui, j'en suis sûre, sera des plus enrichissantes.

INTRODUCTION

L'énergie éolienne est produite au moyen d'un processus aérogénérateur comme la turbine éolienne. Il s'agit d'une tâche peu évidente d'évaluer le potentiel éolien à un site particulier afin de savoir s'il vaut la peine d'y installer une turbine et de pouvoir convaincre la banque qu'un prêt pour l'installation de turbines représente un bon investissement. En effet, on engage généralement des consultants en ingénierie éolienne, lesquels iront installer des mâts de mesure à l'emplacement désiré et analyseront par la suite les données de vitesse et de direction du vent mesurées pour en faire ressortir une évaluation de la puissance du vent à cet endroit. Cependant, le processus étant assez coûteux, des données mesurées ne sont généralement disponibles que sur une courte période et il devient donc plus difficile d'évaluer le potentiel éolien à long terme.

D'autre part, divers outils ont été développés dans le passé afin d'améliorer les prévisions du vent et du potentiel éolien. Entre autres, des données simulées de type "méso-échelle" sont maintenant disponibles par l'entremise de compagnies spécialisées qui utilisent des données environnementales, par exemple l'humidité relative de l'air ou le relief du site, pour simuler des vitesses de vent horaires à l'endroit désiré. Ces méthodes ont, en quelque sorte, fait leur preuve et sont maintenant utilisées par les ingénieurs éoliens dans divers projets. Par contre, on ne retrouve pas dans la littérature de validation exhaustive de ces données simulées permettant de tester l'utilité des données "méso-échelle" dans l'évaluation du potentiel éolien.

Dans le cadre de ce mémoire, on travaillera avec des données fournies par la compagnie *Hatch*, une compagnie d'ingénierie qui possède une équipe complète dédiée à l'analyse de données de vent et d'installation de mâts de mesure. Grâce au programme de stage MITACS, qui a permis le financement d'une partie de mon stage chez *Hatch* durant l'été 2012, j'ai pu compléter quatre mois de recherche sur ces données. Les données m'ont été fournies pour quelques 31

sites dans le monde, situés sur les continents de l'Amérique du Nord, du Sud et de l'Afrique. Pour ces sites, des données de vitesse du vent et de direction du vent mesurées sur des mâts de mesure sont disponibles pour des périodes variant de deux à neuf ans, aux dix minutes. De plus, nous possédons des jeux de données simulées horaires de type "méso-échelle" pour tous ces sites, sur dix ans. Ces données sont elles aussi composées de la vitesse du vent et de la direction sur une base horaire. Nous possédons aussi une troisième source de données pour 15 des sites, soient des données collectées aux mâts de mesure de stations de référence (comme *Environnement Canada*, par exemple, qui collecte plusieurs données environnementales).

Nous utiliserons donc les données disponibles afin de répondre à certaines questions de recherche.

Dans le chapitre 1, nous discuterons d'abord du lien entre les vitesses de vent mesurées aux mâts de mesure et le calcul de la production estimée d'énergie annuelle. On verra les diverses étapes qui mènent au calcul du potentiel éolien, à partir de la collecte des données. On discutera aussi de l'utilisation des courbes de puissance de turbines éoliennes.

Dans le chapitre 2, nous décrirons les divers jeux de données ainsi que les notations utilisées dans le mémoire. Nous présenterons aussi une carte des divers emplacements pour lesquels des données sont disponibles.

Au chapitre 3, le calcul de l'erreur quadratique moyenne d'extrapolation permettra d'évaluer l'erreur due à l'extrapolation des vitesses de vent à une hauteur équivalente à celle des turbines éoliennes. On reverra entre autres les raisons qui nous poussent à vouloir calculer cette erreur. De plus, diverses méthodes d'extrapolation seront comparées.

Au chapitre 4, on discutera de la modélisation de la vitesse du vent. On verra en quoi l'étape de la modélisation de la distribution des vitesses du vent peut influencer l'estimation de la production d'énergie annuelle. Plusieurs tests seront entrepris afin d'évaluer la variabilité des distributions de vent annuelles et mensuelles, pour déterminer s'il serait utile de modéliser les vents de façon séparée dans le temps plutôt que globalement sur plusieurs années.

On étudiera au chapitre 5 la prévision de la vitesse du vent dans le passé à partir de validation croisée. On évaluera dans ce chapitre l'utilité des données

"méso-échelle" pour améliorer les prévisions. Le bootstrap sera quant à lui utilisé et pemettra de tenir compte de la structure de dépendance entre les données, laquelle n'est pas considérée dans les autres chapitres. Il permettra aussi de calculer l'erreur de prévision de la vitesse du vent lorsque des données mesurées sont disponibles sur une trop courte période pour utiliser la validation croisée.

Finalement, on rappellera dans la conclusion les questions de recherche et les résultats importants. On considérera aussi d'autres avenues qui pourraient éventuellement être empruntées.

Chapitre 1

ÉTAPES MENANT AU CALCUL DU POTENTIEL ÉOLIEN D'UN ENDROIT SPÉCIFIQUE

Une méthodologie précise est présentement utilisée par les ingénieurs éoliens afin d'obtenir une estimation de la production d'énergie annuelle. Pour obtenir cette estimation, des données de vitesse et de direction du vent sont d'abord collectées par des anémomètres et des girouettes. Ces instruments sont installés à des mâts de mesure, plantés aux endroits où l'on voudrait prédire le potentiel éolien. Les ingénieurs enregistrent et comptabilisent ces mesures durant une certaine période, dépendant du budget et du temps alloué par les investigateurs. Ce sont ces derniers qui désirent connaître la valeur d'un site en terme d'énergie éolienne.

1.1. Collecte des données

Il est bien important, pour la suite des choses, de comprendre la configuration des anémomètres et des girouettes qui sont installés sur les mâts de mesure de la compagnie. En effet, divers anémomètres sont installés sur ces mâts, généralement disposés à trois hauteurs différentes et doublés de chaque coté du mât. Notez que nous n'utiliserons que les anémomètres d'un côté du mât dans ce mémoire. C'est à partir des mesures prises aux trois hauteurs différentes qu'on peut tenter, par diverses méthodes d'extrapolation, d'évaluer la vitesse du vent à une hauteur plus élevée dans le but de déterminer quelle serait la vitesse du vent équivalente à la hauteur d'une turbine éolienne (généralement installée beaucoup plus haut que les anémomètres sur les mâts de mesure). Pour les analyses qui ne touchent pas à l'extrapolation de la vitesse du vent, nous utilisons tout au long du mémoire les données collectées à l'anémomètre 1, qui est le plus haut sur le mât et dont on possède généralement plus de données sur la période disponible (moins de données manquantes).

Pour la section touchant à l'extrapolation des données, nous utilisons les données collectées aux anémomètres 1, 2 et 3. Ceux-ci sont situés du même coté du mât à trois hauteurs différentes, l'anémomètre 2 étant sous l'anémomètre 1 et l'anémomètre 3 étant au plus bas, sous l'anémomètre 2. Comme le mât de mesure du site 31 possède des anémomètres installés à quatre hauteurs différentes (l'anémomètre 4 étant situé sous l'anémomètre 3 dans ce cas, comparativement au cas où l'anémomètre 4 est de l'autre coté du mât comme à la figure 1.1), celui-ci fera exception lors des tests sur l'extrapolation, dans le chapitre 3. La figure suivante montre, pour les trente premiers sites, la disposition des anémomètres sur le mât de mesure et donne une idée de la différence d'hauteur entre un mât de mesure et une turbine éolienne.

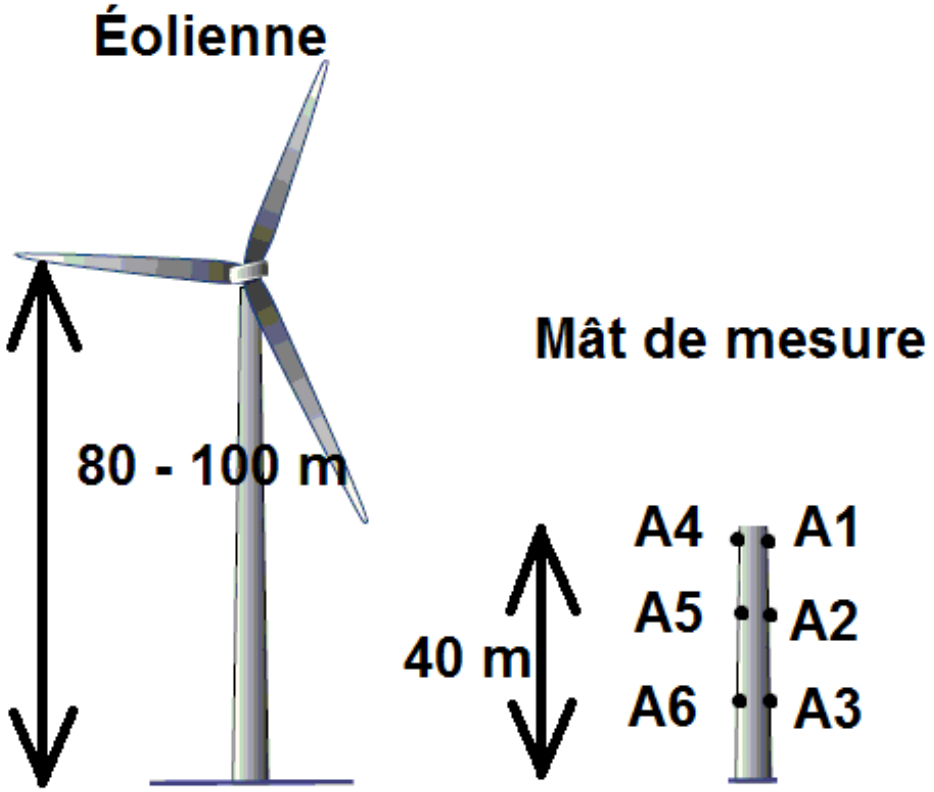

FIGURE 1.1. Représentation d'un mât de mesure classique avec six anémomètres installés à trois hauteurs différentes

1.2. Évaluation du potentiel éolien et prévision de la vitesse du vent à long terme

Nous présentons maintenant certaines des étapes menant à l'estimation de la production d'énergie pour évaluer le potentiel d'un certain site.

1.2.1. Extrapolation des vitesses de vent

Cette étape consiste à extrapoler la vitesse du vent à une hauteur équivalente à celle où l'on installerait une turbine éolienne. En effet, les données de vitesse du vent utilisées sont généralement celles de l'anémomètre 1, qui est le plus haut sur le mât de mesure. Or, cet anémomètre peut être installé à une hauteur de 40 mètres, par exemple, alors qu'on installerait plutôt une turbine à 80 ou 100 mètres. Si l'on désire évaluer la puissance qui résulterait d'une telle turbine, on doit donc évaluer la vitesse du vent à la hauteur de la turbine. On utilisera ainsi, dans le chapitre 3, le calcul de l'erreur quadratique moyenne d'extrapolation afin de déterminer les différences dans l'erreur pour diverses méthodes d'extrapolation.

1.2.2. Modélisation de la distribution de la vitesse du vent

On désire modéliser la distribution de la vitesse du vent de façon paramétrique, par exemple, afin de pouvoir résumer cette distribution par quelques paramètres seulement. De plus, ces paramètres sont utilisés par les ingénieurs éoliens dans le calcul du potentiel éolien d'un site. En effet, la fonction de densité reliée à ces paramètres et à la loi utilisée pour modéliser les vents est utilisée conjointement à la courbe de puissance d'une turbine éolienne pour faire le calcul de l'estimation de la production d'énergie annuelle en kW à un site particulier.

1.2.3. Utilisation de la courbe de puissance pour calculer l'énergie estimée

La courbe de puissance d'une turbine éolienne représente la puissance en kilowatts produite par une turbine éolienne, selon la vitesse du vent (en mètres/seconde) entrant dans la turbine. Aussitôt que nous avons une estimation des probabilités que la vitesse du vent à un site soit dans chaque catégorie de vitesse du vent (0-1 m/s, 1-2 m/s, etc.) faite à partir de notre modélisation de la distribution des vitesses du vent et que nous avons une idée de la puissance créée à partir de chacune de ces catégories de vitesse de vent, grâce à

la courbe de puissance, il nous est possible de calculer l'estimation de la production d'énergie annuelle pour un certain modèle de turbine éolienne à un certain site. On verra plus en détails le calcul de la production d'énergie dans le chapitre 4, où l'on discutera entre autres de la modélisation de la vitesse du vent.

1.2.4. Évaluation de la moyenne du vent à long terme

Il est utile, pour les investigateurs d'un site, d'avoir une estimation de la production d'énergie éolienne annuelle à cet endroit mais il leur faut aussi connaître la tendance à long terme de la vitesse du vent puisque l'année mesurée n'est pas nécessairement représentative de la tendance à long terme. Ainsi, des prévisions de la vitesse du vent dans le passé et l'estimation d'une moyenne de la vitesse du vent long terme leur permet de garantir, en quelque sorte, à la banque qu'il vaut la peine d'installer une turbine éolienne à un certain endroit. Voilà pourquoi on tentera, dans le chapitre 5, d'évaluer l'erreur de prévision de la vitesse du vent à long terme à partir de la validation croisée et du bootstrap, ce dernier tenant compte de la dépendance entre les vitesses de vent mesurées et permettant d'évaluer l'erreur de prévision lorsqu'on possède trop peu de données pour faire une validation croisée.

Chapitre 2

DESCRIPTION DU VOCABULAIRE ET DES JEUX DE DONNÉES UTILISÉS

La vitesse du vent étant très variable dans le temps, l'industrie éolienne doit se baser sur diverses méthodes afin de prévoir les productions futures d'énergie et d'obtenir une évaluation de l'énergie à long terme. Pour ce faire, l'on se basera parfois sur d'autres jeux de données que celles véritablement collectées par des anémomètres, données qui sont disponibles sur une plus longue période, afin de prévoir la tendance sur le long terme des données de vitesse de vent.

Dans ce mémoire, on se concentre sur trois types de jeux de données qui sont disponibles pour l'analyse statistique, que l'on présente maintenant.

2.1. Les jeux de données disponibles et la notation des variables

2.1.1. Les données collectées par la compagnie

Afin d'étudier les prédictions du vent et du potentiel éolien, des données collectées à différents mâts de mesure dans le monde ont été fournies par la compagnie *Hatch*. Les données collectées sont constituées de la vitesse du vent en mètres/seconde et de la direction du vent en degrés. ll s'agit de la première source d'information concernant le potentiel éolien à un endroit dans le monde, et, des divers jeux de données disponibles, sûrement la mesure la plus fiable, puisqu'aucune altération n'y a été portée et que la mesure est prise au site exact où l'on désire obtenir une évaluation du potentiel éolien. Les données ainsi collectées sont disponibles à chaque intervalle de dix minutes, tant pour la vitesse que pour la direction des vents et sont disponibles pour une période

qui varie d'un à neuf ans, dépendamment des sites étudiés. Notons qu'en général, les données collectées sont disponibles sur une période d'un à deux ans seulement. Dans le cadre de cette étude, nous avons recherché des sites pour lesquels la compagnie *Hatch* avait des données collectées sur de plus longues périodes afin d'avoir plus de latitude pour nos analyses.

Comme nous n'utilisons pas les directions du vent dans ce mémoire, nous introduisons maintenant une notation pour la vitesse du vent collectée sans sa direction respective. On notera par $y_{a,i,j}$ la vitesse du vent en m/s à l'anémomètre a, a=1,2,3,4, collectée durant l'heure i, i=0,1,...,N et pour l'indice des dix minutes j, j=1,2,...,6. L'indice j va de 1 à 6 puisque pour chaque heure i on a six périodes de dix minutes indicées de 1 à 6.

Pour ce type de données, nous nous préoccuperons au chapitre 3 des données aux dix minutes seulement. Pour simplifier la notation on utilisera donc plutôt $y_{a,i}$ où i=0,...,N_{10} représentera maintenant l'indice du temps aux dix minutes. On notera donc par $(N_{10}+1)$ le nombre de données de vitesse du vent aux dix minutes disponibles, sans perte de généralité par rapport au site. Au chapitre 4, on utilisera les données aux dix minutes et on se préoccupera seulement de l'anémomètre 1. On utilisera donc y_i où i=0,1,...,N_{10} est l'indice associé aux périodes de dix minutes. Au chapitre 5, on fera la régression des données observées sur les données simulées ou de station de référence. Comme ces deux derniers types de données sont disponibles aux heures seulement, on prendra la moyenne des six observations d'une même heure pour les données mesurées. De plus, on utilisera dans ce chapitre les données de l'anémomètre 1 seulement. On aura donc plutôt les données y_i où

$$y_i = \frac{1}{6}\sum_{j=1}^{6} y_{1,i,j}. \tag{2.1.1}$$

On retrouve à la section 2.2 les détails concernant le traitement des données manquantes lors du calcul de ces moyennes horaires.

2.1.2. Les données simulées de type *méso-échelle*

Il s'agit de l'une des références à plus long terme utilisée par les ingénieurs dans le domaine de l'éolien. Les données sont simulées à partir d'un modèle mathématique sur ordinateur qui tient compte de divers indices environnementaux comme la température extérieure, l'humidité relative, la densité de l'air, etc. Il va sans dire que ces données ne sont disponibles que pour le passé

puisque leur simulation nécessite ces informations. De plus, les données sont simulées pour un site bien précis. Notez que le terme *simulées* fait ici référence à un modèle mathématique déterministe, par opposition à une simulation aléatoire.

Pour obtenir l'un de ces jeux de données, on doit contacter une compagnie spécialisée en simulation des données de vitesse du vent et les prix pour une seule série simulée sur dix ans peuvent atteindre des centaines, voire des milliers de dollars. C'est pourquoi les travailleurs de l'industrie éolienne voudraient s'assurer que ces jeux de données, quoique ne représentant pas des données mesurées, soient fiables et qu'ils peuvent être utilisés comme référence à long terme de la tendance de la vitesse du vent.

Nous utilisons ici les séries de données de vent simulées d'une compagnie particulière afin de tester l'efficacité de ces données. Les données de vitesse du vent et de la direction du vent sont disponibles pour chaque heure, sur dix ans. Ici encore, nous n'utiliserons pas les données de direction du vent dans notre étude. Des jeux de données simulées ont été commandés pour représenter chaque site où des données collectées étaient disponibles. On notera par $x_{s,i}$ la vitesse simulée du vent en m/s pour l'heure i, i=S_{min},$S_{min}+1$,...,-1,0,1,...,N. On remarque donc que la fin de la série des données simulées concordera avec la fin des données collectées, en terme de temps. De plus, S_{min} représentera l'indice de la première heure où l'on possède une valeur de donnée simulée de la vitesse du vent.

2.1.3. Les données d'une station de référence comme celles d'*Environnement Canada*

Il est possible d'obtenir gratuitement des données sur la vitesse et la direction du vent à chaque heure, pour divers endroits au Canada, par l'entremise des stations de mesure d'*Environnement Canada*. Cependant, ces stations de référence ne sont pas disponibles pour tous les sites où des données ont été collectées. De plus, elles peuvent être situées à plusieurs kilomètres des mâts de mesures où la compagnie *Hatch* collecte ses données, faisant en sorte que ces stations de référence représentent souvent des données de moindre valeur pour évaluer le potentiel éolien aux endroits désirés. On voudrait tout de même tester leur utilité dans l'évaluation du potentiel éolien. Notez que ces données sont elles aussi collectées sur des mâts de mesure mais que ces mâts sont beaucoup plus courts que ceux utilisés par la compagnie *Hatch* (10-20 mètres de haut versus 40 mètres).

Pour les sites étudiés hors-Canada, d'autres stations de référence semblables ont aussi été utilisées.

On notera donc par $x_{r,i}$ la vitesse fournie par la station de référence pour l'heure i, i=R_{min},R_{min}+1,...,-1,0,1,...,N. Comme pour les données méso-échelle, la fin de la série concorde avec les données collectées en terme de temps. R_{min} représente l'indice de la première heure où nous possédons une valeur de vitesse du vent collectée à la station de référence. R_{min} peut être inférieur, égal ou supérieur à S_{min}. Cependant, nous utiliserons généralement les deux types de données sur une période de temps commune où les deux sont disponibles. Notez que nous possédons des données provenant de stations de référence pour seulement 15 des 31 sites à l'étude.

2.1.4. Un exemple de statistiques descriptives pour les trois jeux de données

Voici maintenant un résumé descriptif des trois types de jeux de données au site 1. Cela nous permet de voir la différence, surtout en terme de moyenne, des vitesses de vent qui sont collectées aux mâts de la compagnie *Hatch*, de celles collectées par les stations de référence et des données simulées à partir d'un modèle méso-échelle.

TABLEAU 2.1. Statistiques descriptives de la vitesse du vent (m/s) horaire pour les trois jeux de données du site 1 sur une période de temps commune (N=72 704)

Données	Minimum	Q1	Moyenne	Q3	Maximum
Collectées	0,315	4,962	7,086	8,908	22,850
Station de référence	0,000	1,340	2,930	4,190	14,580
Méso-échelle	0,000	4,400	6,586	8,300	24,400

En général, on observe que la moyenne de la vitesse du vent enregistrée aux stations de référence est plus basse que celle des données collectées ou méso-échelle. Cela serait causé par la hauteur des mâts des stations de référence. En effet, la vitesse du vent augmente généralement avec la hauteur et ces mâts sont beaucoup plus courts que ceux utilisés par la compagnie *Hatch*. On peut aussi remarquer que la moyenne de la vitesse du vent des données méso-échelle n'est pas si près de celle des données collectées par *Hatch* bien qu'elles soient planifiées pour simuler le vent à la même hauteur. À la lumière de ces chiffres, on comprend pourquoi l'industrie ne peut pas utiliser les données

simulées méso-échelle seules dans l'estimation de la vitesse du vent passé, tout comme les données provenant des stations de référence. Ces deux jeux de données doivent être utilisés conjointement aux données collectées pour qu'ils soient utiles.

2.2. LE TRAITEMENT DES DONNÉES MANQUANTES

Il est important de noter que pour deux des trois types de données considérées dans ce mémoire, des données manquantes sont présentes et peuvent influencer nos résultats. En effet, pour les données collectées aux mâts de mesure ainsi que pour les données provenant d'une station de référence, jusqu'à 75% et 11,7% des données d'une série peuvent être manquantes, respectivement, et ce pour des périodes de temps très variables. Notez que le site où 75% des données collectées sont manquantes est un cas isolé. On aura généralement entre 0% et 22% de données manquantes aux mâts de mesure de la compagnie. On sait que le gel, entre autres, causerait des arrêts du matériel installé aux mâts de mesure nécessitant souvent une intervention humaine sur place.

Aucune analyse poussée n'a été effectuée dans le cadre de ce mémoire, afin de vérifier si les données manquantes étaient bien réparties (préférablement réparties de manière aléatoire), de sorte à minimiser le biais dans les résultats. Pour le calcul des moyennes aux heures de la série aux dix minutes de données collectées aux mâts de mesure, la moyenne était calculée sur les observations qui ne sont pas manquantes. Par exemple, si au moins l'une des données dix minutes de la prochaine heure sur six n'était pas manquante, on avait une donnée de vitesse du vent horaire non manquante pour cette heure-là, qui consistait en la moyenne des données aux dix minutes disponibles. La moyenne horaire de la vitesse du vent pouvait donc être calculée de la sorte, à l'anémomètre 1 :

$$y_{1,i} = \frac{1}{p_i} \sum_{j=1}^{p_i} y_{1,i,j},$$

où p_i représente le nombre de données aux dix minutes qui ne sont pas manquantes pour l'heure i. Rappelons que $i = 0,...,N$ pour la série de vitesses de vent collectées, où la vitesse moyenne horaire au temps 0 est celle calculée en faisant la moyenne des six prochaines observations aux dix minutes. Par exemple, si l'heure 0 se produit lorsqu'il est 0h00 à un site particulier, la vitesse moyenne correspondante à l'heure 0 sera celle calculée à partir des vitesses de

vent collectées aux heures 0h00, 0h10, 0h20, 0h30, 0h40 et 0h50 qui ne sont pas manquantes. Pour les données manquantes aux dix minutes ou aux heures et sur lesquelles des calculs ont été entrepris dans les chapitres 3, 4 ou 5, notez que les nombres N et N_{10} ont été réajustés de sorte qu'ils représentent bien le nombre de données non manquantes.

2.3. Les divers sites étudiés

Dans l'optique d'analyser plusieurs données de vent provenant de divers milieux, la compagnie *Hatch* a sélectionné quelques 31 sites à travers le monde, ces sites ayant comme point commun des mâts de mesure où des données sur la vitesse et la direction du vent ont été collectées sur une période de plus d'un an. On peut différencier ces sites par diverses caractéristiques physiques, par exemple le fait qu'il y ait de grands arbres autour d'un mât à un site particulier ou plutôt de petits arbres, ce qui influencerait la quantité de vent reçue au mât. Le but de l'analyse n'étant pas de comparer les sites par rapport à leurs caractéristiques respectives, celles-ci ne seront pas présentées. La figure 2.1 présente

Figure 2.1. Différents emplacements des mâts de mesure (triangles rouges et verts)

les différents endroits dans le monde pour lesquels on possède des données de vent tant collectées que simulées. Les endroits représentés par des triangles rouges ou verts seulement sont les sites où l'on possède ces deux types de données et sont donc les sites analysés. Les triangles bleus ne sont pas considérés dans ce mémoire.

Chapitre 3

EXTRAPOLATION DES VITESSES DE VENT

À partir des données de vitesse du vent aux dix minutes des anémomètres 1 ($y_{1,i}$), 2 ($y_{2,i}$) et 3 ($y_{3,i}$), $i = 0,1,...,N_{10}$, il est possible d'estimer la vitesse du vent à une hauteur désirée, équivalente à la hauteur à laquelle on voudrait installer une turbine éolienne. Pour ce faire, on doit se baser sur certaines lois physiques. On utilise ici la loi de puissance pour expliquer le profil du vent.

3.1. LA LOI DE PUISSANCE POUR MODÉLISER LE PROFIL DU VENT

La loi de puissance (Peterson, 1978) explique la relation entre les vitesses du vent et la hauteur à laquelle ces vents sont mesurés. Elle permet donc d'obtenir une estimation de la vitesse du vent à une hauteur désirée, si on possède au moins deux autres vitesses de vent à deux hauteurs différentes.

Loi de puissance :

$$\log\left(\frac{vitesse}{vitesse_{ref}}\right) = \alpha \log\left(\frac{hauteur}{hauteur_{ref}}\right), \tag{3.1.1}$$

où α est le coefficient de cisaillement du vent (c'est pour l'estimation de ce coefficient qu'on a besoin d'au moins deux vitesses de vent à deux hauteurs différentes), *vitesse* et *hauteur* représentent la vitesse du vent que l'on cherche et la hauteur désirée d'extrapolation et $hauteur_{ref}$ et $vitesse_{ref}$ sont la vitesse du vent et la hauteur de référence.

Il est possible de retravailler la formule précédente afin d'obtenir une relation entre le logarithme des vitesses du vent et celui des hauteurs. On obtient ainsi la règle suivante :

$$\log(vitesse) = \underbrace{\log(vitesse_{ref}) - \alpha \log(h_{ref})}_{\text{"ordonnée à l'origine"}} + \underbrace{\alpha}_{\text{"pente"}} \log(h), \tag{3.1.2}$$

où l'on peut remarquer que le coefficient de cisaillement α représente la pente de la régression linéaire simple entre les logarithmes des vitesses et des hauteurs, en prenant pour acquis que la partie notée en (3.1.2) comme l'ordonnée à l'origine est constante, cette quantité dépendant du coefficient de cisaillement ainsi que de la vitesse et de la hauteur d'un point de référence. On tiendra donc compte de cette information afin de pouvoir estimer le coefficient de cisaillement dans la prochaine section.

3.1.1. Les méthodes d'extrapolation et l'estimation du coefficient de cisaillement

On doit d'abord distinguer deux méthodes générales que nous avons utilisées pour extrapoler la vitesse du vent à un certain point : celle déjà utilisée par la compagnie *Hatch*, que nous appellerons la méthode du point de référence et celle que nous proposons dans ce mémoire, appelée la méthode de la régression. Les différentes méthodes seront testées à la prochaine sous-section. Nous commençons d'abord par les présenter.

D'abord, la méthode du point de référence pour estimer le coefficient de cisaillement puis extrapoler les données consiste à dériver de la formule (3.1.1) une formule directe pour obtenir la *vitesse* à la hauteur désirée (*hauteur*) dans la même formule. On transforme donc (3.1.1) de la façon suivante :

$$\begin{aligned} \log\left(\frac{vitesse}{vitesse_{ref}}\right) &= \alpha \log\left(\frac{hauteur}{hauteur_{ref}}\right) \\ \Leftrightarrow \exp\left(\log\left(\frac{vitesse}{vitesse_{ref}}\right)\right) &= \exp\left(\alpha \log\left(\frac{hauteur}{hauteur_{ref}}\right)\right) \\ \Leftrightarrow \frac{vitesse}{vitesse_{ref}} &= \exp\left(\alpha \log\left(\frac{hauteur}{hauteur_{ref}}\right)\right), \end{aligned}$$

ce qui nous mène a la formule d'extrapolation :

$$vitesse = vitesse_{ref} \exp\left(\alpha \log\left(\frac{hauteur}{hauteur_{ref}}\right)\right). \tag{3.1.3}$$

Dans cette même formule, on peut donc utiliser différentes hauteurs et vitesses de référence. Par exemple, si l'on possède des données de vitesse du vent aux anémomètres 1, 2 et 3, et qu'on désire extrapoler à une hauteur où l'on ne possède aucune vitesse du vent collectée, on peut utiliser la vitesse de référence

et la hauteur de référence de l'anémomètre 1, de l'anémomètre 2 ou du troisième anémomètre. Notez que l'anémomètre 1 est généralement celui utilisé pour extrapoler la vitesse du vent à une hauteur plus grande que celle du mât de mesure, puisque qu'il s'agit de l'anémomètre le plus haut sur le mât et qu'il est postulé que les données qui y sont collectées devraient donc être plus représentatives de la vitesse du vent à un point plus haut.

La hauteur et la vitesse de référence ne sont pas les seules quantités que nous pouvons faire varier pour obtenir une extrapolation différente ; on considère aussi deux façons différentes d'estimer le coefficient de cisaillement α, soient le calcul global et le calcul local. Le calcul global est présentement utilisé par la compagnie *Hatch* pour faire les extrapolations, entre autres parce qu'il est plus rapide et pratique à faire. Nous avons suggéré l'utilisation du coefficient de cisaillement local dans leur formule d'extrapolation et comparerons leur performance un peu plus tard. Nous croyons que l'utilisation d'un coefficient de cisaillement local (variant aux dix minutes) mènera possiblement à des prévisions plus précises puisqu'on pourrait penser que le coefficient s'adaptera mieux aux variations de la vitesse du vent dans le temps.

Comme on a pu remarquer en (3.1.2), le coefficient de cisaillement représente, en quelque sorte, la pente de la régression linéaire entre le logarithme des vitesses de vent et le logarithme des hauteurs auxquelles ces vents sont collectés. Le calcul global du coefficient de cisaillement sera donc fait en utilisant comme vitesses dans cette régression les vitesses moyennes aux divers anémomètres sur toute la période collectée. On aura donc, dans la régression linéaire, la variable dépendante v_a qui représente le logarithme de la vitesse moyenne à l'anémomètre a, $a = 1,...,A$, i.e. $v_a = \log(\overline{Y}_a)$ où

$$\overline{Y}_a = \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} y_{a,i}.$$

La variable indépendante utilisée dans la régression est le logarithme de la hauteur de l'anémomètre, soit $u_a = \log(h_a)$. On pourra donc estimer le coefficient de cisaillement global α_G à partir des moindres carrés comme suit :

$$\hat{\alpha}_G = \frac{\sum_{a=1}^{A}(u_a - \overline{u})v_a}{\sum_{a=1}^{A}(u_a - \overline{u})^2}, \tag{3.1.4}$$

où $\overline{u} = \frac{1}{A}\sum_{a=1}^{A} u_a$, avec A le nombre d'anémomètres disponibles utilisés dans le calcul. Notez qu'il aurait été possible aussi de modéliser de façon différente la relation entre le logarithme des hauteurs et des vitesses. Effectivement,

pour le moment, les ingénieurs calculent la vitesse moyenne du vent à différentes hauteurs et modélisent la relation linéaire entre le logarithme des vitesses moyennes sur le logarithme des hauteurs, et c'est donc ce qu'on a fait ici. Or, on sait que la relation est présente au niveau du logarithme de la vitesse par rapport au logarithme de la hauteur, et qu'une régression entre le logarithme des vitesses séparées aux 10 minutes et du logarithme des hauteurs est équivalente à une régression de la moyenne des logarithmes de vitesses aux dix minutes sur le logarithme de la hauteur. On aurait donc aussi pu utiliser la moyenne des logarithmes des vitesses, plutôt que le logarithme de la vitesse moyenne du vent.

On voudrait aussi tester l'extrapolation faite à partir d'un coefficient de cisaillement qui serait ré-estimé pour chaque dix minutes. On appellera ce coefficient le coefficient de cisaillement local. On devra donc modifier le coefficient $\hat{\alpha}_{L,i}$ pour chaque indice i correspondant aux périodes de dix minutes lorsqu'on extrapolera les vitesses de vent à la hauteur h désirée, contrairement au $\hat{\alpha}_G$ qui demeure le même pour chaque extrapolation d'une vitesse de vent. On redéfinit donc un nouveau vecteur pour la variable dépendante de la régression linéaire effectuée pour trouver le coefficient de cisaillement. Il s'agit maintenant de $w_{a,i} = \log(y_{a,i})$, toujours avec la variable indépendante u_a définie plus haut. On peut maintenant estimer le coefficient de cisaillement local aux dix minutes $\alpha_{L,i}$ comme suit :

$$\hat{\alpha}_{L,i} = \frac{\sum_{a=1}^{A}(u_a - \overline{u})w_{a,i}}{\sum_{a=1}^{A}(u_a - \overline{u})^2},\ i = 0,1,...,N_{10}. \tag{3.1.5}$$

On peut donc dire qu'à partir de la formule d'extrapolation (3.1.3) utilisée par la compagnie *Hatch*, on a deux façons d'estimer la vitesse du vent à une hauteur h désirée à partir d'une hauteur de référence h_a et de la vitesse du vent collectée à un anémomètre a plus bas ($y_{a,i}$), soient les méthodes du point de référence globale ou locale :

$$\hat{y}_{a,i,G} = y_{a,i} \exp\left(\hat{\alpha}_G \log\left(\frac{h}{h_a}\right)\right)$$
$$\hat{y}_{a,i,L} = y_{a,i} \exp\left(\hat{\alpha}_{L,i} \log\left(\frac{h}{h_a}\right)\right).$$

Dépendamment du nombre d'anémomètres installés à un mât, on aura donc plusieurs estimations de la vitesse du vent à la hauteur h, à partir de ces deux méthodes. En effet, on verra sur les figures 3.1 et 3.2 que la méthode du point

de référence est équivalente à utiliser un point de référence sur lequel on translate la droite de régression linéaire afin de prédire, au logarithme de la hauteur désirée, le logarithme de la vitesse du vent.

Voyons maintenant la méthode d'extrapolation que nous proposons, soit celle de la régression. Comme il y a une relation linéaire entre le logarithme des vitesses (ou vitesses moyennes) et des hauteurs, on pense à utiliser cette relation afin de prédire à partir de la régression linéaire la vitesse à une certaine hauteur. On n'utilise donc plus la notion de hauteur de référence. On fait donc l'hypothèse que le logarithme de la vitesse du vent (m/s) est relié au logarithme de la hauteur (m) de la façon suivante :

$$\log(v) = \alpha_0 + \alpha_1 \log(h) + \varepsilon.$$

Dans cette formule, α_0 et α_1 sont des coefficients expliquant la relation linéaire entre les deux logarithmes et ε est le bruit du modèle linéaire. Comme on a vu en (3.1.2), α_1 est équivalent au coefficient de cisaillement et α_0 correspond à l'ordonnée à l'origine de la droite de régression. Comme pour la méthode du point de référence, il est possible de calculer le coefficient de cisaillement de deux façons distinctes, qui sont encore la façon globale et locale. On aura donc, ici aussi, deux façons de prédire : la façon globale puis locale, toutes deux à partir de la droite de régression et des coefficients estimés par les formules des moindres carrés.

Supposons donc que nous avons les données collectées à A hauteurs différentes et que nous désirons extrapoler à une hauteur h la vitesse du vent, afin d'obtenir une vitesse estimée $\hat{y}_i$ au temps i, i=0,1,...,N_{10}. On doit d'abord estimer les coefficients de la régression linéaire, pour les deux méthodes. Dans le cas de la méthode globale, on utilise la variable dépendante v_a et la variable indépendante u_a, telles que définies plus tôt. On obtient les estimateurs des coefficients de régression linéaire suivants :

$$\hat{\alpha}_{G,0} = \overline{v} - \hat{\alpha}_{G,1}\overline{u},$$

où $\hat{\alpha}_{G,1}$ est défini en (3.1.4) et $\overline{v} = \frac{1}{A}\sum_{a=1}^{A} v_a$. Pour la méthode locale, on estime ces coefficients pour chaque période de dix minutes :

$$\hat{\alpha}_{L,i,0} = \overline{w}_i - \hat{\alpha}_{L,i,1}\overline{u},$$

où $\hat{\alpha}_{L,i,1}$ est défini en (3.1.5) et $\overline{w}_i = \frac{1}{A}\sum_{a=1}^{A} w_{a,i}$. On peut ensuite prédire au logarithme de la hauteur, le logarithme de la vitesse recherchée. On obtient les

deux façons suivantes d'extrapoler :

$$\hat{y}_G = \exp\left(\hat{\alpha}_{G,0} + \hat{\alpha}_{G,1}\log(h)\right) \tag{3.1.6}$$

$$\hat{y}_{i,L} = \exp\left(\hat{\alpha}_{L,i,0} + \hat{\alpha}_{L,i,1}\log(h)\right).$$

Notez que la valeur de $\hat{y}_G$ ne change pas avec l'indice de temps. On a donc, à partir de la formule (3.1.6), une seule valeur prédite à la hauteur h. On peut donc s'attendre à ce que cette méthode n'offre pas nécessairement de bonnes estimations ponctuelles de la vitesse du vent. Par contre, si on s'intéresse à la vitesse moyenne du vent plutôt qu'à chaque vitesse individuelle aux dix minutes, cette méthode pourrait être adéquate.

Afin de bien comprendre chacune des méthodes utilisées pour l'extrapolation de la vitesse du vent, nous illustrons d'abord à la figure 3.1 les divers cas d'extrapolation pour un coefficient de cisaillement calculé aux dix minutes, soient la méthode du point de référence et la méthode de la régression, pour les cas où des données sont disponibles à deux ou à trois hauteurs différentes. En effet, les sites pour lesquels des données collectées sont disponibles sont ceux où des anémomètres sont installés à trois ou quatre hauteurs. Cependant, comme on désire estimer l'erreur quadratique moyenne d'extrapolation en prédisant la vitesse du vent au plus haut point disponible, on devra garder les données de l'anémomètre le plus haut sur le mât à titre de validation et les calculs de coefficients de cisaillement se feront à partir d'une hauteur de moins. Les données utilisées dans notre exemple sont celles du site 31, pour lequel nous possédons les vitesse de vent à quatre hauteurs différentes. À titre d'illustration, nous allons donc représenter le logarithme des vitesses pour le temps i=113 501 et celui des hauteurs des anémomètres 2, 3 et 4 pour extrapoler à l'anémomètre 1. Le graphique de gauche de la figure 3.1 présentera la méthode du point de référence, pour 2 ou 3 points utilisés et le graphique de droite présentera la méthode de la régression pour extrapoler à partir de 2 ou 3 points. On verra donc, entre autres, la différence entre le fait d'utiliser deux ou trois points afin d'extrapoler. En effet, la compagnie *Hatch* utilise présentement les deux points les plus hauts pour extrapoler même si elle possède des données à plus de deux hauteurs, afin de simplifier la méthode d'extrapolation, et nous proposons d'en utiliser le plus possible (généralement un maximum de trois points à trois hauteurs différentes). Le logarithme de la véritable vitesse du vent collectée à l'anémomètre 1 sera aussi montré afin d'évaluer l'erreur d'extrapolation. Rappelons-nous que ces graphiques présentent les prévisions du vent

sur l'échelle logarithmique. On utilise le logarithme de la hauteur de l'anémomètre 1 pour estimer le logarithme de la vitesse à cet anémomètre. Pour obtenir la vitesse estimée, on doit changer d'échelle en calculant l'exponentielle de la valeur obtenue à partir des diverses méthodes.

On remarque à la figure 3.1 que la méthode du point de référence à deux points

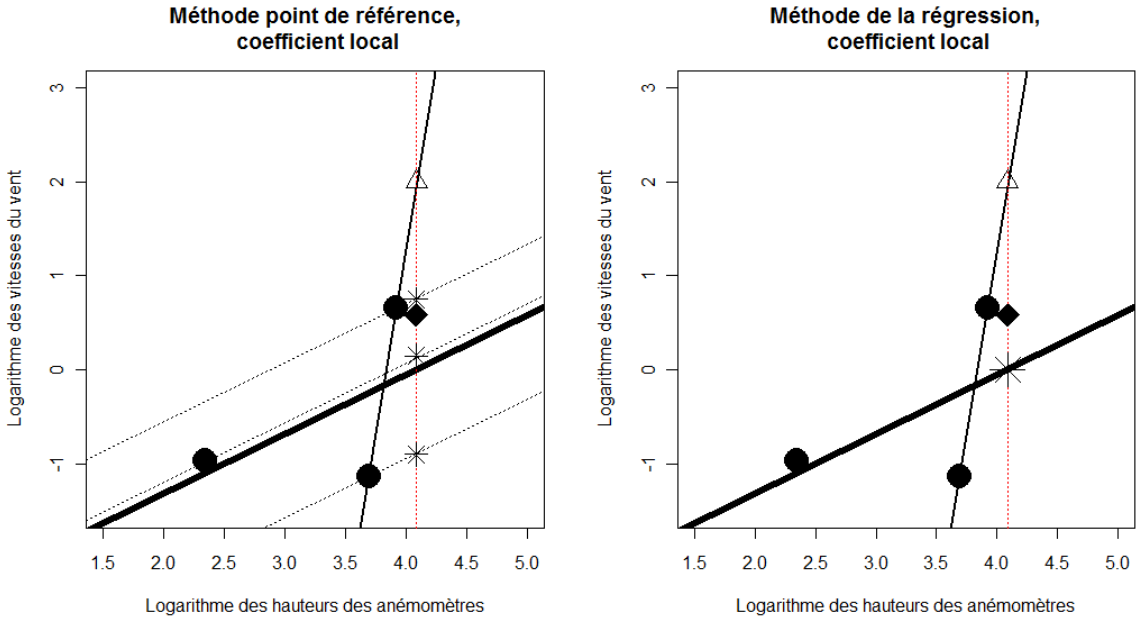

FIGURE 3.1. Les quatre méthodes d'extrapolation à partir d'un coefficient de cisaillement local, au temps i=113 501. La droite noire pleine la moins foncée représente la régression linéaire faite à partir des deux vitesses (anémomètres 2 et 3) et l'autre, celle faite à partir des trois vitesses (anémomètres 2, 3 et 4). Les points noirs sont les logarithmes des vitesses par rapport aux logarithmes des hauteurs pour les trois anémomètres. La ligne pointillée verticale est située en x=log(h1). Les étoiles sont les prévisions à partir des trois différentes hauteurs de référence (graphique de gauche), ou de la régression linéaire à trois points (graphique de droite). Le losange noir est le logarithme de la vitesse mesurée en h1 et les triangles vides sont les prévisions à partir de la régression sur deux points.

et la méthode de la régression à deux points mènent à la même extrapolation, peu importe la hauteur de référence utilisée. Cela est dû au fait que la méthode du point de référence utilise la même droite de régression que la méthode de la régression et la déplace sur chaque point afin de faire une prévision. Ainsi, lorsque cette droite est déplacée sur les deux points qui la définissent, il n'y a aucune différence peu importe le point de référence et la droite reste donc au même endroit qu'avec la méthode de la régression, menant à la même seule

prévision. En ce qui a trait aux méthodes à trois points, remarquez que la méthode du point de référence peut mener à des extrapolations très différentes dépendamment de la hauteur de référence utilisée. Pour le site 31 et le temps i=113 501, on remarque une petite différence entre les extrapolations pour différentes hauteurs de référence (surtout en calculant l'exponentielle des trois prévisions sur le graphique, ce qui nous mènerait à l'estimation de la vitesse du vent, car souvenons-nous que le graphique présente les prévisions sous l'échelle logarithmique). Dans le cas où un des points serait situé assez loin de la droite de régression, on pourrait observer une extrapolation assez différente en utilisant la hauteur de référence correspondante à ce point. Quant à la méthode de la régression à trois points, l'extrapolation demeurera toujours sur la droite de régression.

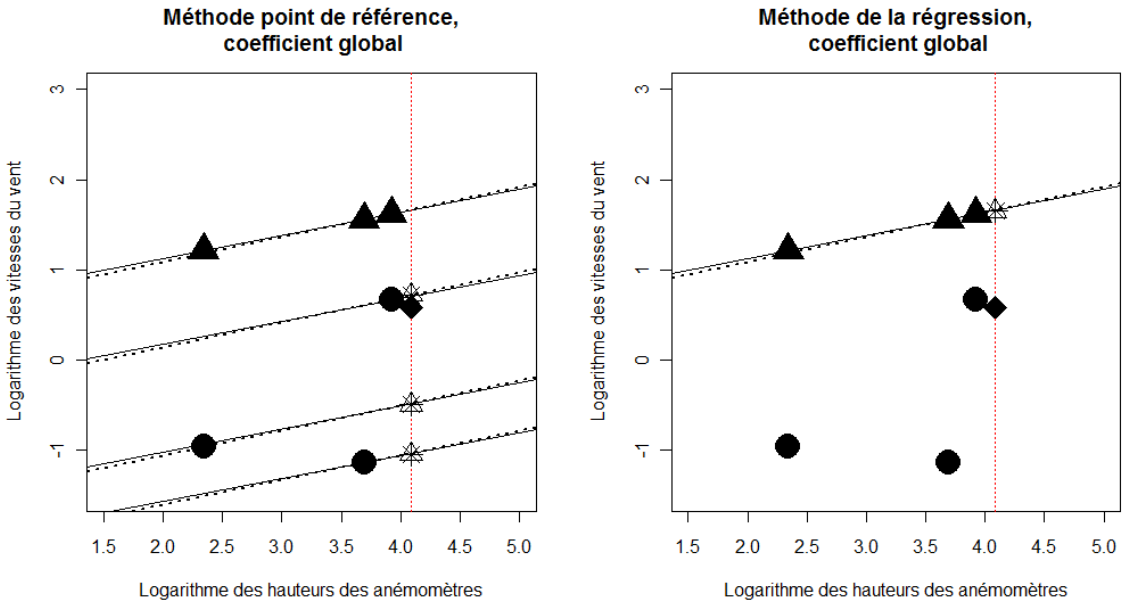

FIGURE 3.2. Les méthodes d'extrapolation à partir d'un coefficient de cisaillement global, au temps i=113501. La droite noire pleine représente la régression linéaire faite à partir des deux vitesses moyennes sur toute la période (anémomètres 2 et 3), et la droite pointillée, celle faite à partir des trois vitesses moyennes sur toute la période (anémomètres 2, 3 et 4). Les points noirs sont les logarithmes des vitesses par rapport aux logarithmes des hauteurs pour les trois anémomètres. Les triangles noirs sont les logarithmes des vitesses moyennes sur toute la période par rapport aux logarithmes des hauteurs pour les trois anémomètres. La ligne pointillée verticale est située en x=log(h1). Les étoiles sont les prévisions à partir des trois différentes hauteurs de référence (graphique de gauche), ou de la régression linéaire à trois points (graphique de droite). Le losange noir est le logarithme de la vitesse mesurée en h1. Les triangles vides sont les prévisions à partir de la régression sur deux points.

Regardons maintenant la figure 3.2, qui illustre chacune des méthodes d'extrapolation lorsqu'un coefficient de cisaillement global est utilisé, toujours à partir des données du site 31 et du temps i=113 501. On doit maintenant noter que pour le graphique de droite, soit la méthode de la régression à deux ou à trois points avec un coefficient de cisaillement global, une seule extrapolation est possible à partir d'une hauteur donnée, quelles que soient les vitesses de vent aux dix minutes. On aura donc, pour toute période de dix minutes, la même valeur extrapolée à l'anémomètre 1. On remarque aussi que les droites de régression sur deux ou trois points varient très peu ici, lorsqu'on utilise la vitesse moyenne sur toute la période. Cela produit donc des prévisions au logarithme

de la hauteur 1 qui sont très similaires quand on utilise un coefficient global, qu'elles soient faites à partir des données de deux ou trois anémomètres.

3.2. Détermination de la méthode optimale d'extrapolation

Voyons maintenant comment nous déterminerons la méthode qui offre les meilleurs résultats en terme d'extrapolation.

Nous utiliserons ici le calcul de l'erreur quadratique moyenne d'extrapolation au plus haut point (l'anémomètre 1) afin d'évaluer la précision de la prévision individuelle, aux dix minutes, de la vitesse du vent. De plus, nous désirons évaluer la qualité de la prévision moyenne sur toute la période de données mesurées disponibles. En effet, les ingénieurs éoliens ont parfois davantage besoin d'une évaluation de cette moyenne à un site que de chaque vitesse ponctuelle horaire ou aux dix minutes. On comparera donc la vraie moyenne de la vitesse du vent à l'anémomètre 1 et la moyenne de la vitesse du vent extrapolée à la hauteur de l'anémomètre 1, soient :

$$\overline{y}_1 = \frac{1}{N_{10}+1}\sum_{i=0}^{N_{10}} y_{1,i}$$

et

$$\overline{\hat{y}}_E = \frac{1}{N_{10}+1}\sum_{i=0}^{N_{10}} \hat{y}_{i,E},$$

où E est la méthode d'extrapolation utilisée. Cela permettra de vérifier laquelle des méthodes serait la plus efficace en terme d'estimation de la vitesse de vent moyenne.

Comme l'on désire calculer l'erreur d'extrapolation et que l'on se basera sur les données mesurées à l'anémomètre 1, on n'utilisera pas dans nos estimations de la vitesse extrapolée les données de l'anémomètre 1. Pour 30 des 31 sites, nous n'avons que trois hauteurs différentes où des données sont mesurées, ne nous laissant la possibilité d'utiliser que deux des trois anémomètres (les anémomètres 2 et 3) pour extrapoler au plus haut anémomètre. Nous utiliserons donc ces vitesses pour faire les régressions linéaires nécessaires et extrapolations, telles que vues à la section précédente. Pour le site 31, nous utiliserons aussi les données de l'anémomètre 4, situé sous l'anémomètre 3. Nous extrapolerons donc à partir des diverses méthodes et calculerons dans un premier temps l'EQM entre les valeurs estimées et les valeurs réelles de vitesse du vent

collectées à l'anémomètre 1.

L'EQM d'extrapolation (EQME) sera calculée comme suit :

$$EQME = \frac{1}{N_{10}+1}\sum_{i=0}^{N_{10}}(y_{1,i} - \hat{y}_{i,E})^2,$$

où E représente la méthode d'extrapolation utilisée. On présente maintenant les EQME pour chaque méthode d'extrapolation et pour chaque site. L'EQME minimale pour chaque site est mise en gras dans les prochains tableaux.

TABLEAU 3.1. EQME calculée à partir des diverses méthodes d'extrapolation, pour les sites 1 à 30 (2 anémomètres disponibles seulement)

Site	$\hat{y}_{2,i,G}$	$\hat{y}_{2,i,L}$	$\hat{y}_G$	$\hat{y}_{i,L}$	Site	$\hat{y}_{2,i,G}$	$\hat{y}_{2,i,G}$	$\hat{y}_G$	$\hat{y}_{i,L}$
1	**0,149**	0,286	9,107	0,286	16	0,076	**0,057**	8,669	**0,057**
2	**0,095**	0,110	28,087	0,110	17	0,032	**0,023**	8,075	**0,023**
3	0,071	**0,069**	11,144	**0,069**	18	**0,185**	0,288	12,656	0,288
4	**0,071**	0,270	8,953	0,270	19	0,032	**0,019**	10,704	**0,019**
5	0,222	**0,148**	12,657	**0,148**	20	0,039	**0,028**	10,026	**0,028**
6	0,097	**0,086**	9,038	**0,086**	21	**0,634**	0,747	18,980	0,747
7	**0,046**	0,053	7,095	0,053	22	0,129	**0,053**	13,658	**0,053**
8	**0,107**	0,166	13,467	0,166	23	1,432	**1,207**	13,352	**1,207**
9	**0,052**	0,067	13,103	0,067	24[1]	-	-	-	-
10	0,165	**0,142**	11,749	**0,142**	25	**0,025**	0,040	16,890	0,040
11	0,163	**0,076**	7,688	**0,076**	26	**0,189**	0,208	18,209	0,208
12	0,030	**0,023**	6,299	**0,023**	27	0,223	**0,195**	27,689	**0,195**
13	**0,303**	3,039	18,188	3,039	28	**1,406**	1,501	17,557	1,501
14	**0,071**	0,085	25,538	0,085	29	0,243	**0,168**	8,905	**0,168**
15	**0,065**	0,069	13,706	0,070	30	**0,126**	0,220	10,995	0,220

[1]Le site 24 ne possède pas d'autres anémomètres que les deux plus hauts donc ce test ne s'applique pas à ce site.

TABLEAU 3.2. EQME calculée à partir des diverses méthodes d'extrapolation, au site 31

Site	Nb de h_i utilisé	$\hat{y}_{4,i,G}$	$\hat{y}_{3,i,G}$	$\hat{y}_{2,i,G}$	$\hat{y}_{4,i,L}$	$\hat{y}_{3,i,L}$	$\hat{y}_{2,i,L}$	$\hat{y}_G$	$\hat{y}_{i,L}$
31	3	2,619	0,273	0,083	0,105	0,212	**0,071**	7,915	0,114
	2	-	-	**0,083**	-	-	0,093	8,252	0,093

Notez que nous avons pu calculer l'EQME à partir de trois hauteurs (anémomètres 2, 3 et 4) ou de deux hauteurs seulement (anémomètres 2 et 3), comme

le fait *Hatch*, pour le site 31, puisque nous possédons des données à quatre hauteurs pour ce site, et que nous avons donc présenté les résultats pour les différents nombres d'anémomètres utilisés. Ces résultats se retrouvent dans le tableau 3.2.

En ce qui a trait à la comparaison des vitesses extrapolées moyennes et de la vitesse collectée moyenne à l'anémomètre 1, on calcule l'erreur relative entre ces deux mesures pour chaque méthode d'extrapolation. L'erreur relative est calculée comme suit :

$$ER = 100 \times \frac{\overline{\hat{y}}_E - \overline{y}_1}{\overline{y}_1}.$$

On présente maintenant les erreurs relatives entre les vitesses de vent moyennes pour chaque méthode d'extrapolation et pour chaque site. En gras, on retrouve pour chaque site l'erreur relative minimale.

TABLEAU 3.3. Erreurs relatives (%) entre la vitesse collectée moyenne à l'anémomètre 1 et les vitesses extrapolées moyennes calculées à partir de chaque méthode, pour les sites 1 à 30 (2 anémomètres disponibles seulement)

Site	$\overline{\hat{y}}_{2,i,G}$	$\overline{\hat{y}}_{2,i,L}$	$\overline{\hat{y}}_G$	$\overline{\hat{y}}_{i,L}$	Site	$\overline{\hat{y}}_{2,i,G}$	$\overline{\hat{y}}_{2,i,L}$	$\overline{\hat{y}}_G$	$\overline{\hat{y}}_{i,L}$
1	2,661	**2,393**	2,661	**2,393**	16	2,287	**2,032**	2,287	**2,032**
2	0,768	**0,538**	0,768	**0,538**	17	0,869	**0,656**	0,869	**0,656**
3	**-2,294**	-2,433	**-2,294**	-2,433	18	0,446	**0,220**	0,446	**0,220**
4	0,409	**0,207**	0,409	**0,207**	19	0,332	**0,226**	0,332	**0,226**
5	1,214	**0,664**	1,214	**0,664**	20	**-0,453**	-0,611	**-0,453**	-0,611
6	**-0,485**	-0,634	**-0,485**	-0,634	21	5,697	**5,347**	5,697	**5,347**
7	0,554	**0,335**	0,554	**0,335**	22	**0,066**	-0,140	**0,066**	-0,140
8	**0,061**	-0,188	**0,061**	-0,188	23	0,402	**-0,314**	0,402	**-0,314**
9	**-0,202**	-0,278	**-0,202**	-0,278	24[1]	-	-	-	-
10	0,669	**0,388**	0,669	**0,388**	25	0,455	**0,293**	0,455	**0,293**
11	**-0,027**	-0,338	**-0,027**	-0,338	26	0,325	**0,107**	0,325	**0,107**
12	**-0,059**	-0,150	**-0,059**	-0,150	27	**-1,085**	-1,264	**-1,085**	-1,264
13	**-0,302**	-1,130	**-0,302**	-1,130	28	4,726	**4,312**	4,726	**4,312**
14	1,634	**1,492**	1,634	**1,492**	29	0,896	**0,479**	0,896	**0,479**
15	**-1,167**	-1,316	**-1,167**	-1,316	30	1,521	**1,261**	1,521	**1,261**

[1]Le site 24 ne possède pas d'autres anémomètres que les deux plus hauts donc ce test ne s'applique pas à ce site.

TABLEAU 3.4. Erreurs relatives (%) entre les vitesses de vent moyennes calculées à partir des diverses méthodes d'extrapolation et la vitesse moyenne à l'anémomètre 1, au site 31

Site	Nb de h_i utilisé	$\overline{\hat{y}}_{4,i,G}$	$\overline{\hat{y}}_{3,i,G}$	$\overline{\hat{y}}_{2,i,G}$	$\overline{\hat{y}}_{4,i,L}$	$\overline{\hat{y}}_{3,i,L}$	$\overline{\hat{y}}_{2,i,L}$	$\overline{\hat{y}}_{G}$	$\overline{\hat{y}}_{i,L}$
31	3	-0,839	-1,113	-0,669	0,298	0,714	**0,147**	-0,669	0,340
	2	-	-	0,497	-	-	**0,010**	0,497	**0,010**

Nous avons encore une fois testé les méthodes d'extrapolation utilisant deux anémomètres (anémomètres 2 et 3) ou trois anémomètres (anémomètres 2, 3 et 4). Remarquez que nous ne notons dans le tableau que les résultats pour une hauteur de référence reliée à l'anémomètre 2 lorsque nous utilisons deux anémomètres pour extrapoler, puisque les résultats sont les mêmes si on utilise la hauteur de référence de l'anémomètre 3 (une régression linéaire sur deux points mène à la même prévision, qu'on se base sur la hauteur reliée au premier ou au deuxième point de la régression, i.e. $\hat{y}_{2,i,G} = \hat{y}_{3,i,G}$ et $\hat{y}_{2,i,L} = \hat{y}_{3,i,L}$ si l'on utilise seulement les anémomètres 2 et 3).

3.2.1. Discussion

Dans ce chapitre, nous voulions vérifier si la méthode présentement utilisée par *Hatch* pour l'extrapolation de la vitesse du vent, soit l'extrapolation à partir d'un coefficient de cisaillement global basée sur deux anémomètres, était adéquate. Nous voulions voir l'effet d'utiliser un coefficient de cisaillement local par rapport à global, et s'il est possible de diminuer l'erreur quadratique moyenne d'extrapolation ainsi que l'erreur relative de la moyenne de la vitesse du vent estimée par rapport à la vraie moyenne, en utilisant un anémomètre de plus pour extrapoler la vitesse du vent.

Tout d'abord, on peut voir dans les tableaux 3.1 et 3.2 que, comme prévu, les méthodes d'extrapolation du point de référence et de régression avec coefficient de cisaillement local mènent toujours aux mêmes EQME si l'on utilise seulement deux anémomètres dans le calcul. Ensuite, si l'on fait le décompte des méthodes où l'on obtient les meilleurs résultats, donc celles menant aux EQME minimales (nombres en gras), on trouve que la méthode du point de référence avec coefficient de cisaillement global offre le meilleur résultat pour 16 des 30 sites où l'on possédait des résultats (en comptant le site 31). La moyenne des erreurs quadratiques moyennes d'extrapolation pour les 30 sites abonde dans le même sens, avec une moyenne d'EQME de $0{,}219$ pour l'extrapolation

$\overline{\hat{y}}_{2,i,G}$, de $13{,}348$ pour $\overline{\hat{y}}_G$, et de $0{,}318$ pour les méthodes $\overline{\hat{y}}_{2,i,L}$ et $\overline{\hat{y}}_{i,L}$ en comptant les extrapolations à partir de deux anémomètres du site 31. Pour le site 31 et l'extrapolation à partir de trois anémomètres, la méthode optimale est celle du point de référence avec coefficient de cisaillement local où la hauteur de référence est h_2 (donc l'anémomètre le plus haut qu'on puisse utiliser). L'utilisation de trois hauteurs plutôt que deux nous mène à de bonnes améliorations de l'erreur quadratique moyenne d'extrapolation (EQME de 0,071 contre 0,083 pour les meilleures méthodes à trois ou deux hauteurs utilisées, respectivement, ce qui équivaut donc à une amélioration de 14%). De plus, on voit que pour chaque anémomètre k, $\hat{y}_{k,i,L}$ fait toujours mieux que $\hat{y}_{k,i,G}$ et que l'EQME de $\hat{y}_{k,i,G}$ diminue à mesure qu'on prend un anémomètre plus haut (équivalent à un k qui diminue), mais que le comportement de $\hat{y}_{k,i,L}$ ne suit pas la même tendance.

En ce qui a trait à l'erreur relative entre la véritable moyenne des vitesses de vent à l'anémomètre 1 et la moyenne des vitesses de vent extrapolées à la hauteur de l'anémomètre 1 à partir des diverses méthodes d'extrapolation, on peut voir dans les tableaux 3.3 et 3.4 que pour 19 des 30 sites où l'on possède des résultats (en comptant le site 31), la meilleure méthode est celle du point de référence avec coefficient de cisaillement local ou de la régression avec coefficient de cisaillement local, qui menaient toujours aux mêmes résultats, comme attendu. Ce sont donc ces deux méthodes qui réussissent le mieux à estimer la moyenne de la vitesse du vent à la hauteur désirée, ici. De plus, la moyenne des erreurs relatives absolues est de 1,085% pour les méthodes $\overline{\hat{y}}_{2,i,G}$ et $\overline{\hat{y}}_G$ et de 0,992% pour les méthodes $\overline{\hat{y}}_{2,i,L}$ et $\overline{\hat{y}}_{i,L}$ (notez qu'il s'agit de faibles erreurs relatives, puisqu'on parle d'erreurs autour de 1%). Si l'on se base sur l'estimation de la moyenne de la vitesse extrapolée, on préférerait donc utiliser un coefficient de cisaillement local plutôt que global, c'est-à-dire recalculer le coefficient de cisaillement à chaque dix minutes au lieu de le calculer à partir des vitesses moyennes sur toute la période où l'on possède des données collectées de vitesses du vent. Notez aussi que tel que prévu, les erreurs quadratiques moyennes d'extrapolation pour les prévisions $\overline{\hat{y}}_G$ sont terribles (on s'y attendait puisque la prévision demeure la même peu importe le temps, pour une même hauteur) mais que la méthode n'est pas si mal pour estimer le vent moyen (elle mène aux mêmes résultats que la méthode du point de référence avec coefficient de cisaillement global). Pour ce qui est du site 31 et de l'extrapolation à partir de trois anémomètres, l'erreur relative minimale est celle

obtenue à partir de la méthode du point de référence avec coefficient de cisaillement local et hauteur de référence h_2. Les meilleurs résultats quant à l'erreur relative pour ce site sont obtenus en n'utilisant que deux hauteurs pour extrapoler à h_1 (erreurs relatives de 0,010$\%$ à deux hauteurs utilisées contre 0,147$\%$ à trois hauteurs utilisées, pour les meilleurs méthodes respectives) mais on remarque que toutes les erreurs relatives sont inférieures à 1,2$\%$ en valeur absolue, donc la méthode importe peu pour ce site. Si on regarde l'ensemble général des sites, on voit que l'erreur relative est faible dans tous les cas et la moyenne estimée à la hauteur h_1 est assez près de la moyenne mesurée à l'anémomètre 1. Les erreurs relatives sont toutes inférieures à 6$\%$, tous les sites et les modèles étant confondus.

De façon générale, on ne trouve pas de méthode qui se démarque vraiment des autres lorsqu'on utilise seulement deux anémomètres pour extrapoler à une troisième hauteur. Il semblerait aussi qu'il y ait un certain avantage à utiliser un anémomètre de plus pour extrapoler ici. Par contre, il faut demeurer prudent avec cette conclusion puisque nous n'avons pu comparer les extrapolations basées sur deux anémomètres à celles sur trois anémomètres que pour un seul site. La méthode présentement utilisée par *Hatch*, soit l'extrapolation à partir de deux hauteurs seulement et d'un seul coefficient de cisaillement global, semble donc être plutôt adéquate puisque l'utilisation d'un coefficient de cisaillement local (qui entraîne des calculs un peu plus complexes) n'a pas grandement amélioré les résultats ici. Or, on a remarqué au site 31 qu'on améliorait la précision de 14$\%$ en utilisant 3 points (en termes d'EQME) alors que pour la moyenne, toutes les erreurs relatives étaient plutôt faibles (sous 1,2$\%$). Il pourrait donc être avantageux pour la compagnie *Hatch* d'utiliser un anémomètre de plus et il aurait été intéressant d'avoir les vitesses de vent à quatre hauteurs différentes pour plus de sites, afin de voir la tendance quant à l'utilisation d'un anémomètre de plus.

Chapitre 4

MODÉLISATION DE LA DISTRIBUTION DE LA VITESSE DES VENTS

La raison première justifiant la modélisation de la distribution de la vitesse des vents est le désir de connaître les paramètres pouvant tenter de résumer à eux seuls cette distribution. On fait d'abord l'hypothèse que les vents peuvent être assez bien modélisés par une loi dont les paramètres ne changeraient pas nécessairement avec le temps. On utilisera les données de vitesse du vent aux dix minutes du plus haut anémomètre de chaque site seulement (anémomètre 1) tout au long de ce chapitre et on fera d'abord l'hypothèse que les données sont indépendantes et identiquement distribuées. La section 4.4.2 se concentrera sur la dépendance entre les observations.

Un grand nombre de méthodes permettant de modéliser la distribution de la vitesse des vents ont été explorées dans la littérature. En effet, plus d'une loi de probabilité permettrait, à partir de la fonction de densité respective, d'approcher cette distribution. La loi la plus couramment utilisée serait celle de *Weibull* (Burton, 2011), qu'on verra dans la prochaine section et qui sera la loi utilisée tout au long du chapitre pour modéliser la vitesse du vent. On retrouve d'autres distributions dans la littérature, comme celle de *Rayleigh-Rice*, et on mentionne dans Drobinski (2012) qu'elle serait préférable à la distribution *Weibull* lorsque les vents sont très variables selon la direction d'observation (donc pour des vents qui ne seraient pas isotropiques). Dans ce chapitre, nous tenterons de vérifier l'ajustement de la *Weibull* sur la distribution des vents, en plus de tester si les paramètres de la *Weibull* devraient être réajustés pour chaque année ou chaque mois plutôt que de façon globale. Nous discuterons aussi de l'estimation de la production d'énergie annuelle à partir de la *Weibull* et de la courbe de puissance associée aux turbines éoliennes.

4.1. LA LOI DE *Weibull* ET SES CARACTÉRISTIQUES

En plus d'être la distribution utilisée par la compagnie *Hatch* pour modéliser celle de la vitesse des vents lors du calcul de la production d'énergie, il est dit à maintes reprises dans la littérature que la distribution des vitesses de vent peut être représentée (ou encore, caractérisée) par la fonction de densité reliée à la loi de *Weibull*. Revoyons maintenant un résumé des caractéristiques de la loi *Weibull*.

La loi de *Weibull* à deux paramètres est une loi de probabilité continue caractérisée par les paramètres de forme, noté k, et d'échelle, noté λ. La fonction de densité de la *Weibull* est la suivante :

$$f_{k,\lambda}(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(\frac{x}{\lambda})^k}.$$

La fonction de répartition de la *Weibull*, quant à elle, est définie par :

$$F_{k,\lambda}(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}.$$

L'espérance et la variance d'une variable aléatoire X de la distribution *Weibull* sont respectivement données par :

$$\begin{aligned}\mathbb{E}_{k,\lambda}(X) &= \lambda\Gamma\left(1+\frac{1}{k}\right)\\ \mathbb{V}ar_{k,\lambda}(X) &= \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right].\end{aligned}$$

La figure 4.1 montre quatre exemples de fonctions de densité d'une *Weibull*.

Il est intéressant de noter que la transformation vue dans le chapitre 3, soit l'extrapolation à partir de la formule (3.1.3) d'un jeu de données de loi *Weibull*, mènera à un nouveau jeu extrapolé aussi de loi *Weibull*. Soit Z_1,Z_2,...,Z_n un jeu de données distribuées selon une *Weibull*(k,λ) à une hauteur h_{ref} et E_1,E_2,...,E_n ces mêmes données une fois extrapolées à une hauteur h à partir de la formule (3.1.3). On a donc ceci :

$$\begin{aligned}E_i =& Z_i \exp(\alpha(log(h) - log(h_{ref})))\\ =& Z_i \exp(cte),\end{aligned}$$

où $cte = \alpha(log(h) - log(h_{ref}))$ est constant dans le cas où le coefficient de cisaillement demeure le même pour tout temps i. Ainsi, E_i est *Weibull*(k,$\lambda \exp(cte)$). Une fois nos vitesses de vent extrapolées, on peut donc toujours modéliser la distribution de la vitesse du vent par une *Weibull*, ce qui est utile lors de l'estimation de la production d'énergie. En effet, les ingénieurs éoliens feront leur

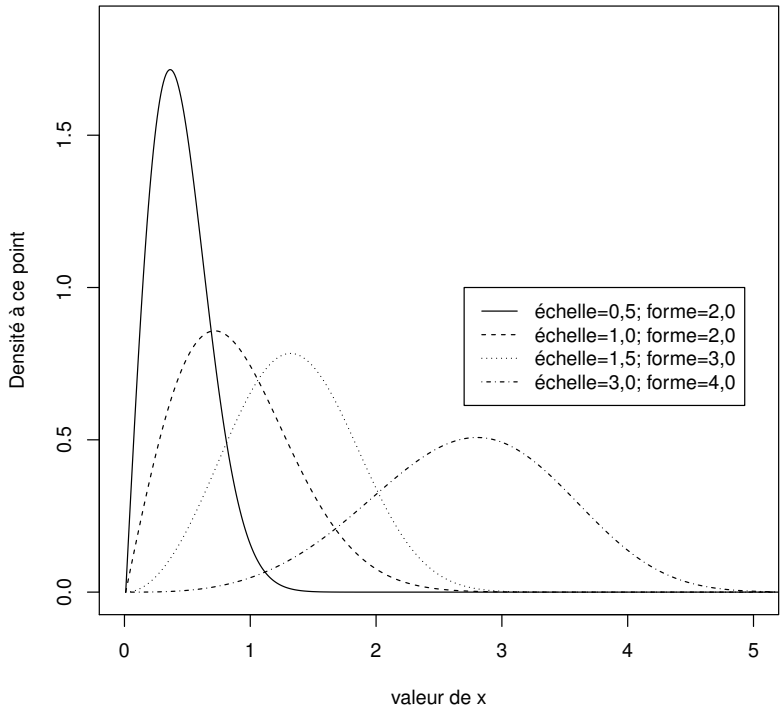

FIGURE 4.1. Exemples de la fonction de densité d'une distribution *Weibull* pour divers paramètres

estimation à partir des données extrapolées à la hauteur de la turbine qu'ils désirent installer. Notez que nous ne considérerons pas l'extrapolation avant l'estimation de la production d'énergie dans ce mémoire. Généralement, comme la hauteur à laquelle est mesurée la vitesse du vent est plus basse que celle d'une turbine éolienne, les ingénieurs vont d'abord extrapoler la vitesse à la hauteur correspondante à la turbine, puis modéliser la distribution une fois extrapolée afin d'estimer la production d'énergie à partir d'une turbine particulière. Or, ici, ces deux étapes seront vues de façon séparée et l'estimation de la production d'énergie sera faite sur des données non extrapolées.

4.2. L'ESTIMATION DES PARAMÈTRES

Soit X_1,X_2,...,X_n un échantillon indépendant et identiquement distribué selon une loi *Weibull*(k,λ). Différentes méthodes statistiques nous permettent d'estimer les deux paramètres (de forme et d'échelle) de la *Weibull* qu'on désire ajuster sur la distribution de la vitesse des vents aux dix minutes en mètres/seconde.

Dans ce mémoire, nous utilisons les paramètres estimés à partir de la fonction de vraisemblance. Il est possible de trouver ces estimateurs à partir d'un logiciel statistique ou en résolvant l'équation (4.2.1) par rapport à $\hat{k}$ puis en calculant $\hat{\lambda}$ à partir de la formule (4.2.2), pour arriver aux mêmes résultats (Johnson, 1994).

$$\hat{k} = \left(\frac{\sum_{i=1}^{n} X_i^{\hat{k}} \log X_i}{\sum_{i=1}^{n} X_i^{\hat{k}}} - \frac{1}{n} \sum_{i=1}^{n} \log X_i \right)^{-1} \tag{4.2.1}$$

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^{\hat{k}} \right)^{\frac{1}{\hat{k}}}, \tag{4.2.2}$$

où n est le nombre d'observations sur lesquelles est ajustée la distribution.

Une fois que les estimateurs du maximum de vraisemblance (EVM) sont trouvés, il est possible de dessiner par-dessus l'histogramme des vitesses de vent la courbe de densité de la *Weibull* ajustée. La figure 4.2 présente un exemple de l'ajustement d'une fonction de densité de *Weibull* sur la distribution des vitesses du vent en mètres/seconde. Les vitesses de vent utilisées sont celles du site 1, où environ neuf ans de données sont disponibles.

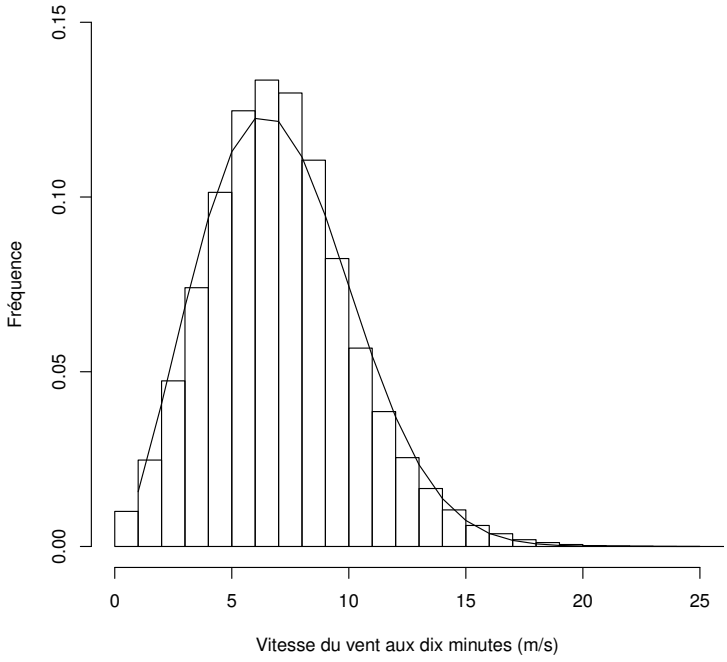

FIGURE 4.2. Exemple d'ajustement de *Weibull* sur l'histogramme des vitesses du vent aux dix minutes au site 1, sur neuf ans ($\hat{k} = 2{,}423$, $\hat{\lambda} = 7{,}987$)

4.3. VÉRIFICATION DE L'AJUSTEMENT DE LA *Weibull* SUR LA DISTRIBUTION DE LA VITESSE DES VENTS

En premier lieu, nous désirons vérifier l'ajustement de la *Weibull* sur la distribution des vitesses de vent : à quel point cette distribution s'ajuste bien, afin de l'utiliser pour estimer celle de la vitesse des vents ? On fait donc un test d'adéquation du khi-deux pour vérifier si la *Weibull* convient à la modélisation des vitesses de vent aux dix minutes.

4.3.1. Tests d'adéquation pour la distribution *Weibull*

Ici, nous avons utilisé le test du khi-deux. Celui-ci s'applique bien dans notre cas, puisque nous pouvons utiliser ce test sur des intervalles d'une largeur de 1 m/s entre 0 et 25 m/s, sachant que la fonction de puissance est discrétisée de cette façon et que nous l'utiliserons un peu plus tard. Pour utiliser ce test, nous devions estimer les probabilités théoriques de se retrouver dans chacune des classes de la distribution de la vitesse du vent, en se basant sur la

Weibull ajustée sur les données collectées de vitesse du vent. Pour ce faire, il fallait d'abord estimer les paramètres de la *Weibull*. Nous devions aussi calculer les fréquences de la vitesse du vent pour chacune des classes afin de les comparer aux probabilités théoriques estimées. Nous avons utilisé, pour chaque site, les classes de vitesses du vent collectées suivantes : $[0,1[$ m/s, $[1,2[$ m/s,..., $[24,+\infty[$ m/s.

Pour pouvoir comparer notre statistique de test à un certain quantile de la χ^2, et donc pour avoir une distribution asymptotique qui soit χ^2 avec le bon nombre de degrés de liberté (nombre de catégories$-1-$nombre de paramètres), nous avons estimé les paramètres de la *Weibull* en maximisant la vraisemblance multinômiale, i.e. celle basée sur les estimations des probabilités multinômiales des 25 classes (Rice, 2007). La probabilité que la vitesse Y appartienne à la classe allant de a à b m/s est :

$$\begin{aligned} P(Y \in [a,b[) &= P(Y < b) - P(Y < a) \\ &= F_{k,\lambda}(b) - F_{k,\lambda}(a) \\ &= \left(1 - \exp\left(-\left(\frac{b}{\lambda}\right)^k\right)\right) - \left(1 - \exp\left(-\left(\frac{a}{\lambda}\right)^k\right)\right) \\ &= \exp\left(-\left(\frac{a}{\lambda}\right)^k\right) - \exp\left(-\left(\frac{b}{\lambda}\right)^k\right). \end{aligned} \tag{4.3.1}$$

Pour trouver les estimateurs du maximum de vraisemblance de k et λ pour ce test, on devait donc maximiser la formule suivante, par rapport à ces deux mêmes paramètres :

$$\begin{aligned} \prod_{i=0}^{N_{10}} p_i &= \prod_{j=1}^{25} p_j^{n_j}, \text{ j=1,...,25 le nombre de classes} \\ &= \left(\exp\left(-\left(\frac{0}{\lambda}\right)^k\right) - \exp\left(-\left(\frac{1}{\lambda}\right)^k\right)\right)^{n_1} \cdot \ldots \cdot \\ &\quad \left(\exp\left(-\left(\frac{24}{\lambda}\right)^k\right)\right)^{n_{25}} \end{aligned}$$

où les n_j représentent le nombre de vitesses de vent collectées aux dix minutes contenues dans chaque classe j. Ainsi, n_{25} contient le nombre d'observations supérieures ou égales à 24 m/s.

Pour maximiser cette vraisemblance, nous avons utilisé un algorithme basé sur une application de la méthode de Nelder et Mead (1965). La méthode de Nelder et Mead est un algorithme d'optimisation non-linéaire cherchant à minimiser une fonction continue dans un espace à plus d'une dimension. C'est la fonction *optim* du progiciel *Stats* (voir le lien internet en référence), du logiciel R qui nous a permis cette optimisation. On pouvait donc obtenir les estimateurs $\hat{k}$ et $\hat{\lambda}$, que nous avons remplacés dans la formule (4.3.1) afin d'obtenir les probabilités estimées de se retrouver dans chaque classe, les $\hat{p}_j$, j=1,...,25.

Il ne reste, par la suite, qu'à comparer à partir du test du Khi-deux les probabilités estimées pour chaque classe aux probabilités mesurées (équivalentes aux fréquences mesurées de la vitesse du vent pour chaque classe respective).

Pour faire le test, on calcule la statistique du χ^2, définie de la façon suivante :

$$T = \sum_{j=1}^{25} \frac{((N_{10}+1)f_j - (N_{10}+1)\hat{p}_j)^2}{(N_{10}+1)\hat{p}_j}, \tag{4.3.2}$$

où $(N_{10}+1)$ représente toujours le nombre total de vitesses de vent disponibles (la longueur de la série chronologique sans les observations manquantes), les f_j représentent les probabilités empiriques (fréquences pour chaque classe), et les $\hat{p}_j$, les probabilités théoriques estimées à partir des paramètres trouvés avec le maximum de vraisemblance.

L'hypothèse nulle pour le test est l'hypothèse selon laquelle les probabilités que la vitesse prenne les valeurs dans les classes 1 à 25 proviennent d'une loi *Weibull*. La statistique trouvée en (4.3.2) suit asymptotiquement, sous l'hypothèse nulle et sous l'hypothèse que les données sont i.i.d., une loi du χ^2 à $25 - 1 - 2 = 22$ degrés de liberté. On rejettera donc l'hypothèse nulle si $T > \chi^2_{22;0,05} = 33{,}924$ pour un test avec niveau de significativité de 5%.

On appliquera ce test pour tous les sites afin de vérifier si la modélisation de la vitesse des vents sur toute la période disponible, par la *Weibull* est appropriée.

4.3.1.1. *Résultats*

TABLEAU 4.1. Résultats aux tests du Khi-deux

Site	N_{10}	Statistique de test	Valeur-p	Rejet de H_0 (Oui/non)
1	444 777	6 956,84	<0,001	Oui
2	281 117	5 766,37	<0,001	Oui
3	222 688	2 553,78	<0,001	Oui
4	320 903	5 563,79	<0,001	Oui
5	49 689	230,63	<0,001	Oui
6	178 278	2 714,39	<0,001	Oui
7	261 566	51 530,94	<0,001	Oui
8	51 979	262,71	<0,001	Oui
9	275 917	4 731,91	<0,001	Oui
10	243 024	3 556,63	<0,001	Oui
11	126 242	31 377,06	<0,001	Oui
12	51 980	1 755,10	<0,001	Oui
13	195 796	1 411,86	<0,001	Oui
14	264 175	4 667,13	<0,001	Oui
15	159 580	2 096,91	<0,001	Oui
16	174 216	7 917,83	<0,001	Oui
17	53 629	1 004,43	<0,001	Oui
18	48 507	1 187,39	<0,001	Oui
19	63 246	2 033,06	<0,001	Oui
20	61 271	3 637,26	<0,001	Oui
21	116 714	765,06	<0,001	Oui
22	51 626	1 189,15	<0,001	Oui
23	118 120	7 280,38	<0,001	Oui
24	97 640	5 625,07	<0,001	Oui
25	35 758	2 428,91	<0,001	Oui
26	14 085	192,34	<0,001	Oui
27	21 289	739,56	<0,001	Oui
28	52 082	710,42	<0,001	Oui
29	83 844	2 221,16	<0,001	Oui
30	122 229	4 310,35	<0,001	Oui
31	207 621	3 211,29	<0,001	Oui

4.3.1.2. *Discussion*

On peut voir dans le tableau 4.1 que l'hypohèse nulle est largement rejetée dans tous les cas. Effectivement, rappelons que les statistiques de test présentées dans ce tableau (comprises entre 192,34 et 51 530,94) sont comparées à la valeur 33,92 pour déterminer du rejet ou non de H_0. C'est donc dire qu'on

rejette l'hypothèse que les observations sont indépendantes et identiquement distribuées selon une *Weibull* dont les paramètres sont fixes sur toute la période et que les probabilités que la vitesse prenne les valeurs dans les classes 1 à 25 ne proviendraient pas d'une loi *Weibull* à paramètres fixes sur la période, selon ces tests. Par contre, il faut réaliser qu'avec une taille d'échantillon aussi grande (entre 14 085 et 444 777 observations), la puissance du test est énorme pour détecter même de petites différences. Il est aussi possible que les rejets de H_0 soient dus à la variabilité de la distribution de la vitesse du vent dans le temps (par exemple, d'un mois à l'autre ou d'une année à l'autre). En effet, les paramètres pourraient changer davantage dans le temps que ce à quoi l'on peut s'attendre en utilisant des données indépendantes identiquement distribuées sur toute la période, ce que nous ne pourrions pas remarquer en ne modélisant la distribution de la vitesse du vent que de façon globale sur tout $i = 0,...,N_{10}$. La modélisation serait donc inadéquate non pas à cause de l'utilisation d'une *Weibull* mais plutôt parce qu'on ne considère pas la variation de la distribution dans le temps.

4.4. Variation des paramètres de la *Weibull* en fonction du temps

Comme nous désirons, entre autres, évaluer l'incertitude reliée à l'estimation de la production d'énergie à long terme, il serait intéressant de porter une attention particulière à la variation des paramètres de la *Weibull* ajustée sur la distribution des vitesses du vent en fonction du temps. En effet, une grande variation des paramètres pourrait signifier que l'on doive ajuster une *Weibull* différente chaque année ou chaque mois, puisque la distribution change trop dans le temps, ou encore considérer une modélisation plus complexe où les paramètres pourraient changer dans le temps. Présentement, les ingénieurs utilisent une même distribution *Weibull*, laquelle a été trouvée en ajustant sur le jeu de données de vitesses de vent aux dix minutes complet une *Weibull* à partir des estimateurs du maximum de vraisemblance des paramètres de forme et d'échelle. Voyons maintenant comment l'on pourrait justifier l'utilisation de *Weibull* séparées pour chaque année ou chaque mois.

4.4.1. Comparaison de la variance des paramètres globaux, annuels et mensuels

On pense que les paramètres des *Weibull* ajustées sur diverses années ou même divers mois varient de façon assez considérable. Serait-il préférable de

modéliser de façon séparée chaque distribution de vent annuelle ou mensuelle ? Si nous ajustons à chaque début de nouvelle année des paramètres de *Weibull* différents, on peut obtenir des estimations séparées de la variance des estimateurs pour les paramètres. Il en va de même pour chaque mois. On se demande donc si les paramètres changent beaucoup chaque an/mois et si oui, cela affectera certainement la variance asymptotique des paramètres. Si non, on risque de voir des variances assez semblables d'années en années ou de mois en mois.

Afin de vérifier si l'ajustement de *Weibull* séparées est justifié, nous allons donc comparer la variance des paramètres globaux, des paramètres provenant de distributions annuelles et celle des paramètres provenant de distributions mensuelles. Pour les paramètres mensuels ou annuels, on pourra calculer la variance à partir de deux estimateurs différents, soient la variance échantillonnale et la variance asymptotique estimée. En ce qui concerne les paramètres globaux, comme ils sont estimés sur la distribution complète et que l'on ne possède donc qu'une paire de paramètres, on ne pourra qu'estimer la variance asymptotique.

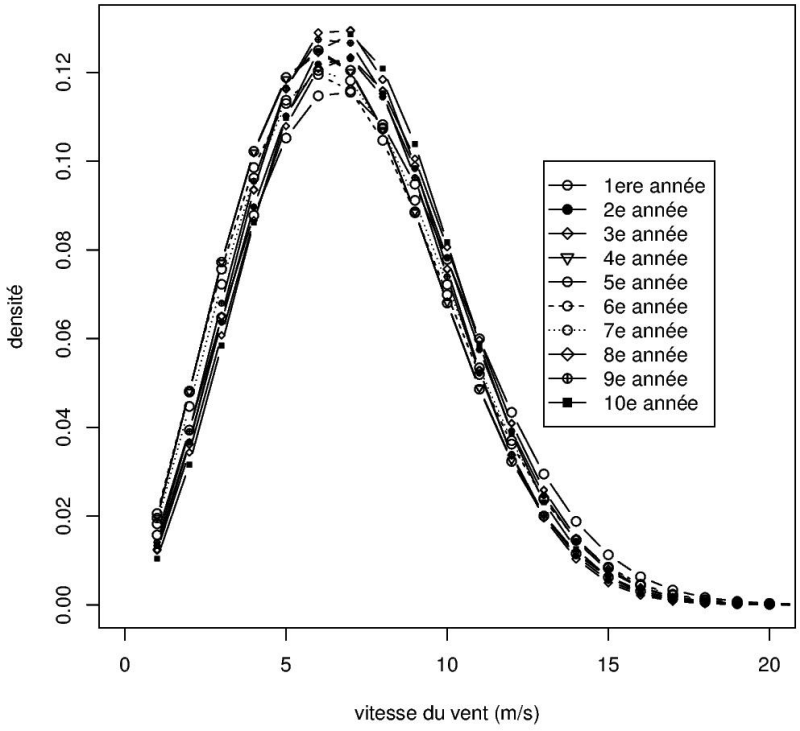

FIGURE 4.3. Ajustements de lois *Weibull* pour plusieurs années, au site 1

On peut retrouver à la figure 4.3 les diverses fonctions de densité des *Weibull* ajustées sur les distributions de vitesse du vent annuelles au site 1, à l'aide des estimateurs du maximum de vraisemblance des deux paramètres. La première année est l'année 2003 et la dixième, 2012.

Voyons d'abord la notation utilisée pour les données aux dix minutes regroupées par mois ou par année et pour les paramètres qui seront ajustés sur chaque distribution annuelle ou mensuelle. Ici encore, on utilise les données de vitesse du vent aux dix minutes de l'anémomètre 1. On notera maintenant $Y_{A_{i,j,l}}$ la donnée de vitesse de vent aux dix minutes collectée durant l'année i, $i = 1,...,A$, pour le j^e mois de l'année i, $j = 1,...,12$ et pour ce mois de cette année précise, durant le l^e dix minutes, $l = 1,...,N_{i,j}$. On comprend donc que $N_{i,j}$ est le nombre de vitesses de vent aux dix minutes collectées durant le j^e mois de l'année i.

Ensuite, afin de faciliter la notation des données ayant été collectées durant un même mois, on considérera sans perte de généralité qu'il y a pour chaque site un nombre M de mois sur lesquels les vitesses aux dix minutes s'étendent. On notera M_j l'indice des mois lorsqu'on ne tient pas compte de l'année, $M_j = 1,...,M$. On pourra donc faire correspondre une certaine combinaison des $(A_{i,j})$ à l'indice M_j et l'indice des dix minutes, $l = 1,...,N_{i,j}$ deviendra maintenant $l = 1,...,N_{M_j}$ de sorte que le nombre de données aux dix minutes qui soient contenues dans le mois M_j soit maintenant noté N_{M_j}.

On fait ensuite l'hypothèse que les $Y_{M_j,1},...,Y_{M_j,N_{M_j}}$, i.e. les données aux dix minutes collectées durant le mois M_j, sont i.i.d. et distribuées selon une *Weibull*$(\boldsymbol{\theta}_{M_j})$ ou encore, de façon équivalente, une *Weibull*(λ_{M_j}, k_{M_j}). On estimera ces deux derniers paramètres à partir de la méthode du maximum de vraisemblance, tels que vus à la section 4.2, afin d'obtenir $\hat{\boldsymbol{\theta}}_{M_j}$.

Comme on possède plusieurs estimateurs (un pour chaque mois), on peut d'abord calculer la variance échantillonnale des estimateurs. Or, si l'on veut demeurer cohérent, on doit calculer la variance échantillonnale des paramètres de forme et d'échelle multipliés par la racine du nombre d'observations, afin de pouvoir comparer les diverses estimations de variance (entre autres parce que les paramètres annuels ou globaux ont été ajustés sur un nombre différent d'observations et que l'estimation de leur variance asymptotique tient compte du nombre d'observations ; on doit donc réajuster pour cela). De plus, si on pense seulement aux paramètres mensuels, on fait l'hypothèse que les données sont i.i.d. à l'intérieur d'un mois puisqu'on ajuste sur ces données une distribution *Weibull* en croyant que chacune des vitesses de vent proviendrait

d'une telle distribution. Puis, si les paramètres sont identiques d'un mois à l'autre et que la taille de l'échantillon est la même, alors les $(\hat{\lambda}_{M_j}, \hat{k}_{M_j})$ sont i.i.d. Nous devons donc absolument utiliser le même nombre d'observations pour chaque mois où l'on ajustera une *Weibull* aux vitesses de vent mesurées, mais chaque mois ne contient malheureusement pas le même nombre d'observations à cause de données manquantes ou du nombre de jours qui varie entre les mois. Nous avons donc décidé de faire un tirage aléatoire sans remise de q vitesses de vent aux dix minutes dans chaque mois et d'utiliser ces q données pour l'ajustement des *Weibull* mensuelles, plutôt que les N_{M_j} données disponibles variant pour chaque mois. Certains des mois présentant plusieurs vitesses de vent manquantes, q a été défini comme suit :

$$q = \min\{N_{M_j} | M_j \in 1,...,M \text{ et } N_{M_j} > \frac{1}{2} \times 6 \times 24 \times 31\}.$$

Nous avons donc choisi, parmi les mois où au moins la moitié des données étaient disponibles, le mois où il y avait le moins de données de vitesses de vent aux dix minutes disponibles. Ainsi, sans perte de généralité, notez que $N_{M_j} = q$ vitesses de vent aux dix minutes ont été utilisées pour l'ajustement de *Weibull* mensuelles et pour les estimations de variance asymptotique ou échantillonnale des paramètres, et que les mois où moins de q vitesses de vent ont été mesurées n'ont pas été pris en compte dans les calculs.

Définissons maintenant les variances échantillonnales que nous calculerons pour les paramètres mensuels :

$$\hat{V}_{ech,\hat{k}_M} = \frac{q}{M-1} \sum_{j=1}^{M} (\hat{k}_{M_j} - \overline{\hat{k}}_M)^2,$$

$$\hat{V}_{ech,\hat{\lambda}_M} = \frac{q}{M-1} \sum_{j=1}^{M} (\hat{\lambda}_{M_j} - \overline{\hat{\lambda}}_M)^2,$$

où

$$\overline{\hat{k}}_M = \frac{1}{M} \sum_{j=1}^{M} \hat{k}_{M_j} \text{ et}$$

$$\overline{\hat{\lambda}}_M = \frac{1}{M} \sum_{j=1}^{M} \hat{\lambda}_{M_j}.$$

De plus, si q est grand, on a aussi que :

$$\hat{\boldsymbol{\theta}}_{M_j} \approx \mathbb{N}\left(\boldsymbol{\theta}_{M_j}, \frac{\mathbb{I}^{-1}_{M_j}(\boldsymbol{\theta}_{M_j})}{q}\right)$$

$$\Leftrightarrow \hat{\boldsymbol{\theta}}_{M_j}\sqrt{q} \approx \mathbb{N}\left(\boldsymbol{\theta}_{M_j}\sqrt{q}, \mathbb{I}_{M_j}^{-1}(\boldsymbol{\theta}_{M_j})\right)$$
$$\approx \mathbb{N}\left(\boldsymbol{\theta}_{M_j}\sqrt{q}, \mathbb{I}_{obs,M_j}^{-1}\right), \text{ où j=1,...,M,}$$

où l'information de Fisher est définie de façon générale comme :

$$\mathbb{I}_{M_j}(\boldsymbol{\theta}_{M_j}) = E\left[\left(\frac{\partial \log f(\mathbf{Y}_{M_j};\boldsymbol{\theta}_{M_j})}{\partial \boldsymbol{\theta}_{M_j}}\right)^2\right]. \tag{4.4.1}$$

L'information de Fisher observée est dénotée $\mathbb{I}_{obs,M_j}$ et est obtenue à partir de $\mathbb{I}_{M_j}(\hat{\boldsymbol{\theta}}_{M_j})$ où $\hat{\boldsymbol{\theta}}_{M_j}$ est l'estimateur du maximum de vraisemblance des paramètres de la *Weibull* $\boldsymbol{\theta}_{M_j}$ calculé à partir de $\mathbf{Y}_{M_j}$ le vecteur composé des observations sur lesquelles on ajuste une distribution, donc les vitesses de vent aux dix minutes du j^e mois dans notre cas. On écrit donc l'estimateur de la variance asymptotique des paramètres mensuels (multipliés par la racine carrée de q) de la façon suivante :

$$\hat{\mathbf{V}}_{asy,M_j} = \mathbb{I}_{obs,M_j}^{-1} = \begin{pmatrix} \widehat{\mathbb{V}\text{ar}(\hat{k}_{M_j})} & \widehat{\mathbb{C}\text{ov}(\hat{k}_{M_j},\hat{\lambda}_{M_j})} \\ \widehat{\mathbb{C}\text{ov}(\hat{\lambda}_{M_j},\hat{k}_{M_j})} & \widehat{\mathbb{V}\text{ar}(\hat{\lambda}_{M_j})} \end{pmatrix}. \tag{4.4.2}$$

On utilisera en fait la moyenne des variances asymptotiques mensuelles comme valeur de variance des paramètres mensuels à comparer aux autres mesures de variance des paramètres :

$$\bar{\hat{\mathbf{V}}}_{asy,M} = \frac{1}{M}\sum_{j=1}^{M}\hat{\mathbf{V}}_{asy,M_j}.$$

Notez que la variance ci-haut est une matrice 2x2 et que nous n'utiliserons que les entités de la diagonale, qui représentent respectivement les estimations de la variance des estimateurs du paramètre k et du paramètre λ ayant été multipliés par la racine carrée de q.

On fera la même chose avec les données aux dix minutes de chaque année afin de comparer les variances échantillonnales et asymptotiques annuelles aux variances mensuelles ou globales. On considère maintenant $Y_{A_i,1},...,Y_{A_i,N_{A_i}}$ où $A_i = 1,...,A$ est l'année sur laquelle des données aux dix minutes sont collectées pour un site particulier et N_{A_i}, le nombre de vitesses de vent aux dix minutes collectées durant l'année i, sans perte de généralité, et on note que pour $Y_{A_i,m}$ la m^e vitesse de vent aux dix minutes récoltée durant l'année i, correspond une combinaison des indices (j,l) de sorte que les deux notations soient équivalentes.

On fait ensuite l'hypothèse que les $Y_{A_i,1},...,Y_{A_i,N_{A_i}}$, i.e. les données aux dix minutes collectées durant l'année A_i, sont i.i.d. et distribuées selon une *Weibull*($\boldsymbol{\theta}_{A_i}$) ou encore une *Weibull*(k_{A_i},λ_{A_i}). On estime ces deux paramètres à partir des estimateurs du maximum de vraisemblance, toujours tels que vus à la section 4.2, et l'on obtient $\hat{\boldsymbol{\theta}}_{A_i}$. On estime d'abord la variance à partir de la variance échantillonnale des paramètres annuels multipliés par $\sqrt{N_{A_i}}$, le nombre d'observations disponibles pour chaque année. Or, comme dans le cas des mois, certaines années présentent plusieurs données manquantes et nous avons besoin d'estimer les paramètres sur des distributions de même taille. Nous avons donc décidé, ici encore, de faire un tir aléatoire sans remise de p vitesses de vent aux dix minutes dans chaque année et d'utiliser ces p données pour l'ajustement des *Weibull* annuelles, plutôt que les N_{A_i} données disponibles. Cette fois, p a été défini de la façon suivante :

$$p = min\{N_{A_i} | A_i \in 1,...,A \text{ et } N_{A_i} > \frac{1}{2} \times 6 \times 24 \times 365\}.$$

On peut donc calculer la variance échantillonnale des paramètres annuels multipliés par $\sqrt{p}$ de la façon suivante :

$$\hat{V}_{ech,\hat{k}_A} = \frac{p}{A-1} \sum_{i=1}^{A} (\hat{k}_{A_i} - \overline{\hat{k}}_A)^2,$$

$$\hat{V}_{ech,\hat{\lambda}_A} = \frac{p}{A-1} \sum_{i=1}^{A} (\hat{\lambda}_{A_i} - \overline{\hat{\lambda}}_A)^2,$$

où

$$\overline{\hat{k}}_A = \frac{1}{A} \sum_{i=1}^{A} \hat{k}_{A_i} \text{ et}$$

$$\overline{\hat{\lambda}}_A = \frac{1}{A} \sum_{i=1}^{A} \hat{\lambda}_{A_i}.$$

Si p est grand, on a que :

$$\hat{\boldsymbol{\theta}}_{A_i} \approx \mathbb{N}\left(\boldsymbol{\theta}_{A_i}, \frac{\mathbb{I}_{A_i}^{-1}(\boldsymbol{\theta}_{A_i})}{p}\right)$$

$$\Leftrightarrow \hat{\boldsymbol{\theta}}_{A_i}\sqrt{p} \approx \mathbb{N}\left(\boldsymbol{\theta}_{A_i}\sqrt{p}, \mathbb{I}_{A_i}^{-1}(\boldsymbol{\theta}_{A_i})\right)$$

$$\approx \mathbb{N}\left(\boldsymbol{\theta}_{A_i}\sqrt{p}, \mathbb{I}_{obs,A_i}^{-1}\right), \text{ où i=1,...,A,}$$

où on redéfinit l'information de Fisher et l'information de Fisher observée de la même façon qu'en (4.4.1) et (4.4.2) mais pour les paramètres annuels. On

utilisera, tout comme pour la variance mensuelle, la moyenne des variances asymptotiques annuelles comme valeur de variance asymptotique des paramètres annuels à comparer aux autres mesures de variance des paramètres :

$$\bar{\hat{\mathbf{V}}}_{asy,A} = \frac{1}{A}\sum_{i=1}^{A} \hat{\mathbf{V}}_{asy,A_i}.$$

Notez encore une fois que la variance ci-haut est une matrice 2x2 et que nous n'utiliserons que les entités de la diagonale, qui représentent respectivement les estimations de la variance des estimateurs du paramètre k et du paramètre λ ayant été multipliés par la racine carrée de p.

Finalement, aux variances asymptotiques et échantillonnales mensuelles ou annuelles, nous allons comparer la variance asymptotique globale estimée, soit celle des paramètres ajustés sur la distribution complète des vitesses de vent aux dix minutes. On fait l'hypothèse que toutes les vitesses de vent aux dix minutes $Y_{A_{i,j,l}}$, pour $A_i = 1,...,A$, $j = 1,...,12$, $l = 1,...,N_{i,j}$, sont i.i.d. *Weibull*(k_G,λ_G). On estime les deux paramètres, toujours à partir des estimateurs du maximum de vraisemblance, et l'on obtient $\hat{\boldsymbol{\theta}}_G$. On fait encore une fois l'hypothèse de normalité suivante pour les paramètres de la distribution globale :

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_G &\approx \mathbb{N}\left(\boldsymbol{\theta}_G, \frac{\mathbb{I}_G^{-1}(\boldsymbol{\theta}_G)}{N_{10}+1}\right) \\
\Leftrightarrow \hat{\boldsymbol{\theta}}_G\sqrt{N_{10}+1} &\approx \mathbb{N}\left(\boldsymbol{\theta}_G\sqrt{N_{10}+1}, \mathbb{I}_G^{-1}(\boldsymbol{\theta}_G)\right) \\
&\approx \mathbb{N}\left(\boldsymbol{\theta}_G\sqrt{N_{10}+1}, \mathbb{I}_{obs,G}^{-1}\right),
\end{aligned}$$

où l'information de Fisher globale et l'information de Fisher observée globale sont toujours définies de la même façon que précédemment mais $\mathbf{Y}_G$ devient le vecteur composé de toutes les vitesses de vent aux dix minutes qu'on a pu collecter à l'anémomètre 1. On écrit donc l'estimateur de la variance asymptotique des paramètres globaux de la façon suivante :

$$\hat{\mathbf{V}}_{asy,G} = \mathbb{I}_{obs,G}^{-1} = \begin{pmatrix} \widehat{\mathbb{V}\text{ar}(\hat{k}_G)} & \widehat{\mathbb{C}\text{ov}(\hat{k}_G,\hat{\lambda}_G)} \\ \widehat{\mathbb{C}\text{ov}(\hat{\lambda}_G,\hat{k}_G)} & \widehat{\mathbb{V}\text{ar}(\hat{\lambda}_G)} \end{pmatrix},$$

où l'on utilise encore une fois les entités de la diagonale comme estimations de la variance.

On obtient donc finalement cinq mesures de la variance pour chaque paramètre, soient les éléments en première ligne et première colonne des matrices $\hat{\mathbf{V}}_{asy,G}$, $\bar{\hat{\mathbf{V}}}_{asy,A}$, $\tilde{\hat{\mathbf{V}}}_{asy,M}$, ainsi que les mesures $\hat{V}_{ech,\hat{k}_A}$ et $\hat{V}_{ech,\hat{k}_M}$ pour la variance

de $\hat{k}$ et les éléments en dernière ligne et dernière colonne des matrices $\hat{\mathbf{V}}_{asy,G}$, $\bar{\hat{\mathbf{V}}}_{asy,A}$, $\bar{\hat{\mathbf{V}}}_{asy,M}$, ainsi que les valeurs de $\hat{V}_{ech,\hat{\lambda}_A}$ et $\hat{V}_{ech,\hat{\lambda}_M}$ pour la variance de $\hat{\lambda}$. Les tableaux suivants présentent la comparaison des variances pour chaque site, de façon séparée pour chaque paramètre.

4.4.1.1. *Résultats*

TABLEAU 4.2. Comparaison des diverses mesures de variance du paramètre de forme k

Site	p (ans)	q (mois)	$\hat{V}_{asy,G}$	$\bar{\hat{V}}_{asy,A}$	$\bar{\hat{V}}_{asy,M}$	$\hat{V}^1_{ech,\hat{k}_A}$	$\hat{V}_{ech,\hat{k}_M}$
1	24 869	2 383	3,46	3,52	4,37	363,89	240,31
2	26 052	2 549	1,45	1,49	1,81	152,11	238,73
3	13 753	3 900	2,95	2,95	3,51	51,78	273,92
4	31 524	2 346	3,12	3,17	3,89	258,55	239,76
5	5 986	2 376	2,88	3,80	4,12	355,35	243,24
6	43 245	2 367	3,60	3,73	4,50	541,94	169,42
7	14 974	2 352	3,37	3,44	4,30	237,36	297,18
8	9 085	3 765	2,96	3,26	3,38	263,44	266,33
9	25 850	3 243	3,07	3,09	3,79	144,39	302,96
10	13 324	3 378	3,60	3,82	4,62	197,68	419,13
11	14 865	3 594	5,05	4,89	6,19	1338,85	987,30
12	4 240	3 796	3,95	4,85	5,05	422,92	392,35
13	35 148	3 281	1,80	1,80	2,19	546,99	341,60
14	15 096	2 372	2,07	2,12	3,06	59,99	372,67
15	16 731	2 526	2,78	2,82	3,33	130,63	189,72
16	27 375	2 958	2,67	2,69	3,29	91,52	134,10
17	24 572	2 664	1,76	1,80	2,17	5,64	40,77
18	14 591	2 310	3,70	4,01	5,31	377,87	345,36
19	24 048	3 918	2,03	2,06	2,53	17,96	672,57
20	21 888	4 104	2,36	2,41	3,06	22,70	659,31
21	24 934	4 310	2,42	2,53	3,19	476,29	101,76
22	19 378	2 862	2,56	2,79	3,27	1 303,42	212,57
23	19 416	2 446	2,85	2,93	3,22	121,39	167,41
24	16 674	2 697	4,15	4,58	5,05	1 459,22	441,34
25	9 519	4 011	2,13	2,46	3,21	1 050,07	283,70
26	14 085	4 120	2,21	2,21	2,65	-	86,23
27	21 289	2 649	2,29	2,29	2,48	-	44,85
28	12 594	3 941	2,62	2,92	3,15	541,29	230,88
29	10 944	2 448	3,55	4,00	4,71	46,79	165,37
30	29 423	2 314	2,79	2,78	3,78	206,07	171,42
31	25 339	2 508	2,15	2,19	2,63	109,79	146,26

[1]Les cases vides sont les sites où il n'y a pas assez d'observations pour obtenir une variation annuelle.

TABLEAU 4.3. Comparaison des diverses mesures de variance du paramètre d'échelle λ

Site	p (ans)	q (mois)	$\hat{V}_{asy,G}$	$\bar{\hat{V}}_{asy,A}$	$\bar{\hat{V}}_{asy,M}$	$\hat{V}^{1}_{ech,\hat{\lambda}_A}$	$\hat{V}_{ech,\hat{\lambda}_M}$
1	24 869	2 383	12,04	11,97	12,03	1 201,59	1 791,02
2	26 052	2 549	37,61	37,36	34,59	9 353,24	8 456,28
3	13 753	3 900	15,16	15,46	13,93	1 287,72	4 657,65
4	31 524	2 346	11,48	11,32	11,42	1 765,99	1 636,82
5	5 986	2 376	15,50	14,35	15,53	1 804,03	4 864,89
6	43 245	2 367	11,17	10,67	11,17	380,03	1 064,72
7	14 974	2 352	9,18	9,23	8,49	1 712,51	1 609,66
8	9 085	3 765	16,96	15,40	16,81	567,33	2 724,53
9	25 850	3 243	16,60	16,53	15,65	1 392,49	3 191,62
10	13 324	3 378	14,68	14,50	12,79	4 636,30	5 784,21
11	14 865	3 594	9,06	9,20	8,36	4 119,81	24 00,78
12	4 240	3 796	7,86	6,96	6,96	446,14	3 217,10
13	35 148	3 281	24,20	24,02	22,23	13 121,06	6 306,30
14	15 096	2 372	34,17	34,46	28,06	6 703,15	15 636,58
15	16 731	2 526	17,28	17,51	16,79	3 521,49	2 046,93
16	27 375	2 958	10,98	10,99	10,03	751,75	2 788,72
17	24 572	2 664	10,33	10,31	9,30	7 907,32	2 945,67
18	14 591	2 310	15,89	15,09	17,14	16 330,78	377,87
19	24 048	3 918	13,92	14,08	13,48	5 030,83	2 463,52
20	21 888	4 104	12,92	13,11	12,23	5 978,81	2 858,05
21	24 934	4 310	27,21	25,80	24,33	13 871,73	6 560,63
22	19 378	2 862	17,39	16,79	16,30	4 050,34	3 390,51
23	19 416	2 446	16,75	16,44	18,26	145,22	575,92
24	16 674	2 697	11,23	10,76	11,81	33,70	967,63
25	9 519	4 011	21,62	20,61	17,95	16 200,89	13 982,28
26	14 085	4 120	35,42	35,42	31,14	-	16 113,60
27	21 289	2 649	35,43	35,43	32,89	-	6 129,90
28	12 594	3 941	23,06	22,68	21,61	13 283,79	5 596,19
29	10 944	2 448	11,27	11,03	10,01	1 980,34	2 347,37
30	29 423	2 314	14,80	14,83	15,13	1 015,86	2 285,58
31	25 339	2 508	11,00	10,79	10,65	338,92	1 820,69

[1]Les cases vides sont les sites où il n'y a pas assez d'observations pour obtenir une variation annuelle.

4.4.1.2. *Discussion*

Suite aux résultats obtenus aux tests du Khi-deux précédemment, on pense que la distribution de la vitesse du vent serait trop variable dans le temps pour n'être modélisée que de façon globale. On a donc voulu comparer diverses mesures de variance afin de vérifier si les paramètres annuels ou mensuels varient

davantage que ce à quoi l'on pourrait s'attendre en se basant sur la variance de ces mêmes paramètres pour la distribution globale. En regardant les tableaux 4.2 et 4.3, on a pu voir tant pour le paramètre de forme que d'échelle, que la variance échantillonnale explose par rapport aux variances asymptotiques estimées. De plus, les variances asymptotiques globales, annuelles et mensuelles sont toutes assez semblables alors que la différence entre les variances échantillonnales annuelles et mensuelles est davantage marquée.

Ainsi, comme on voit une grosse différence entre les variances échantillonnales et asymptotiques, on se demande maintenant si cela peut réellement nous permettre de conclure que les paramètres varient de façon considérable (trop pour n'ajuster qu'une seule *Weibull* sur la distribution de la vitesse du vent) ou s'il n'y a pas une autre raison qui nous échappe, qui mènerait à de tels résultats, aussi différents entre les estimateurs de la variance. Prenons par exemple le site 1, où la variance asymptotique du paramètre de forme annuel multiplié par la racine de p est d'environ 3 et la variance échantillonnale du paramètre de forme annuel multiplié par la racine de p, plutôt de l'ordre de 300 (et on trouve à peu près le même rapport pour les variances du paramètre d'échelle à ce site). Comme on a utilisé partout le même nombre d'observations pour l'ajustement d'une *Weibull* (p dans le cas des paramètres annuels), on voit que même en prenant l'échelle originale et en divisant par p les variances obtenues, on obtient des variances échantillonnales et asymptotiques pour $\hat{k}$ de l'ordre de $\frac{300}{p}$ et $\frac{3}{p}$ respectivement, qui sont toujours très différentes. On se demande donc si cette différence est trop importante pour qu'il ne s'agisse que de la variabilité des paramètres dans le temps, et si cela pourrait être dû à la présence de dépendance entre les observations, menant à des estimateurs qui seraient davantage variables. Effectivement, l'on pourrait peut-être s'attendre à des variances très différentes dans le cas où l'on estimerait la variance de paramètres ajustés sur une distribution de données dépendantes par rapport à indépendantes, mais il ne s'agit que d'une hypothèse pour le moment. Pour vérifier cette hypothèse ainsi que celle que les paramètres varient trop pour n'utiliser que des paramètres globaux, nous allons donc faire une simulation à partir du bootstrap et des paramètres ajustés sur la distribution globale de la vitesse du vent, ainsi que les paramètres annuels et mensuels. Nous allons donc utiliser des données indépendantes entre elles, générées sous des modèles précis selon l'hypothèse à tester : notre hypothèse nulle sera que la distribution du vent devrait être modélisée par une *Weibull* avec paramètres globaux sur toute la

période (et donc que des paramètres globaux suffiraient lors de la modélisation). En ce qui a trait aux hypothèse alternatives, on aura dans un premier temps (première hypothèse alternative) que la distribution pourrait plutôt être décomposée en plusieurs distributions annuelles de la vitesse de vent qui proviennent chacune d'une *Weibull* différente. La deuxième hypothèse alternative sera plutôt que la distribution sur neuf ans peut être décomposée en 108 distributions mensuelles (neuf années multipliées par douze mois) qui proviendraient toutes de *Weibull* avec des paramètres différents. Pour voir à quels résultats mènerait l'hypothèse nulle (en faisant aussi l'hypothèse que les données sont indépendantes), nous allons procéder comme suit :

Algorithme pour la variance sous H_0

Rappelons que nous avons utilisé q=2 383 données par mois, p=24 869 données par année et 444 777 données au total au site 1 pour nos ajustements de *Weibull* (voir tableau 4.2). Sous H_0, on utilise les paramètres globaux $\hat{k}_G, \hat{\lambda}_G$ ajustés précédemment au site 1 :

(1) Simulation d'un jeu de données i.i.d de taille 444 777 à partir d'une $Weibull(\hat{k}_G,\hat{\lambda}_G)$

(2) Ajustement de paramètres globaux sur les données simulées en (1) et estimation de leur variance asymptotique

(3) Séparation des 444 777 données en neuf années et utilisation de seulement p=24 869 données par année (choisies de façon aléatoire dans les 444 777/9≈49 420 données par année disponibles)

(4) Ajustement de paramètres annuels sur chacune de ces distributions annuelles de p données et estimation de la variance échantillonnale des neuf paramètres d'échelle et des neuf paramètres de forme, ainsi que de la variance asymptotique (moyenne des variances asymptotiques de chacun des neuf paramètres d'échelle et de forme)

(5) Séparation des 444 777 données en 108 mois et utilisation de seulement q=2 383 données par mois (choisies de façon aléatoire dans les 444 777/108≈4 118 données disponibles pour chaque mois)

(6) Ajustement de paramètres mensuels sur chacune de ces distributions mensuelles de q données et estimation de la variance échantillonnale des 108 paramètres d'échelle et de forme ainsi que de la variance asymptotique (moyenne des variances asymptotiques de chacun des 108 paramètres d'échelle et de forme)

(7) Reproduire les étapes (1) à (6) 1 000 fois pour obtenir 1 000 valeurs de variance asymptotique globale, annuelle et mensuelle et le même nombre de valeurs de variance échantillonnale annuelle et mensuelle, et ce pour chacun des deux paramètres d'une *Weibull*

Nous obtenons donc 1 000 variances asymptotiques globales, annuelles, mensuelles et 1 000 variances échantillonnales annuelles et mensuelles sous l'hypothèse nulle, pour chacun des paramètres. Pour cette hypothèse, nous ne présentons pas les résultats puisque nous avons obtenu, dans tous les cas, des résultats très similaires à ceux obtenus dans les tableaux 4.2 et 4.3 pour la variance asymptotique globale, annuelle et mensuelle des paramètres (c'est-à-dire des variances tant asymptotiques qu'échantillonnales tournant autour de 3,5 pour le paramètre de forme et de 12,0 pour le paramètre d'échelle). Ainsi, en ce qui a trait aux variances asymptotiques, on ne peut pas dire que H_0 ne soit pas plausible. Par contre on voit que les variances échantillonnales (tant annuelles que mensuelles) obtenues en 4.2 et en 4.3 sont très différentes des valeurs obtenues par simulation. Dans les simulations, les variances échantillonnales tant annuelles que mensuelles sont en moyenne proches des variances asymptotiques. Allons donc voir plus loin, à partir des simulations basées sur les hypothèses alternatives, si l'on n'obtiendrait pas des valeurs similaires à celles trouvées avec les vraies valeurs de vitesse de vent.

Algorithme pour la variance sous $H_{A,1}$

Rappelons encore que nous avons utilisé q=2 383 données par mois, p=24 869 données par année et 444 777 données au total au site 1. Sous $H_{A,1}$, on utilise les paramètres annuels $\hat{k}_{A_i}$,$\hat{\lambda}_{A_i}$, i=1,...,9 ajustés précédemment au site 1 :

(1) Simulation d'un jeu de données de taille 444 777 à partir de 444 777/9$\approx$ 49 420 données simulées de chacune des distributions *Weibull*($\hat{k}_{A_i}$,$\hat{\lambda}_{A_i}$), i=1,...,9, mises bout à bout.

(2) Ajustement de paramètres globaux sur toute la distribution de 444 777 données créée en (1) et estimation de leur variance asymptotique

(3) Séparation des 444 777 données en neuf années et utilisation de seulement p=24 869 données par année (choisies de façon aléatoire dans les 49 420 données disponibles)

(4) Ajustement de paramètres annuels sur chacune de ces distributions annuelles de p données et estimation de la variance échantillonnale des neuf paramètres d'échelle et des neuf paramètres de forme, ainsi que

de la variance asymptotique (moyenne des variances asymptotiques de chacun des neuf paramètres d'échelle et de forme)

(5) Séparation des 444 777 données en 108 mois et utilisation de seulement q=2 383 données par mois (choisies de façon aléatoire dans les 4 118 données disponibles pour chaque mois)

(6) Ajustement de paramètres mensuels sur chacune de ces distributions mensuelles de q données et estimation de la variance échantillonnale des 108 paramètres d'échelle et de forme ainsi que de la variance asymptotique (moyenne des variances asymptotiques de chacun des 108 paramètres d'échelle et de forme)

(7) Reproduire les étapes (1) à (6) 1 000 de fois pour obtenir 1 000 valeurs de variance asymptotique globale, annuelle et mensuelle et le même nombre de valeurs de variance échantillonnale annuelle et mensuelle, et ce pour chacun des deux paramètres d'une *Weibull*

Ici, il est intéressant de présenter la distribution des variances échantillonnales annuelles et mensuelles. En effet, la distribution des variances asymptotiques globales, annuelles et mensuelles est demeurée semblable à celle sous H_0, c'est-à-dire que les distributions tournent encore autour de 3,5 pour le paramètre de forme et de 12,0 pour le paramètre d'échelle. On présente donc à la figure 4.4 la distribution des variances échantillonnales annuelles et mensuelles sous $H_{A,1}$, pour chacun des paramètres, ainsi que les valeurs trouvées à partir des données réelles dans les tableaux 4.2 et 4.3 (représentées sous forme de barres verticales pointillées sur les histogrammes en 4.4). On voit bien que pour les paramètres annuels, les variances échantillonnales sont, pour les deux paramètres, assez près de celles obtenues à partir des données réelles. Les valeurs réelles sont même contenues dans les distributions obtenues à partir du bootstrap. Par contre, pour les distributions de variances échantillonnales mensuelles, les valeurs obtenues plus tôt et présentées dans les tableaux 4.2 et 4.3 étaient beaucoup plus grandes que les valeurs obtenues par bootstrap. Les valeurs réelles ne sont pas contenues dans les distributions, pour les deux paramètres, et cela semble indiquer que la variance échantillonnale de nos paramètres est plus grande que ce qu'on pourrait trouver si des paramètres changeant annuellement suffisaient lors de l'ajustement de *Weibull* sur la distribution de la vitesse du vent.

Comme on vient de voir que d'ajuster des *Weibull* différentes à chaque année ne mène toujours pas à des variances mensuelles des paramètres aussi grandes

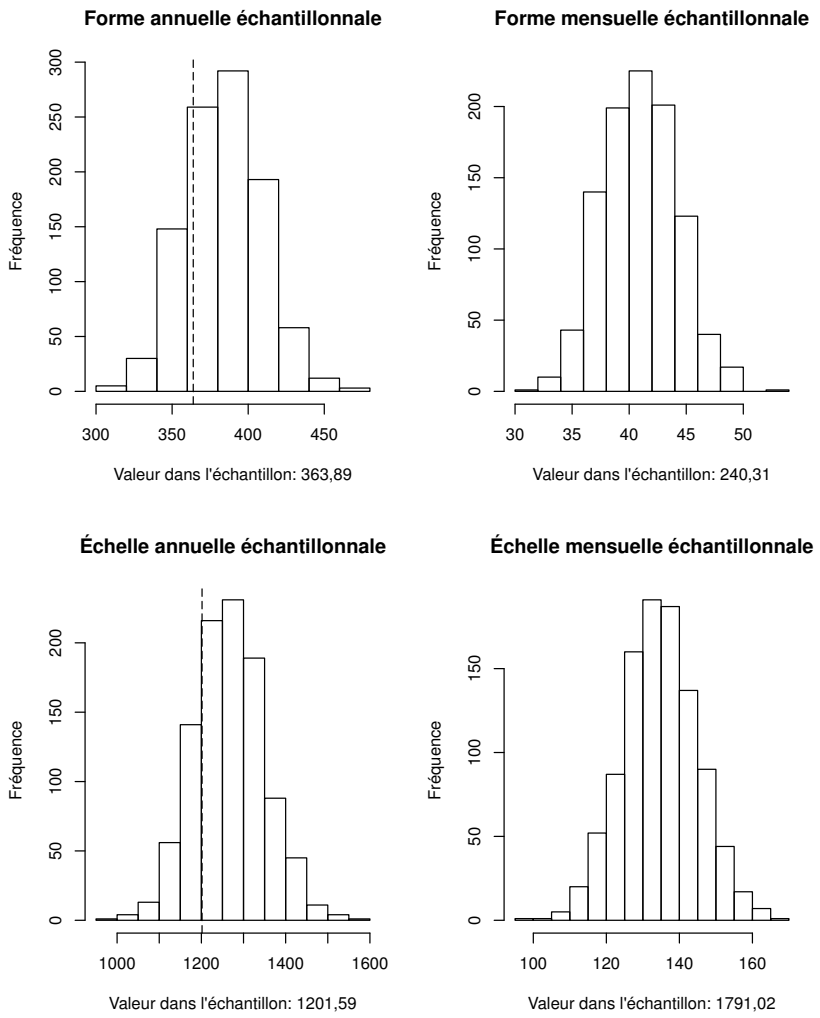

FIGURE 4.4. Distributions des 1 000 variances échantillonnales annuelles et mensuelles pour les deux paramètres de *Weibull*, provenant des ajustements sur des distributions simulées bootstrap à partir de paramètres annuels

que celles observées, on veut aussi tester l'hypothèse alternative 2, c'est-à-dire des paramètres différents devraient être ajustés sur chacune des distributions mensuelles de vitesses du vent. Pour ce faire, nous présentons maintenant l'algorithme de simulation utilisé pour vérifier si les variances obtenues sous cette

hypothèse (en faisant encore l'hypothèse que les données sont indépendantes) sont représentatives de ce qu'on a observé.

Algorithme pour la variance sous $H_{A,2}$

Rappelons toujours que nous avons utilisé q=2 383 données par mois, p=24 869 données par année et 444 777 données au total au site 1. Sous $H_{A,2}$, on utilise les paramètres mensuels $\hat{k}_{M_j}$,$\hat{\lambda}_{M_j}$, j=1,...,108 ajustés précédemment au site 1 :

(1) Simulation d'un jeu de données de taille 444 777 à partir de 444 777/108$\approx$ 4 118 données simulées de chacune des distributions *Weibull*($\hat{k}_{M_j}$,$\hat{\lambda}_{M_j}$), j=1,...,108, mises bout à bout.

(2) Ajustement de paramètres globaux sur toute la distribution de 444 777 données créée en (1) et estimation de leur variance asymptotique

(3) Séparation des 444 777 données en neuf années et utilisation de seulement p=24 869 données par année (choisies de façon aléatoire parmi les 49 420 données disponibles pour chaque année)

(4) Ajustement de paramètres annuels sur chacune de ces distributions annuelles de p données et estimation de la variance échantillonnale des neuf paramètres d'échelle et des neuf paramètres de forme, ainsi que de la variance asymptotique (moyenne des variances asymptotiques de chacun des neuf paramètres d'échelle et de forme)

(5) Séparation des 444 777 données en 108 mois et utilisation de seulement q=2 383 données par mois (choisies de façon aléatoire parmi les 4 118 données disponibles pour chaque mois)

(6) Ajustement de paramètres mensuels sur chacune de ces distributions mensuelles de q données et estimation de la variance échantillonnale des 108 paramètres d'échelle et de forme ainsi que de la variance asymptotique (moyenne des variances asymptotiques de chacun des 108 paramètres d'échelle et de forme)

(7) Reproduire les étapes (1) à (6) 1 000 fois pour obtenir 1 000 valeurs de variance asymptotique globale, annuelle et mensuelle et le même nombre de valeurs de variance échantillonnale annuelle et mensuelle, et ce pour chacun des deux paramètres d'une *Weibull*

Dans le cas de l'hypothèse alternative 2, on obtient encore des distributions des variances asymptotiques tournant autour de 3,5 pour le paramètre de forme et de 12,0 pour le paramètre d'échelle, et les estimateurs de la variance asymptotique obtenus dans les tableaux 4.2 et 4.3 sont contenus (ou très près d'être

contenus) dans les distributions des 1 000 variances trouvées par simulations. Or, en ce qui concerne les distributions des variances échantillonnales, on obtient plutôt les résultats présentés à la figure 4.5.

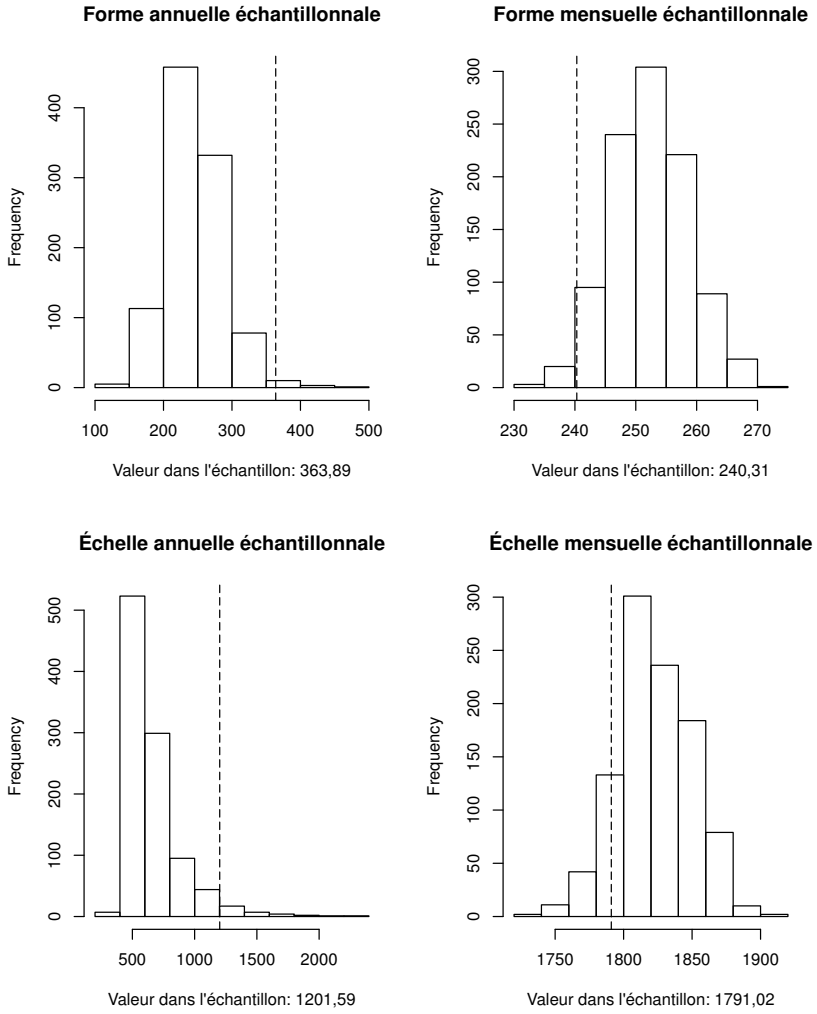

FIGURE 4.5. Distributions des 1 000 variances échantillonnales annuelles et mensuelles pour les deux paramètres de *Weibull*, provenant des ajustements sur des distributions simulées bootstrap à partir de paramètres mensuels

On voit maintenant que les valeurs obtenues plus tôt (barres pointillées dans le graphique) sont contenues dans les distributions, tant pour la distribution des variances des paramètres annuels que mensuels. Cela indique donc que, même en simulant des données indépendantes, mais à partir de distributions différentes chaque mois, on obtient des valeurs pour les variances qui sont similaires à ce qu'on a obtenu dans la réalité. Ainsi, s'il y a de la dépendance entre les observations, ce ne serait peut-être pas la cause des variances échantillonnales si différentes des variances asymptotiques (environ cent fois plus grandes).

On pense plutôt que des paramètres globaux, dans un premier temps, ne permettent pas de capturer la variabilité dans la distribution de la vitesse de vent, dans le temps. Dans un deuxième temps, on a remarqué à partir des résultats sous l'hypothèse $H_{A,1}$ que des paramètres annuels ne sembleraient pas non plus capturer toute la variabilité de la distribution de la vitesse du vent et que cette distribution varie encore davantage qu'aux années, puisque les variances échantillonnales mensuelles obtenues pour chacun des paramètres, à partir des vraies données, étaient respectivement de 240,31 et 1 791,02 alors qu'on obtenait des variances moyennes de 41,01 et 134,70 respectivement, sous $H_{A,1}$. Après avoir vu à la figure 4.5 que des paramètres variant mensuellement mènent à des résultats assez près de ceux obtenus à partir des données mesurées, on pense qu'il est possible que des paramètres changés chaque mois mènent à une meilleure modélisation de la distribution de la vitesse du vent. Bien sûr, il nous est impossible, juste avec ces résultats, de nous assurer que de tels paramètres variant mensuellement représentent la solution à tout problème de modélisation. Nous pensons aussi qu'il existe de la dépendance entre les observations, puisque la vitesse du vent à un moment précis sera probablement plus semblable à celle une heure plus tard qu'à la vitesse du vent une année plus tard, donc il est possible que les simulations ne soit pas représentatives de la réalité puisque les données simulées étaient indépendantes et faisaient donc abstraction de la structure de corrélation entre les observations. C'est pourquoi nous voulons investiguer davantage sur la dépendance entre les observations.

4.4.2. La dépendance entre les observations

Jusqu'à maintenant, nous avons traité les observations de vitesse du vent comme indépendantes. Or, de façon logique, on peut s'attendre à ce que la série des vitesses de vent à chacun des sites soit composée d'observations qui

sont dépendantes entre elles, jusqu'à un certain délai dans le temps. En effet, si l'on mesure la vitesse du vent aujourd'hui à un site particulier, et qu'on retourne la mesurer une heure plus tard, on s'attend à ce que la vitesse n'ait pas changé drastiquement sous des conditions météorologiques plutôt normales. Par contre, il est possible que si l'on retourne mesurer la vitesse du vent dans plusieurs mois au même site, la tendance du vent ait beaucoup changé et que la vitesse ne soit plus autant reliée aux mesures prises aujourd'hui. Le temps durant lequel il y a encore un lien (de la dépendance) entre les observations de vitesse du vent est ce que nous aimerions étudier dans cette section.

Pour étudier la dépendance, nous utilisons la série des vitesses de vent au site 1. Nous utilisons aussi un outil très utile quand vient le temps d'analyser des séries chronologiques : un graphique d'autocorrélation. Celui-ci nous permettra de vérifier, pour ce site, jusqu'à quel délai de temps (en heures, puis en jours) la série de vitesses du vent est-elle encore corrélée. Chaque barre présentée dans les graphiques nous indiquera la corrélation (un indice contenu entre -1 et 1), pour différents délais. Notez que la série de vitesses de vent est plutôt stationnaire, elle ne présente pas de tendance particulière dans le temps et les vitesses de vent oscillent plutôt autour de la moyenne long terme. Nous avons donc utilisé directement la série des vitesses moyennes aux heures ou aux jours sans y appliquer de transformation.

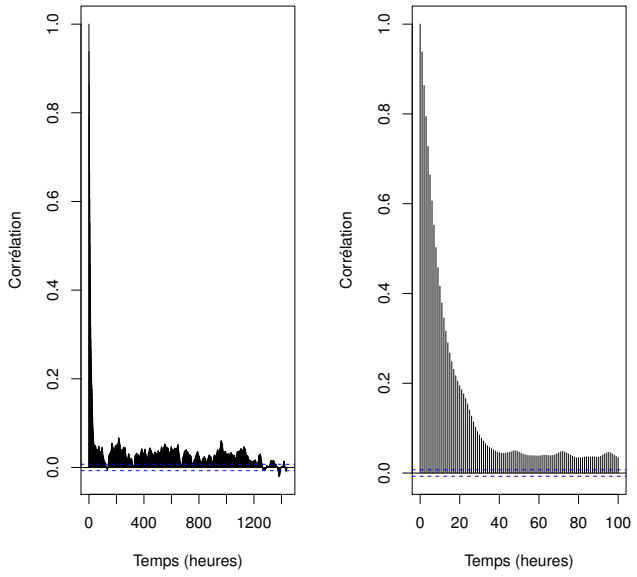

FIGURE 4.6. Graphique d'autocorrélation de la série des vitesses de vent moyennes aux heures, au site 1

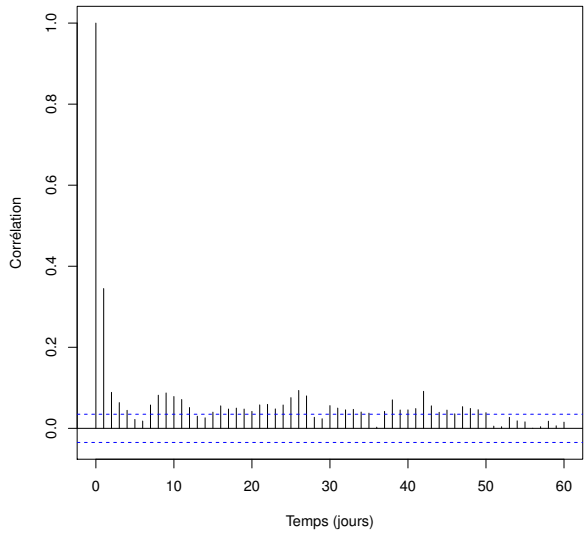

FIGURE 4.7. Graphique d'autocorrélation de la série des vitesses de vent moyennes aux jours, au site 1

En ce qui a trait à la corrélation entre les vitesses aux heures, on voit à droite de la figure 4.6 que la corrélation est bien présente, au moins jusqu'au délai 100 (près de quatre jours). Si on regarde la figure 4.7, on peut aussi voir que la corrélation entre les vitesses moyennes aux jours est significative sur un délai de plus de quatre jours (elle oscille ensuite autour de 0). Ces résultats nous aident à mieux comprendre la structure de corrélation entre les observations. De plus, cela va dans le même sens que ce que l'on croyait, c'est-à-dire qu'il existe bien de la dépendance entre les observations.

Comme nous venons de voir qu'il existe une dépendance non négligeable sur un délai d'environ trois à quatre jours, nous aimerions maintenant reproduire le test du khi-deux de la section 4.3.1 en utilisant la moyenne des vitesses de vent sur trois jours plutôt que les vitesses de vent aux 10 minutes comme en 4.3.1, pour quelques sites. Nous croyons que l'effet de la dépendance entre les observations sera amoindri en utilisant les moyennes et que les résultats du test pourraient maintenant être plus fiables. Il sera intéressant de remarquer si les résultats sont différents de ce qu'on avait avec les données aux dix minutes, et si la *Weibull* semble maintenant appropriée pour modéliser la distribution de la vitesse du vent. Nous avons donc refait le test à partir des observations moyennes aux trois jours de quelques-uns des sites présentant des nombres d'observations variables. Aux sites 18, 24, 25 et 26, l'hypothèse nulle n'a pas été rejetée pour des nombres d'observations aux trois jours allant de 32 à 226

(valeurs-p respectives de 0,996, 0,304, 0,348 et 0,921). Par contre, pour les sites 1, 15 et 30, on rejetait l'hypothèse nulle du test (valeurs-p<0,001) pour 1029, 369 et 282 observations, respectivement. Il est donc possible qu'on ne détecte pas la différence entre la densité de *Weibull* et la probabilité empirique d'être dans chaque catégorie à cause d'un trop petit nombre d'observations moyennes aux trois jours, pour les premiers sites. On pense donc toujours que la *Weibull* n'est pas complètement appropriée pour modéliser la distribution du vent, même en supprimant une portion de la dépendance entre les observations.

4.5. Comparaison du calcul de l'énergie estimée à partir de la distribution expérimentale et de l'ajustement de Weibull globale

Généralement, le calcul de la production d'énergie estimée est fait à partir des paramètres de la *Weibull* ajustée sur la distribution des vitesses du vent et d'une courbe de puissance fournie par les fabricants de turbines éoliennes. La courbe de puissance est en fait un graphe qui représente la puissance de sortie d'une éolienne à différentes vitesses de vent. On retrouve à la figure 4.8 la courbe de puissance discrétisée en intervalles de 1 m/s utilisée dans ce mémoire, reliée à une turbine éolienne classique installée par la compagnie *Hatch*, soit la turbine *Enercon E-82 E2*. Remarquez que la puissance devient nulle à partir d'une vitesse de plus de 25 m/s. Cela est dû au fait que les vitesses supérieures à 25 m/s deviennent dangereuses pour la turbine et qu'il est préférable d'arrêter les turbines éoliennes à partir d'une telle vitesse, afin d'éviter les bris de matériel.

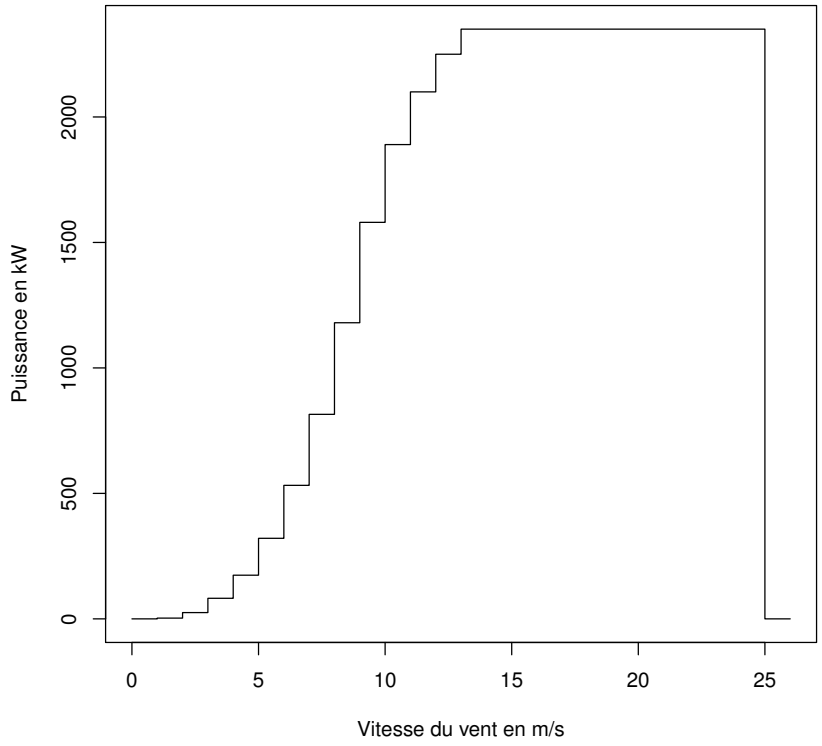

FIGURE 4.8. Courbe de puissance de la turbine éolienne E-82 E2

Le calcul de l'estimation de la production d'énergie en kW pour le nombre d'heures dans une année (donc kW/an) à partir de la courbe de puissance et des paramètres de la *Weibull* se fait de la façon suivante :

$$Prod_1 = \sum_{j=1}^{25} puis_j \cdot \hat{P}(puis_j) \cdot 365{,}25 \cdot 24$$

$$= \sum_{j=1}^{25} puis_j \cdot \exp\left[\left(-\frac{j-1}{\hat{\lambda}}\right)^{\hat{k}} - \exp\left(\left(-\frac{j}{\hat{\lambda}}\right)^{\hat{k}}\right)\right] \cdot 365{,}25 \cdot 24 \qquad (4.5.1)$$

dans le cas où l'on utilise les probabilités estimées d'être dans chaque classe. Les valeurs $puis_j$ sont données sur la figure 4.8 ; il s'agit de la puissance correspondant à la classe j, en kilowatts. Notez que le calcul s'arrête pour la classe de vitesse du vent 24-25 m/s puisque la puissance devient nulle après 25 m/s et que le reste (les vitesses de vent supérieures à 25 m/s) n'est donc pas comptabilisé dans la somme. Par contre, la *Weibull* est bel et bien ajustée sur la distribution de toutes les vitesses de vent car la probabilité d'être dans chaque

classe se doit de tenir compte des vitesses supérieures à 25 m/s. On multiplie ainsi chaque probabilité d'être dans un intervalle (par exemple, 0-1 m/s) par la puissance respective produite par cette catégorie de vitesse de vent, puis par le nombre d'heures présentes dans une année afin d'obtenir une estimation de la production d'énergie sur une année complète en kW (donc en kW/an, équivalent à kW/h multiplié par le nombre d'heures dans une année).

Notez qu'il est aussi possible d'utiliser la fonction de répartition expérimentale de la distribution de la vitesse du vent mesurée plutôt que d'utiliser la fonction de densité de la *Weibull* ajustée. Dans ce cas, la formule pour l'estimation de la production d'énergie en kW/an devient :

$$Prod_2 = \sum_{j=1}^{25} puis_j \cdot \frac{n_j}{N_{10}} \cdot 365{,}25 \cdot 24,$$

où n_j représente le nombre d'observations de vitesse du vent aux dix minutes dans la classe j (j=1,...,25), la classe j=1 représentant des vitesses qui vont de 0 à 1 m/s et la classe j=25 représentant des vitesses allant de 24 à 25 m/s cette fois, puisque la puissance devient nulle après. Le dénominateur ici se doit d'être le nombre total de vitesses de vent aux dix minutes, afin de pouvoir comparer la fraction obtenue ici (n_j/N_{10}) à la probabilité obtenue à partir de la *Weibull* ajustée plus tôt sur toute la distribution des observations aux dix minutes. On pense que le fait d'utiliser la distribution expérimentale peut nous donner une idée de l'ampleur de la différence entre l'estimation de la *Weibull* et la réalité. On peut donc utiliser la distribution expérimentale dans le calcul de la production d'énergie estimée pour se donner une valeur de référence à laquelle comparer l'estimation faite à partir de la *Weibull*. Le tableau 4.4 présente, pour chaque site étudié, la production d'énergie estimée à partir de l'ajustement de densité de *Weibull* et la production d'énergie estimée à partir des fréquences de vent expérimentales. Les paramètres $\hat{k}$ et $\hat{\lambda}$ utilisés dans le calcul de la production d'énergie estimée à partir de la *Weibull* (4.5.1) sont ceux calculés à partir de la distribution globale de la vitesse du vent aux dix minutes ($\hat{k}_G$ et $\hat{\lambda}_G$).

4.5.1. Résultats

TABLEAU 4.4. Comparaison des estimations de la production d'énergie annuelle faites à partir de la densité *Weibull* ajustée ou des fréquences empiriques de la distribution des vitesses de vent *Weibull* annuelles

Site	N_{10}	Estimation de la production d'énergie (kW/an)		Erreur relative p/r à fréquences (%)
		À partir de la *Weibull*	À partir des fréquences	
1	444 777	7 505 120	7 397 724	1,45
2	281 117	8 419 031	8 866 938	-5,05
3	222 688	7 954 900	7 900 730	0,69
4	320 903	6 526 109	6 380 852	2,28
5	49 689	7 893 300	7 812 491	1,03
6	178 278	7 224 776	7 131 771	1,30
7	261 566	5 800 907	5 546 591	4,59
8	51 979	8 522 101	8 589 171	-0,78
9	275 917	9 012 572	8 925 977	0,97
10	243 024	9 279 775	9 236 803	0,47
11	126 242	7 967 984	8 015 773	-0,60
12	51 983	5 543 743	5 408 930	2,49
13	195 796	7 557 156	7 611 996	-0,72
14	264 175	10 157 508	10 452 137	-2,82
15	159 580	8 447 073	8 354 194	1,11
16	174 216	5 422 102	5 228 319	3,71
17	53 629	3 465 501	3 377 229	2,61
18	48 507	9 616 664	9 855 781	-2,43
19	63 246	5 358 472	5 120 893	4,64
20	61 271	5 903 203	5 903 203	6,28
21	116 714	10 354 168	10 286 547	0,66
22	51 626	8 079 849	7 870 135	2,66
23	118 120	7 793 500	8 170 885	-4,62
24	97 640	8 438 529	8 371 262	0,80
25	35 758	8 677 819	8 170 706	6,21
26	14 085	11 311 648	11 055 074	2,32
27	21 289	11 471 330	11 064 096	3,68
28	52 082	9 999 613	9 836 322	1,66
29	83 844	7 586 470	7 561 606	0,33
30	122 229	7 562 620	7 430 818	1,77
31	207 621	4 284 831	4 192 896	2,19

4.5.2. Discussion

On remarque en regardant le tableau 4.4 que l'erreur relative entre les deux estimations de production d'énergie oscille entre $-5{,}05\%$ et $6{,}28\%$. De plus, pour 7 des 31 sites, l'erreur relative est négative et l'estimation à partir de la *Weibull* sous-estime la production d'énergie calculée à partir des fréquences réelles. Pour 24 des 31 sites, l'estimation par la *Weibull* représente donc une sur-estimation de la production d'énergie annuelle trouvée à partir des fréquences réelles de la vitesse du vent mesurée au site. Si on tient plutôt compte de l'erreur relative absolue (la dernière colonne du tableau ci-dessus en absolu), on trouve une moyenne d'erreur relative de 2,35%, avec écart-type de 1,72. La différence entre les deux méthodes est donc plutôt considérable (on parle de dizaines et de centaines de milliers de kW/an de différence), et il pourrait être préférable d'estimer la production d'énergie annuelle toujours à partir de la distribution empirique, puisque celle-ci devrait mieux représenter la distribution du vent annuelle que la distribution de la *Weibull* ajustée sur les vitesses de vent, d'autant plus que l'utilisation de la *Weibull* mène à une sur-estimation de la production d'énergie dans la majorité des cas (menant du même coup à un déficit d'énergie par rapport aux attentes pour un site particulier).

4.6. Conclusion sur la modélisation de la vitesse du vent

De façon globale, les tests que nous avons faits dans ce chapitre nous ont pointé un problème par rapport à la modélisation de la vitesse du vent. En effet, la distribution globale de la vitesse du vent (celle de toutes les vitesses disponibles mesurées à un site) ne semble pas être bien modélisée par une seule distribution *Weibull* si l'on se base sur les tests du Khi-deux faits dans ce chapitre. Rappelons-nous que que les résultats obtenus aux tests du Khi-deux étaient, selon nous, sûrement dûs à la très grande puissance du test, lorsqu'autant de données appartiennent à la distribution modélisée. Or, si les observations sont dépendantes entre elles pour un certain délai de temps, ce qui semble être le cas selon les résultats en section 4.4.2, cette puissance est moindre et il devient difficile de déterminer si la *Weibull* est inappropriée pour modéliser la distribution de la vitesse du vent ou si ce n'est que question de puissance.

Nous avons aussi pensé que l'utilisation de paramètres de *Weibull* variant dans le temps pourrait permettre de pallier à la mauvaise modélisation de la distribution de vitesses de vent. La comparaison des variances des paramètres à

celles obtenues lors de simulations avec des paramètres globaux, annuels ou mensuels nous a permis de voir qu'avec une distribution de vitesses du vent pouvant être bien modélisée par une seule *Weibull*, nous obtiendrions des variances des paramètres bien différentes de celles observées. La comparaison nous a aussi permis de voir que même en utilisant des distributions différentes chaque année, nous obtiendrions des variances mensuelles échantillonnales encore beaucoup plus petites que celles que nous avons observées. Il semble que les paramètres ajustés chaque mois fassent un peu mieux. De plus, l'étude de la variance des paramètres nous a permis de trouver que les différences observées entre les variances asymptotiques et échantillonnales n'étaient pas nécessairement dues à la dépendance entre les observations. Par contre, nous n'avons pas tenu compte de structure de dépendance quelconque entre les données, dans le sens où nous avons comparé les résultats observés à des résultats simulés où les données utilisées étaient indépendantes et avons conclus en se basant là-dessus. Il serait donc intéressant, dans de futures études, de trouver une façon de juger de la qualité de la modélisation des vitesses de vent de façon séparée, en tenant aussi compte de la dépendance entre les observations, par exemple dans les simulations.

En ce qui a trait à l'estimation de la production d'énergie annuelle en kW/an à partir de la *Weibull* ou des fréquences de la vitesse du vent mesurée, nous avons trouvé des différences assez considérables entre les deux estimations pour la plupart des sites. La compagnie *Hatch* pratique présentement l'estimation de la production d'énergie à partir de l'ajustement d'une distribution *Weibull* et nous voulions voir à quel point leur estimation pourrait être influencée par l'utilisation des fréquences de la vitesse du vent mesurée. Bien entendu, nous ne savons jamais si la prochaine année à un site sera similaire aux années qui viennent de passer, en termes de vitesses de vent, et nous ne pouvons pas être certains que les vitesses de vent mesurées sont plus représentatives que celles obtenues par la modélisation de la distribution du vent à partir de la *Weibull*. Or, il fût intéressant de démontrer qu'il existe bel et bien une différence entre les deux méthodes et d'autres études pourraient se pencher sur les avantages d'utiliser la distribution des vitesses de vent mesurées, qui ne nécessite aucune modélisation ou encore sur la modélisation non paramétrique de la distribution des vitesses de vent à partir de différentes méthodes.

Chapitre 5

PRÉVISION DE LA VITESSE DU VENT PASSÉE ET ÉVALUATION DE LA VARIABILITÉ DE CETTE PRÉVISION

5.1. Utilité des prévisions de la vitesse du vent sur des années passées

Les ingénieurs éoliens tentent de prédire les données de vent à long terme afin d'avoir un indice de la moyenne de la vitesse du vent à long terme plutôt que sur quelques années seulement. Par exemple, s'ils ont installé un mât de mesure il y a trois ans et qu'ils collectent des données depuis ce temps, ils ont une mesure de la vitesse moyenne du vent pour les dernières années mais les banquiers veulent une évaluation de la vitesse moyenne du vent sur un plus long horizon, généralement dix ans, pour mieux tenir compte des variations annuelles. Ces ingénieurs cherchent donc des moyens permettant de prédire le vent, à plus long terme, à partir d'autres sources de données. La méthode *MCP* (Measure-Correlate-Predict), souvent utilisée dans le domaine de l'éolien, consiste en une façon de prédire le vent dans le passé, au-delà de la période de données mesurées sur des mâts, à partir de données disponibles sur une plus longue période, par exemple les données provenant d'*Environnement Canada* ou les données simulées.

5.2. La méthode MCP

5.2.1. Définition de la méthode

La méthode MCP est composée de trois étapes :
– La collecte des données ("*measure*"),

- la régression linéaire entre les données mesurées sur un court terme et les données simulées ou d'une source comme *Environnement Canada* sur le long terme dont on garde la partie correspondante au court terme ("*correlate*") et
- la prévision dans le passé des vitesses de vent collectées non disponibles à partir de la régression linéaire déjà faite ("*predict*").

C'est cette méthode qui permet aux ingénieurs de prédire la vitesse du vent dans le passé. Par exemple, si deux années de données collectées sont disponibles, mais que l'on désire obtenir une idée du vent moyen sur dix ans, on prédit à partir d'une régression linéaire simple et des données simulées (ou d'autres disponibles, comme celles d'*Environnement Canada*) sur les huit années passées manquantes les vitesses de vent collectées puis on calcule la moyenne de la vitesse du vent sur dix ans à partir de deux années réelles et de huit années prédites. La prochaine section présente les modèles statistiques utilisés pour la prévision de la vitesse du vent.

5.2.2. Les modèles linéaires utilisés pour faire les prévisions

Rappelons que nous utiliserons, tout au long de ce chapitre, les données collectées moyennes aux heures de l'anémomètre 1 qu'on note par y_i, $i = 0,1,...,N$ ou encore $\mathbf{Y}$. Nous considérerons trois modèles de prévision et commencerons d'abord par présenter le cas général. Soit $\mathbf{X}$ la matrice composée d'une première colonne pleine de 1 et d'autres colonnes qui représenteront les prédicteurs dans la régression linéaire qui nous permettra de prédire la vitesse du vent dans le passé. Ici, $\mathbf{X}$ aura donc deux ou trois colonnes, dépendamment de si l'on utilise seulement les vitesses de vent méso-échelle (ce que *Hatch* veut investiguer), seulement les vitesses de vent provenant d'une station de référence (ce qu'ils font présentement), ou les deux prédicteurs ensemble dans une régression linéaire multiple (ce qui, selon nous, pourrait être mieux).

On supposera d'abord que les données mesurées $\mathbf{Y}$ sont reliées aux prédicteurs par le modèle linéaire suivant :

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

où $\mathbf{Y}$ est de taille $(N+1)$, $\boldsymbol{\beta}$ est le vecteur de dimension p des paramètres inconnus du modèle, la matrice $\mathbf{X}$, de taille $(N+1)$ par p, a été définie plus tôt et $\boldsymbol{\varepsilon}$ est le vecteur des erreurs de moyenne 0. Notez que la dimension de chacune des colonnes de $\mathbf{X}$ est de $(N+1)$, puisque nous gardons la partie

commune des données mesurées, méso-échelle et provenant d'une station de référence, et que ces deux derniers jeux de données couvrent complètement la période où des données mesurées moyennes aux heures sont disponibles. On a donc toujours les temps $i = 0,...,N$ où les trois types de données sont disponibles.

Il est possible d'estimer les coefficients $\boldsymbol{\beta}$ à partir de la formule des moindres carrés, soit de la façon suivante :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{5.2.1}$$

C'est avec ces coefficients que nous pourrons prédire la vitesse du vent en dehors du domaine de régression, à partir de la formule suivante de prévision faite à partir d'une régression linéaire :

$$\hat{y}^p_{nouveau} = \mathbf{x}'_{nouveau}\hat{\boldsymbol{\beta}},$$

pour $\mathbf{x}_{nouveau}$ un vecteur composé des prédicteurs (la première entité valant 1), pour un temps ($nouveau$) qui n'est pas contenu dans $0,1,...,N$ et pour lequel on désire une prévision.

Maintenant, voyons les trois cas considérés. Nous utiliserons d'abord les vitesses de vent méso-échelle seules dans une régression linéaire simple où la variable dépendante est la vitesse de vent collectée à l'anémomètre 1. La matrice $\mathbf{X}$ sera donc composée d'une colonne pleine de 1 et d'une colonne composée de $N+1$ vitesses de vent méso-échelle. On obtiendra, à partir de la formule (5.2.1), deux coefficients qu'on notera $\hat{\beta}_{s,0}$ et $\hat{\beta}_{s,1}$. On définira comme suit la prévision faite à partir de ces coefficients et des vitesses de vent méso-échelle :

$$\hat{y}_{s,i} = \hat{\beta}_{s,0} + \hat{\beta}_{s,1}x_{s,i},\ i = S_{min},..., -1.$$

On pourra faire la même chose à partir des données provenant d'une station de référence, en modifiant la matrice $\mathbf{X}$ de sorte que la seconde colonne soit composée de $(N+1)$ vitesses de vent fournies par une station de référence située près du site. On estime encore, à partir de la formule (5.2.1), les coefficients expliquant la relation linéaire entre les données mesurées et celles de la station, puis on obtient deux coefficients qu'on notera maintenant $\hat{\beta}_{r,0}$ et $\hat{\beta}_{r,1}$. Il est donc possible de faire des prévisions de la vitesse du vent en dehors du domaine, de la façon suivante :

$$\hat{y}_{r,i} = \hat{\beta}_{r,0} + \hat{\beta}_{r,1}x_{r,i},\ i = R_{min},..., -1.$$

Finalement, on considère un troisième modèle où la matrice **X** est composée de trois colonnes : une colonne pleine de 1 et deux colonnes de prédicteurs, soient les vitesses de vent simulées méso-échelle et les vitesses récoltées aux stations de référence. La formule (5.2.1) nous fournira donc trois coefficients estimés, qu'on notera $\hat{\beta}_0$, $\hat{\beta}_s$ et $\hat{\beta}_r$, correspondant à la colonne de 1, aux vitesses méso-échelle et aux vitesses de la station de référence, respectivement. On pourra donc prédire la vitesse du vent mesurée en dehors du domaine de régression, à l'aide de la formule suivante :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_s x_{s,i} + \hat{\beta}_r x_{r,i},\ \text{i=max}\{S_{min},R_{min}\},...,\text{-1}.$$

5.3. Validation croisée sur les premières années de données disponibles à partir des deux dernières années

On comparera maintenant les trois modèles vus dans la section précédente en terme d'erreurs de prévision dans le passé à l'aide de la validation croisée. On verra alors l'utilité des données simulées méso-échelle pour les prévisions à long terme. Pour ce faire, définissons d'abord les deux types d'échantillons qui seront utilisés dans la validation croisée.

5.3.1. Les échantillons de validation et d'apprentissage

Les ingénieurs éoliens n'ont généralement en leur possession que deux ou trois années de données collectées aux mâts de mesure. C'est avec ces données qu'ils estiment le potentiel éolien à un site et l'on voudrait donc connaître la part de l'erreur de prévision de la vitesse du vent qui entre en compte dans l'estimation du potentiel éolien, lorsqu'un si petit nombre de données est disponible. L'on se mettra donc dans la position où seulement deux années de données sont disponibles, afin de faire les prévisions de la vitesse du vent passé. L'échantillon d'apprentissage sera composé des deux dernières années de données des trois types de séries de vent. C'est sur cet échantillon qu'on fera les différentes régressions linéaires vues dans la section 5.2.2. Pour faciliter la compréhension, l'on introduira maintenant la notation $i = 0, 1,..., N - N_{2ans}, N - N_{2ans} + 1,..., N$ pour l'indice de temps horaire, où N_{2ans} est le nombre d'observations disponibles durant les deux dernières années où des données ont été collectées. Ainsi, l'échantillon d'apprentissage sera composé des vitesses de vent correspondant aux indices $i = N - N_{2ans} + 1,..., N$ (ou encore $i \in E_{app}$), et ce tant pour les données simulées, des stations de référence et collectées. La taille de E_{app} sera notée N_{app} (égale à N_{2ans}). Notez que nous ne possédons pas de données provenant de stations de référence pour plusieurs des sites.

Ces données sont disponibles pour seulement 15 sites sur 31. Notez aussi que certains des sites où nous possédons des données provenant d'une station de référence n'avaient des données collectées que sur deux ou trois années. Pour ces sites, où l'on possède tout de même au moins deux ans de données, on gardera une plus petite partie de données pour l'échantillon d'apprentissage (environ le quart des données disponibles sur la période complète et les autres trois quarts pour faire la validation). Pour simplifier les notations, l'on notera par N_{2ans} la taille de l'échantillon d'apprentissage, dans tous les cas même ceux où l'on utilise moins de deux ans. Une validation croisée sera aussi faite pour les 16 autres sites où nous n'avons pas de données de référence, afin d'avoir une idée de l'erreur de prévision faite à partir de la régression où l'on utilise seulement les données méso-échelle.

L'échantillon de validation sera quant à lui composé du restant de chacune des séries y_i, $x_{s,i}$ et $x_{r,i}$ (donc les indices de temps $i = 0,...,N - N_{2ans}$, ou encore $i \in E_{valid}$). L'on notera par N_{valid} la taille de l'échantillon de validation. Comme son nom l'indique, cet échantillon servira à valider les différentes régressions linéaires faites sur les séries d'apprentissage, en calculant les prévisions de la vitesse du vent passée et en les comparant à ce qui est réellement collecté sur le site, à l'anémomètre 1.

5.3.2. Application de la validation croisée

Comme l'on désire estimer l'erreur de prévision des différentes régressions linéaires de la section 5.2.2, on estime les coefficients $\beta_{s,0}$, $\beta_{r,0}$, β_0, $\beta_{s,1}$, $\beta_{r,1}$, β_r et β_s des trois modèles de régression à partir des séries de données de l'échantillon d'apprentissage et des formules des moindres carrés montrées dans cette même section. On obtient donc les formules de prévision suivantes :

$$\hat{y}_{s,i} = \hat{\beta}_{s,0} + \hat{\beta}_{s,1} x_{s,i},$$
$$\hat{y}_{r,i} = \hat{\beta}_{r,0} + \hat{\beta}_{r,1} x_{r,i}$$

et

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_s x_{s,i} + \hat{\beta}_r x_{r,i},$$

qu'on utilise afin de prédire aux indices $i = 0,...,N - N_{2ans}$ (donc sur l'échantillon de validation) la vitesse du vent collectée. Or, pour ces indices, on possède aussi la vitesse de vent moyenne horaire de l'anémomètre 1, y_i. Il est donc possible d'estimer l'erreur quadratique moyenne de prévision pour chacune des trois méthodes de prévision :

$$\widehat{EQM}_s^{cv} = \frac{1}{N_{valid}} \sum_{i \in E_{valid}} (y_i - \hat{y}_{s,i})^2,$$

$$\widehat{EQM}_r^{cv} = \frac{1}{N_{valid}} \sum_{i \in E_{valid}} (y_i - \hat{y}_{r,i})^2 \text{ et}$$

$$\widehat{EQM}_{s,r}^{cv} = \frac{1}{N_{valid}} \sum_{i \in E_{valid}} (y_i - \hat{y}_i)^2.$$

On cherchera donc à connaître la meilleure méthode de régression, soit celle offrant la plus petite erreur quadratique moyenne de prévision. On comparera les modèles à l'aide de la racine carrée de l'erreur quadratique moyenne estimée.

5.3.3. Résultats

Les résultats des validations croisées entreprises pour tous les 31 sites sont présentés dans le tableau 5.1, à la page suivante. En gras, on retrouve les erreurs quadratiques moyennes minimales entre celles trouvées à partir des trois divers modèles de régression linéaire.

5.3.4. Discussion

Dans un premier temps, on peut voir à partir des résultats que la méthode présentement utilisée par *Hatch*, soit de prévoir la vitesse du vent à partir des données de stations de référence seulement, mène aux pires résultats sauf pour le site 19. À part ce site, le modèle avec les données méso-échelle seules pour prévoir les vitesses mesurées fait toujours mieux que le modèle présentement utilisé. Cependant, cela ne veut pas dire qu'on devrait simplement remplacer les données de station de référence par les données méso-échelle. En effet, en définissant l'amélioration relative comme suit pour le modèle à deux prédicteurs par rapport à celui où l'on utilise seulement les données méso-échelle :

$$100 \times \frac{\sqrt{\widehat{EQM}_s^{cv}} - \sqrt{\widehat{EQM}_{s,r}^{cv}}}{\sqrt{\widehat{EQM}_s^{cv}}},$$

et en définissant de la même façon l'amélioration relative du modèle à deux prédicteurs par rapport à celui où on utilise seulement les données de la station de référence (on change seulement s par r dans $\sqrt{\widehat{EQM}_s^{cv}}$, dans la formule ci-haut), on a pu trouver une amélioration relative de 8% (minimum de 0,1%, maximum de 16,2%) lorsqu'on ajoutait les données de station de référence aux

données méso-échelle déjà dans le modèle alors que l'utilisation des deux variables augmente de 26,7% la précision par rapport à ce qu'ils font présentement (minimum de 13,8%, maximum de 54,8%). Les deux types de données sont donc utiles pour la prévision de la vitesse du vent mesurée à partir d'une régression linéaire (multiple dans ce cas).

TABLEAU 5.1. Racines carrées des erreurs quadratiques moyennes de prévision par validation croisée pour chaque site et chaque groupe de prédicteurs dans la régression linéaire

Site	N_{valid}	N_{app}	Données simulées	Données de station de référence[1]	Deux types de données[1]
1	55 184	17 521	1,766	2,450	**1,725**
2	21 827	8 762	3,200	3,299	**2,844**
3	19 605	17 521	2,100	2,401	**1,914**
4	34 532	17 521	1,825	2,273	**1,661**
5	5 998	2 008	1,916	2,218	**1,683**
6	12 329	17 521	1,873	-	-
7	25 949	17 521	1,406	1,962	**1,334**
8	5 998	2 613	2,243	-	-
9	28 994	17 521	2,590	-	-
10	23 008	17 521	2,429	-	-
11	3 673	17 521	1,965	-	-
12	5 999	2 671	1,365	-	-
13	15 115	17 521	2,803	-	-
14	26 718	17 521	3,039	-	-
15	9 096	17 521	2,078	-	-
16	11 532	17 521	1,761	-	-
17	5 887	2 989	1,872	1,890	**1,578**
18	5 999	2 030	2,169	3,504	**2,167**
19	7 999	2 086	2,701	2,628	**2,263**
20	7 999	1 771	2,555	2,558	**2,191**
21	15 999	3 321	2,199	4,784	**2,164**
22	5 999	2 585	2,498	-	-
23	14 999	4 698	2,156	-	-
24	11 999	4 721	1,856	3,018	**1,854**
25	3 999	1 988	2,392	-	-
26	1 599	788	2,673	-	-
27	2 299	1 230	2,572	-	-
28	5 999	2 719	2,411	-	-
29	8 999	3 833	1,846	2,634	**1,754**
30	2 741	17 521	1,954	2,497	**1,824**
31	16 863	17 521	1,845	1,943	**1,617**

[1]Les cases vides représentent les sites où nous ne possédions pas de données de stations de référence.

5.3.5. Estimation de la moyenne de la vitesse du vent long terme et erreur relative

En plus d'estimer l'erreur quadratique moyenne de prévision à partir de la validation croisée, il nous est possible d'estimer la moyenne long terme de la vitesse du vent à partir des prévisions du vent dans l'échantillon de validation et des vitesses de vent mesurées de l'échantillon d'apprentissage. Rappelons la formule de la moyenne de la vitesse de vent long terme qui est utilisée dans la méthode MCP.

$$\hat{\bar{y}} = \frac{1}{N+1}\left\{\sum_{i\in E_{valid}} \hat{y}_i + \sum_{i\in E_{app}} y_i\right\}.$$

Nous comparons donc cette estimation (dans laquelle $\hat{y}_i$ est l'estimation de y_i trouvée à partir des trois différents modèles de prévision vus précédemment) à la véritable vitesse moyenne du vent à l'anémomètre 1 pour chaque site, soit

$$\bar{y} = \frac{1}{N+1}\sum_{i=0}^{N} y_i.$$

Le tableau 5.2 à la page suivante présente l'erreur relative par rapport à la véritable moyenne de la vitesse à l'anémomètre 1.

5.3.6. Discussion

Dans les sections précédentes, nous avons regardé l'estimation de l'erreur quadratique moyenne de prévision afin de déterminer laquelle des combinaisons de divers prédicteurs (données méso-échelle, de stations de référence ou les deux ensemble) menait à la plus petite EQMP. Cependant, *Hatch* utilise plutôt la moyenne long terme de la vitesse du vent comme indicateur de la valeur d'un site en terme d'énergie éolienne. Il était donc intéressant ici de comparer les différentes estimations de la moyenne long terme afin de déterminer laquelle des méthodes semble offrir la meilleure estimation. Nous avons trouvé, pour le modèle avec données méso-échelle seulement, une erreur relative absolue moyenne de $1{,}604\%$ (écart-type de $1{,}238$) si on utilise seulement les 15 sites comparables et de $2{,}089\%$ (écart-type de $2{,}178$) si on utilise tous les 31 sites. Pour l'estimation à partir des observations provenant de stations de référence, l'erreur relative absolue moyenne était de $1{,}203\%$ (écart-type de $1{,}337$). Quant au modèle à deux prédicteurs, nous avons trouvé une erreur relative absolue moyenne de $1{,}090\%$ (écart-type de $1{,}119$). Notez donc que le modèle avec les données de station de référence comme seul prédicteur fait généralement

TABLEAU 5.2. Erreurs relatives (%) entre la vitesse moyenne long terme estimée à partir de chacun des trois modèles de régression et la vitesse moyenne des vitesses de vent mesurées à l'anémomètre 1

Site	N_{valid}	N_{app}	Données simulées	Données de station de référence[1]	Deux types de données[1]
1	55 184	17 521	-0,444	1,223	**0,015**
2	21 827	8 762	**-0,364**	1,645	1,645
3	19 605	17 521	-0,495	**0,005**	-0,421
4	34 532	17 521	-0,715	-0,875	**-0,321**
5	5 998	2 008	2,516	4,219	**1,981**
6	12 329	17 521	-0,243	-	-
7	25 949	17 521	1,160	-0,878	**0,524**
8	5 998	2 613	1,557	-	-
9	28 994	17 521	1,748	-	-
10	23 008	17 521	2,808	-	-
11	3 673	17 521	0,641	-	-
12	5 999	2 671	-0,069	-	-
13	15 115	17 521	-3,552	-	-
14	26 718	17 521	-0,146	-	-
15	9 096	17 521	-0,007	-	-
16	11 532	17 521	-0,039	-	-
17	5 887	2 989	-1,326	3,703	**0,255**
18	5 999	2 030	-1,120	**-0,042**	-0,354
19	7 999	2 086	3,362	**-0,724**	0,830
20	7 999	1 771	3,740	**-0,579**	-1,062
21	15 999	3 321	0,796	**-0,208**	-1,231
22	5 999	2 585	-8,010	-	-
23	14 999	4 698	2,945	-	-
24	11 999	4 721	2,147	**0,851**	2,037
25	3 999	1 988	-4,824	-	-
26	1 599	788	-3,935	-	-
27	2 299	1 230	-1,363	-	-
28	5 999	2 719	8,812	-	-
29	8 999	3 833	-3,730	**2,791**	-4,351
30	2 741	17 521	-0,197	0,153	**-0,080**
31	16 863	17 521	1,943	**-0,145**	1,249

[1]Les cases vides représentent les sites où nous ne possédions pas de données de stations de référence.

mieux que celui avec les données méso-échelle ici, et que c'est le seul de nos tests qui présente de tels résultats. Par contre, les données méso-échelle ne sont pas à bannir puisqu'elles apportent encore un peu d'information dans le modèle à deux prédicteurs lors de la prévision de la vitesse du vent passée. En

effet, le modèle à deux prédicteurs est le meilleur en terme d'erreur relative absolue moyenne sur tous les sites.

5.4. Calcul de la variabilité des prévisions à partir du bootstrap

Les estimateurs de la validation croisée de la section précédence estiment (5.4.1), l'erreur quadratique moyenne de prévision :

$$\begin{aligned} EQMP =& E\left(\frac{1}{N_{valid}}\sum_{i\in E_{valid}}(y_i-\hat{y}_i)^2\right) \\ =& \frac{1}{N_{valid}}E\left((\mathbf{Y}_{valid}-\hat{\mathbf{Y}}_{valid})'(\mathbf{Y}_{valid}-\hat{\mathbf{Y}}_{valid})\right). \end{aligned} \tag{5.4.1}$$

Ils ne peuvent être calculés que si nous avons les valeurs de y_i dans l'ensemble de validation, ce qui n'est généralement pas le cas. Afin de voir comment nous pourrions estimer l'EQMP, faisons l'hypothèse, comme nous l'avons fait jusqu'à maintenant, que les observations sont indépendantes, de même variance. Afin d'alléger la notation, dénotons par $\mathbf{X}$ et $\mathbf{X}_v$ les matrices de variables explicatives pour les observations d'apprentissage et de validation, respectivement. Voyons comment il est possible de développer la formule de l'erreur quadratique moyenne de prévision si l'on postule

$$\begin{aligned} \mathbf{Y}_{app} =& \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \\ \mathbf{Y}_{valid} =& \mathbf{X}_v\boldsymbol{\beta}+\boldsymbol{\varepsilon}_v \end{aligned}$$

où $E(\boldsymbol{\varepsilon})=\mathbf{0}_{N_{app}}$, $E(\boldsymbol{\varepsilon}_v)=\mathbf{0}_{N_{valid}}$, $Var(\boldsymbol{\varepsilon})=\sigma^2\mathbb{I}_{N_{app}\times N_{app}}$, $Var(\boldsymbol{\varepsilon}_v)=\sigma^2\mathbb{I}_{N_{valid}\times N_{valid}}$ et $\boldsymbol{\varepsilon}$ et $\boldsymbol{\varepsilon}_v$ sont indépendants.

Lors de la validation croisée, nous n'avons que $\mathbf{Y}_{app}$ puisque $\mathbf{Y}_{valid}$ n'est généralement pas observé. On estime donc $\boldsymbol{\beta}$ par les moindres carrés sur le bloc d'apprentissage, soit

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_{app}. \tag{5.4.2}$$

On prédit les valeurs de $\mathbf{Y}_{valid}$ à l'aide de

$$\hat{\mathbf{Y}}_{valid}=\mathbf{X}_v\hat{\boldsymbol{\beta}}=\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_{app}.$$

Pour calculer l'EQMP dans (5.4.1), notons que :

$$\begin{aligned} \hat{\mathbf{Y}}_{valid}-\mathbf{Y}_{valid} =& \mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-(\mathbf{X}_v\boldsymbol{\beta}+\boldsymbol{\varepsilon}_v) \\ =& \mathbf{X}_v\boldsymbol{\beta}+\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}-\mathbf{X}_v\boldsymbol{\beta}-\boldsymbol{\varepsilon}_v \end{aligned}$$

$$=\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}_v.$$

De cette façon, on trouve que

$$\begin{aligned}(\hat{\mathbf{Y}}_{valid}-\mathbf{Y}_{valid})'(\hat{\mathbf{Y}}_{valid}-\mathbf{Y}_{valid}) =&\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}-\\ &2\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\boldsymbol{\varepsilon}_v+\boldsymbol{\varepsilon}_v'\boldsymbol{\varepsilon}_v.\end{aligned}$$

Maintenant, si $E(\boldsymbol{\varepsilon})=\boldsymbol{\mu}$ et $Var(\boldsymbol{\varepsilon})=\boldsymbol{V}$, alors $E(\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon})=tr(\boldsymbol{AV})+\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$ (Searle, 1971), de telle sorte qu'avec les hypothèses posées précédemment, on a

$$\frac{1}{N_{valid}}E\left((\hat{\mathbf{Y}}_{valid}-\mathbf{Y}_{valid})'(\hat{\mathbf{Y}}_{valid}-\mathbf{Y}_{valid})\right) \tag{5.4.3}$$

$$=\frac{1}{N_{valid}}\left(\sigma^2 tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')+0+\sigma^2 tr(\mathbb{I}_{N_{valid}})\right)$$

$$=\sigma^2\left(1+\frac{tr((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v)}{N_{valid}}\right). \tag{5.4.4}$$

Bien entendu, cela demeure vrai en autant que le modèle tienne tant pour le vecteur $\mathbf{Y}_{app}$ que pour le vecteur $\mathbf{Y}_{valid}$.

Il est donc possible d'estimer l'EQMP avec cette dernière formule en remplaçant σ^2 par son estimateur basé sur la somme des carrés des résidus de la régression calculée sur les données appartenant à E_{app}.

Maintenant, un des problèmes qui se pose dans la démarche précédente est l'hypothèse d'indépendance entre les erreurs. Nous avons vu dans le chapitre précédent (section 4.4.2) que l'hypothèse de l'indépendance des vitesses de vent mesurées dans le temps n'est pas compatible avec les données et il existe donc une certaine structure de corrélation pour les erreurs du modèle de régression linéaire entre les données mesurées et divers prédicteurs. Postulons donc maintenant ceci :

$$\begin{pmatrix}\mathbf{Y}_{app}\\ \mathbf{Y}_{valid}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \mathbf{X}_v\end{pmatrix}\boldsymbol{\beta}+\begin{pmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_v\end{pmatrix}$$

où on fait maintenant les hypothèses

$$E\begin{pmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_v\end{pmatrix}=\begin{pmatrix}\mathbf{0}_{N_{app}}\\ \mathbf{0}_{N_{valid}}\end{pmatrix}\text{ et }Var\begin{pmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_v\end{pmatrix}=\begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}. \tag{5.4.5}$$

Voyons à quoi ressemble la formule de l'erreur quadratique moyenne de prévision développée selon ces hypothèses.

On utilise les mêmes estimateurs des coefficients de la régression linéaire, soient

ceux présentés en (5.4.2). Les étapes menant au développement de l'erreur quadratique moyenne de prévision sont les mêmes, sauf au moment où la matrice de variance des erreurs change. Reprenons donc à partir de (5.4.3) :

$$\frac{1}{N_{valid}} E\left((\hat{\mathbf{Y}}_{valid} - \mathbf{Y}_{valid})'(\hat{\mathbf{Y}}_{valid} - \mathbf{Y}_{valid})\right)$$

$$= \frac{1}{N_{valid}} E\left[\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\boldsymbol{\varepsilon}_v + \boldsymbol{\varepsilon}_v'\boldsymbol{\varepsilon}_v\right].$$

Grâce à Searle (1971), on sait que :

$$E\left(\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right)$$
$$= tr\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V_{11}}\right)$$
$$\text{et} \quad E\left(\boldsymbol{\varepsilon}_v'\boldsymbol{\varepsilon}_v\right) = tr\left(\mathbf{V_{22}}\right).$$

De plus, on peut calculer la valeur de $-2E\left(\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\boldsymbol{\varepsilon}_v\right)$ en créant de nouvelles matrices, soient

$$\boldsymbol{\varepsilon}_{nouveau} = \begin{pmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}_v \end{pmatrix} \quad \text{et} \quad \mathbf{B} = \begin{pmatrix} \mathbf{0}_{N_{app}\times N_{app}} & \mathbf{A} \\ \mathbf{0}_{N_{valid}\times N_{app}} & \mathbf{0}_{N_{valid}\times N_{valid}} \end{pmatrix}$$

où

$$\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'.$$

Alors rechercher $-2E\left(\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\boldsymbol{\varepsilon}_v\right)$ revient à calculer ceci :

$$-2E\left(\boldsymbol{\varepsilon}_{nouveau}'\mathbf{B}\,\boldsymbol{\varepsilon}_{nouveau}\right) = -2tr\left(\mathbf{B}\begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}\right)$$
$$= -2tr\left(\begin{pmatrix} \mathbf{0}_{N_{app}\times N_{app}} & \mathbf{A} \\ \mathbf{0}_{N_{valid}\times N_{app}} & \mathbf{0}_{N_{valid}\times N_{valid}} \end{pmatrix}\begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}\right)$$
$$= -2tr\left(\mathbf{A}\mathbf{V_{21}}\right) = -2tr\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{V_{21}}\right).$$

Finalement, on peut regrouper tout ensemble pour montrer que l'erreur quadratique moyenne de prévision sous les hypothèses posées précédemment devient :

$$\frac{1}{N_{valid}}\left\{tr\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V_{11}}\right]\right.$$
$$\left. -2tr\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{V_{21}}\right] + tr\left(\mathbf{V_{22}}\right)\right\}. \tag{5.4.6}$$

On remarque assez rapidement que, comparativement à l'erreur quadratique moyenne de prévision développée sous l'hypothèse d'indépendance des erreurs, on obtient ici plusieurs termes qui ne s'annulent plus entre eux. On aura entre autres les coefficients à l'intérieur des partitions de la matrice de variance-covariance à estimer (ceux des matrices $\mathbf{V}_{11}$, $\mathbf{V}_{12}$, $\mathbf{V}_{21}$ et $\mathbf{V}_{22}$), contrairement à l'autre développement où nous avions simplement σ^2 à estimer. Ainsi, l'on pourrait maintenant utiliser des modèles de séries chronologiques afin d'estimer ces coefficients en ajustant un modèle à nos données. Or, l'on ne connaît pas bien la tendance de la corrélation entre les données et l'on ne sait pas si des modèles classiques de séries chronologiques s'ajustent bien aux données. Une option s'offrant à nous et qui règle alors le problème est l'estimation de l'erreur quadratique moyenne de prévision à partir du bootstrap par bloc. En utilisant des blocs de données consécutives, nous pourrons capturer en bonne partie la structure de dépendance entre les résidus des diverses régressions sans avoir à estimer les paramètres d'un modèle postulé pour la matrice de variance-covariance. Commençons d'abord par revoir en quoi consiste le bootstrap.

5.4.1. Le bootstrap utilisé sur des données indépendantes et identiquement distribuées

Le bootstrap est une méthode d'inférence statistique introduite par Efron en 1979, basée sur le rééchantillonnage des données observées (boostrap non paramétrique) ou la simulation de données provenant d'une distribution paramétrisée particulière (bootstrap paramétrique). Cette méthode a été introduite par Efron en 1979 . Elle a pour objectif l'estimation de certaines caractéristiques de la distribution d'une statistique, par exemple, sa dispersion.

Considérons le cas où l'on possède un échantillon composé de n données $X_1,...,X_n$, qu'on considère comme i.i.d. de fonction de distribution F. Supposons que l'on désire estimer la dispersion d'un certain paramètre θ_F, lequel est estimé par la statistique $\hat{\theta}_F = T_n(X_1,...,X_n)$, qui dépend de l'échantillon observé. Afin d'estimer la dispersion, on doit donc estimer la distribution de notre estimateur $\hat{\theta}_F$. Si l'on doit estimer la distribution de notre statistique qui dépend de l'échantillon, on a besoin d'observer cette statistique plusieurs fois et c'est à ce moment que le bootstrap entre en jeu. On aura le choix entre tirer de façon aléatoire et avec remise B échantillons de n données parmi notre échantillon initial $X_1,...,X_n$, ou, si la distribution F est paramétrisée, $F(\lambda)$, d'estimer les paramètres de cette distribution, λ_n, et de tirer un échantillon de la

distribution $\hat{F}_n = F(\hat{\lambda}_n)$ avec les valeurs estimées des paramètres afin de simuler pour chaque bootstrap un échantillon de n observations, dans les deux cas, cela nous permettant de calculer sur chacun de ces échantillons (créés de façon paramétrique ou non) la statistique $T^*_{b,n}(X^*_{b,1},...,X^*_{b,n})$, où $b = 1,...,B$. La distribution de ces B statistiques bootstrap représentera donc une estimation de la distribution de la statistique T_n, de sorte qu'en calculant sur cette distribution estimée la variance échantillonnale, on obtiendra une estimation de la variance de l'estimateur T_n. On peut faire de même pour d'autres caractéristiques de la distribution qu'on voudrait estimer, par exemple l'erreur quadratique moyenne de prévision.

5.4.2. Le bootstrap pour évaluer l'erreur quadratique moyenne de prévision à partir d'une régression linéaire sur des données i.i.d.

Dans le cas de la régression linéaire où l'hypothèse d'indépendance entre les vitesses de vent mesurées tiendrait, il est possible de décrire un algorithme où le bootstrap nous permettrait d'estimer l'erreur quadratique moyenne de prévision, toujours à partir de la régression linéaire multiple où $\mathbf{X} = \mathbf{X}_{app}$ est la matrice composée des prédicteurs et $\mathbf{Y}_{app}$ le vecteur composé de la variable dépendante dans la régression. On définit aussi la matrice

$$\mathbf{X}_N = \begin{pmatrix} \mathbf{X} \\ \mathbf{X}_{valid} \end{pmatrix}$$

et on rappelle que $\mathbf{Y}$ représente le vecteur des vitesses de vent mesurées sur toute la période disponible, soit pour les temps $0,1,...,N$.

La quantité qu'on veut estimer ici est l'erreur quadratique moyenne de prévision, c'est-à-dire

$$EQMP = E\left(\frac{1}{N_{valid}} \sum_{i \in E_{valid}} (y_i - \hat{y}_i)^2\right). \tag{5.4.7}$$

On doit donc obtenir plusieurs jeux bootstrap de prévision du vent dans le passé (le passé correspondant ici aux temps $i \in E_{valid}$) afin de calculer plusieurs fois le terme dans l'espérance et d'estimer son espérance par sa moyenne. L'algorithme va comme suit :

Algorithme bootstrap pour l'estimation de l'EQMP dans le cas d'observations i.i.d.

(1) Régresser $\mathbf{Y}_{app}$ sur $\mathbf{X}$ afin d'obtenir les coefficients *initiaux* $\hat{\boldsymbol{\beta}}$ à partir de la formule (5.4.2).

(2) Calculer les résidus de la régression faite en (1).

(3) Tirer de facon aléatoire avec remise $(N+1)$ résidus parmi ceux calculés à l'étape précédente, de sorte à les mettre les uns après les autres et à ainsi créer une série d'erreurs bootstrap $\boldsymbol{\varepsilon}^*$ de longueur $(N+1)$.

(4) Créer les observations bootstrap comme suit :

$$\mathbf{Y}^* = \mathbf{X}_N \hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}^*,$$

où l'on peut décomposer $\mathbf{Y}^*$ en deux vecteurs $\mathbf{Y}^*_{valid}$ et $\mathbf{Y}^*_{app}$.

(5) Régresser $\mathbf{Y}^*_{app}$ sur $\mathbf{X}$ et obtenir les coefficients bootstrap $\hat{\boldsymbol{\beta}}^*$.

(6) Calculer les prévisions bootstrap dans le passé comme suit :

$$\hat{\mathbf{Y}}^*_{valid} = \mathbf{X}_{valid} \hat{\boldsymbol{\beta}}^* .$$

(7) Calculer la moyenne des carrés des résidus de la régression bootstrap comme suit :

$$\widehat{EQMP}^* = \frac{1}{N_{valid}} \left(\mathbf{Y}^*_{valid} - \hat{\mathbf{Y}}^*_{valid} \right)' \left(\mathbf{Y}^*_{valid} - \hat{\mathbf{Y}}^*_{valid} \right).$$

(8) Reproduire les étapes (3) à (7) un nombre B de fois pour obtenir B moyennes des carrés des résidus de régression $\widehat{EQMP}^*$.

(9) Calculer $\widehat{EQMP}^{boot}$ la moyenne des B estimations qui devient l'estimateur bootstrap de l'erreur quadratique moyenne de prévision.

5.4.2.1. *Le bootstrap par bloc*

Nous venons de voir comment il est possible d'estimer l'erreur quadratique moyenne des prévisions faites à partir d'une régression linéaire, dans le cas où la variable dépendante dans la régression est considérée comme i.i.d. Or, tel que mentionné plus tôt, il existe une structure de corrélation assez complexe à modéliser entre les vitesses de vent mesurées. Il est possible de tenir compte de la dépendance entre les erreurs en modifiant légèrement l'algorithme bootstrap présenté dans la section précédente. En effet, regrouper les erreurs bootstrap provenant de la régression dans des blocs permettrait de tenir compte de cette dépendance. Voyons maintenant notre nouvelle façon de procéder :

Algorithme bootstrap pour l'estimation de l'EQMP dans le cas d'observations dépendantes entre elles

(1) Régresser $\mathbf{Y}_{app}$ sur $\mathbf{X}$ afin d'obtenir les coefficients *initiaux* $\hat{\boldsymbol{\beta}}$ à partir de la formule (5.4.2).

(2) Calculer les résidus de la régression faite en (1).

(3) Créer $N_{app} - l + 1$ blocs de résidus de longueur l à partir des résidus calculés à l'étape précédente. Chaque bloc h est composé des l résidus consécutifs $(\hat{\varepsilon}_h,...,\hat{\varepsilon}_{h+l-1})$ et les blocs se chevauchent de sorte que chaque bloc est décallé d'une unité de temps par rapport au bloc précédent.

(4) Tirer avec remise parmi ces blocs de longueur l un nombre s de blocs de résidus qu'on colle ensemble, les uns après les autres, afin de créer une série de résidus de longueur $N+1$ (donc $s = \lceil(\frac{N+1}{l})\rceil$). Si la longueur de la série composée des blocs mis les uns après les autres dépasse $N+1$, on coupe à $N+1$ résidus. La série composée des blocs collés les uns après les autres est utilisée comme série d'erreurs bootstrap $\boldsymbol{\varepsilon}^*$.

(5) Créer les observations bootstrap comme suit :

$$\mathbf{Y}^* = \mathbf{X}_N\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}^*, \tag{5.4.8}$$

où l'on peut décomposer $\mathbf{Y}^*$ en deux vecteurs $\mathbf{Y}^*_{valid}$ et $\mathbf{Y}^*_{app}$.

(6) Régresser $\mathbf{Y}^*_{app}$ sur $\mathbf{X}$ et obtenir les coefficients bootstrap $\hat{\boldsymbol{\beta}}^*$.

(7) Calculer les prévisions bootstrap dans le passé comme suit :

$$\hat{\mathbf{Y}}^*_{valid} = \mathbf{X}_{valid}\hat{\boldsymbol{\beta}}^*. \tag{5.4.9}$$

(8) Calculer la moyenne des carrés des résidus de la régression bootstrap comme suit :

$$\widehat{EQMP}^{*,l} = \frac{1}{N_{valid}}\left(\mathbf{Y}^*_{valid} - \hat{\mathbf{Y}}^*_{valid}\right)'\left(\mathbf{Y}^*_{valid} - \hat{\mathbf{Y}}^*_{valid}\right).$$

(9) Reproduire les étapes (3) à (8) un nombre B de fois pour obtenir B moyennes des carrés des résidus de régression $\widehat{EQMP}^{*,l}$.

(10) Calculer la moyenne $\widehat{EQMP}^{boot,l}$ des B estimations $\widehat{EQMP}^{*,l}$ qui devient l'estimateur bootstrap de l'erreur quadratique moyenne de prévision.

On notera l'estimateur de l'EQMP (5.4.7) par $\widehat{EQMP}^{boot,l}$ dans le cas où la matrice $\mathbf{X}_N$ est composée d'une colonne de 1, d'une colonne de vitesses de vent méso-échelle et d'une troisième colonne de vitesses de vent provenant de station de référence. On notera par $\widehat{EQMP}^{boot,l}_s$ l'estimateur de l'EQMP dans le cas où la matrice $\mathbf{X}_N$ est composée d'une colonne de 1 et d'une colonne de vitesses de vent méso-échelle seulement. Finalement, on notera par $\widehat{EQMP}^{boot,l}_r$ l'estimateur de l'EQMP dans le cas où la matrice $\mathbf{X}_N$ est composée d'une colonne de 1 et d'une colonne de vitesses de vent provenant de station de référence seulement. On testera différentes tailles de blocs l afin de vérifier l'effet du changement de la taille sur les résultats. On pourra comparer les estimations d'erreurs

quadratiques moyennes obtenues à celles obtenues par validation croisée dans la section précédente, afin de voir quelle taille mènerait aux résultats les plus similaires. On rapportera donc dans le tableau 5.3 la différence absolue relative entre la racine de $\widehat{EQMP}^{boot,l}$ (ou $\widehat{EQMP}_s^{boot,l}$, ou $\widehat{EQMP}_r^{boot,l}$) et la racine carrée de l'EQM trouvée à partir de la validation croisée, pour les trois modèles respectifs et pour chaque taille de bloc considérée. La différence absolue relative (DAR) sera calculée de la façon suivante pour les trois modèles :

$$DAR_{s,r}^{l} = 100 \times \frac{\sqrt{\widehat{EQMP}^{boot,l}} - \sqrt{\widehat{EQM}_{s,r}^{cv}}}{\sqrt{\widehat{EQM}_{s,r}^{cv}}},$$

$$DAR_{s}^{l} = 100 \times \frac{\sqrt{\widehat{EQMP}_s^{boot,l}} - \sqrt{\widehat{EQM}_s^{cv}}}{\sqrt{\widehat{EQM}_s^{cv}}},$$

$$DAR_{r}^{l} = 100 \times \frac{\sqrt{\widehat{EQMP}_r^{boot,l}} - \sqrt{\widehat{EQM}_r^{cv}}}{\sqrt{\widehat{EQM}_r^{cv}}}.$$

Voyons maintenant les tailles de blocs que nous avons considérées ici.

5.4.2.2. *Tailles des blocs d'erreurs bootstrap*

La démarche décrite dans la section précédente sera reproduite pour trois différentes tailles l de bloc de résidus. On s'attend à mieux capturer la structure de dépendance lorsqu'on utilise des blocs plus longs. Notez que l'utilisation de blocs de longueur l implique qu'on fait l'hypothèse que les éléments de la matrice de variance-covariance dans (5.4.5) qui sont à une distance supérieure à $l-1$ de la diagonale sont égaux à 0. Cependant, plus les blocs sont longs, moins on possède de blocs pour estimer la distribution conjointe de dimension l. Il s'agit donc d'un compromis que de choisir une longueur de bloc. Dans notre cas, on testera d'abord la méthode avec des blocs de longueur 1 (un). Notez que l'utilisation de blocs de longueur 1 dans le bootstrap par bloc devrait mener, pour un nombre de bootstraps assez grand, à des résultats très similaires à ceux obtenus lors du calcul de l'erreur quadratique moyenne à partir des formules des moindres carrés, pour lesquelles l'hypothèse d'indépendance entre les données doit être faite ; voir (5.4.4). En effet, une structure de dépendance contenue dans des blocs de longueur 1 est équivalente à considérer l'indépendance entre les données. Ensuite, on testera le bootstrap pour des blocs de taille $l = N_{app}^{\frac{1}{3}}$ (Davison & Hinkley, 1997), donc $l \approx 26$ pour les sites où N_{app} est équivalent au nombre d'heures contenues dans deux ans. Finalement, le calcul sera

fait pour des blocs de taille 200 afin de voir l'impact sur les résultats lorsque des blocs plus longs sont utilisés. Il est intéressant de remarquer que des blocs de taille $26 \approx 24$ couvriront environ une structure de dépendance entre les données qui s'étend sur 24 heures, donc une journée, alors qu'une structure de dépendance s'étendant sur 200 données correspond environ à considérer qu'il y a de la dépendance jusqu'à un délai d'une semaine entre les vitesses de vent mesurées. Aussi, nous avons vu dans la section 4.4.2 que la corrélation entre les observations était bien présente sur un délai de plus de quatre jours, et les blocs de taille 200 couvriront au moins ces quatres jours où la corrélation est plus forte.

5.4.2.3. *Résultats*

TABLEAU 5.3. Différence absolue relative (DAR) entre la racine carrée de l'erreur quadratique moyenne de prévision sur 1 000 bootstraps et la racine carrée de l'erreur quadratique trouvée par validation croisée, pour chaque taille de bloc et chaque groupe de prédicteurs dans la régression (%)

Site	N_{valid}	N_{app}	Taille blocs (l)	DAR_s^l	DAR_r^l [1]	$DAR_{s,r}^l$ [1]
1	55 184	17 521	1	-2,95	-4,29	-2,33
			26	-2,92	-4,26	-2,30
			200	-3,19	-4,58	-2,60
2	21 827	8 762	1	-5,39	-1,41	-3,76
			21	-5,48	-1,26	-3,66
			200	-5,30	-1,68	-3,74
3	1 9605	17 521	1	3,72	0,83	1,43
			26	3,73	0,79	1,46
			200	3,63	0,68	1,33
4	34 532	17 521	1	-1,33	-3,24	-0,56
			26	-1,30	-3,20	-0,52
			200	-1,34	-3,39	-0,65
5	5 998	2 008	1	-5,66	-12,42	-11,37
			13	-5,54	-12,44	-11,27
			200	-7,64	-14,11	-13,26
6	12 329	17 521	1	5,36	-	-
			26	5,43	-	-
			200	5,37	-	-
7	25 949	17 521	1	1,08	-2,80	-0,86
			26	1,09	-2,84	-0,90
			200	0,76	-3,09	-1,17

[1] Les cases vides représentent les sites où nous ne possédions pas de données de stations de référence.

(Suite)

Site	N_{valid}	N_{app}	Taille blocs (l)	DAR_s^l	DAR_r^l [1]	$DAR_{s,r}^l$ [1]
8	5 998	2 613	1	-8,67	-	-
			14	-8,91	-	-
			200	-10,54	-	-
9	28 994	17 521	1	-4,32	-	-
			26	-4,28	-	-
			200	-4,55	-	-
10	23 008	17 521	1	-2,80	-	-
			26	-2,82	-	-
			200	-2,99	-	-
11	3 673	17 521	1	-9,64	-	-
			26	-9,57	-	-
			200	-10,00	-	-
12	5 999	2 671	1	6,12	-	-
			14	6,23	-	-
			200	4,80	-	-
13	15 115	17 521	1	-2,82	-	-
			26	-2,86	-	-
			200	-2,90	-	-
14	26 718	17 521	1	-4,37	-	-
			26	-4,34	-	-
			200	-4,56	-	-
15	9 096	17 521	1	1,29	-	-
			26	1,25	-	-
			200	1,01	-	-
16	11 532	17 521	1	-2,65	-	-
			26	-2,63	-	-
			200	-2,75	-	-
17	5 887	2 989	1	-1,92	24,96	10,46
			15	-1,80	25,07	10,57
			200	-4,23	23,58	8,79
18	5 999	2 030	1	-3,05	-5,11	-3,44
			13	-2,82	-4,93	-3,25
			200	-4,87	-6,71	-5,45
19	7 999	2 086	1	-7,40	-12,71	-7,73
			13	-7,99	-12,89	-7,87
			200	-9,73	-14,72	-9,53
20	7 999	1 771	1	-3,41	-11,48	-7,31
			13	-3,67	-11,67	-7,20
			200	-7,10	-15,66	-10,28
21	15 999	3 321	1	-11,76	-30,82	-10,46
			15	-11,58	-30,72	-10,20
			200	-13,33	-31,02	-11,94

[1] Les cases vides représentent les sites où nous ne possédions pas de données de stations de référence.

(Suite)

Site	N_{valid}	N_{app}	Taille blocs (l)	DAR_s^l	DAR_r^l [1]	$DAR_{s,r}^l$ [1]
22	5 999	2 585	1	26,37	-	-
			14	26,25	-	-
			200	24,05	-	-
23	14 999	4 698	1	12,02	-	-
			17	12,11	-	-
			200	10,97	-	-
24	11 999	4 721	1	-10,22	-7,41	-10,15
			17	-10,13	-7,45	-10,12
			200	-11,35	-8,97	-11,11
25	3 999	1 988	1	-13,33	-	-
			13	-13,32	-	-
			200	-17,28	-	-
26	1 599	788	1	-13,41	-	-
			10	-13,19	-	-
			200	-16,83	-	-
27	2 299	1 230	1	-4,34	-	-
			11	-3,92	-	-
			200	-6,58	-	-
28	5 999	2 719	1	9,23	-	-
			14	9,54	-	-
			200	7,99	-	-
29	8 999	3 833	1	9,91	9,98	15,24
			16	10,02	10,11	15,31
			200	9,41	9,32	14,90
30	2 741	17 521	1	2,41	8,03	2,00
			26	2,35	8,12	1,71
			200	2,03	7,72	1,85
31	16 863	17 521	1	-4,30	9,64	0,33
			26	-4,38	9,58	0,31
			200	-4,61	9,59	0,11

[1] Les cases vides représentent les sites où nous ne possédions pas de données de stations de référence.

5.4.2.4. *Discussion*

On remarque, pour une majorité des sites, que la longueur des blocs ne change pas beaucoup la différence absolue relative des racines carrées des erreurs quadratiques moyennes par rapport à celles obtenues par validation croisée. On voit aussi que le modèle menant aux EQMP calculées à partir du bootstrap se rapprochant le plus de l'EQM calculée avec la validation croisée est plutôt difficile à déterminer, puisque pour 7 des 15 sites où l'on peut comparer

les trois modèles, il s'agit du modèle à deux prédicteurs (moyenne des $DAR_{s,r}$ absolues de 5,78%, $l = N_{app}^{1/3}$), mais que ce modèle est suivi de très près par celui où l'on utilise seulement les données méso-échelle (5 sites sur 15 où les résultats sont les plus similaires à la validation croisée, avec une moyenne des DAR_s absolues de 6,50%, $l = N_{app}^{1/3}$ pour les 31 sites ou une moyenne de 4,99% sur les 15 sites comparables seulement). Notez que la moyenne des DAR_r absolues était plutôt de 9,69% pour $l = N_{app}^{1/3}$.

Suite à cette analyse, nous aurions aimé trouver une différence dans les résultats lorsque nous faisions varier la taille des blocs dans le bootstrap. On croit que la différence a été camouflée par un des termes dans l'erreur quadratique moyenne de prévision. Effectivement, en prenant en exemple la formule de l'EQMP sous l'hypothèse d'indépendance entre les données, on pouvait voir à partir de (5.4.4) que le terme de variance σ^2 multiplié par 1 domine le terme $\sigma^2 \left(\frac{tr((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v'\mathbf{X}_v)}{N_{app}} \right)$, ce dernier étant d'ordre $\frac{\sigma^2}{N_{app}}$, avec un N_{app} très grand (rappelons que les erreurs quadratiques moyennes de prévision sont calculées ici à partir des N_{app} observations de l'échantillon d'apprentissage). On ne peut pas aussi clairement différencier les termes dans la formule (5.4.6), formule où l'on tenait compte de la covariance entre les données, mais on pense que le même phénomène se produit et que cela pourrait peut-être expliquer pourquoi on ne détecte aucune différence au niveau de la taille des blocs dans le bootstrap. Effectivement, la variance d'une prévision faite à partir d'une régression linéaire peut être décomposée en la variance due à la prochaine observation autour de la droite de régression et la variance due à l'estimation de la droite de régression (en considérant une régression linéaire simple). Comme nous possédons ici un très grand nombre d'observations pour faire la régression linéaire (peu importe le modèle considéré), la variance due à l'estimation de la droite de régression risque d'être plutôt faible par rapport à celle due à la prochaine observation autour de la droite (plus d'observations mènent à une estimation moins variable de la droite de régression). Or, on s'attend à ce que la taille des blocs influence davantage la variance de l'estimation de la droite que celle de la prochaine observation. Par exemple, nous savons que des observations dépendantes entre elles mèneront à une plus grande variance de la droite de régression que des données indépendantes (la dépendance entre les données nous donnant accès à moins d'information différente sur ces dernières).

Nous sommes donc maintenant intéressés à voir l'effet de la taille de bloc sur la variance de la droite de régression, ou encore de façon semblable sur la variance de la prévision moyenne (cette prévision se trouvant sur la droite de

régression et qui demeure une quantité d'intérêt pour les ingénieurs de chez *Hatch*, pour qui une estimation de la variance pourrait être utile).

5.4.2.5. *Évaluation de la variance de la moyenne long terme estimée*

Comme la méthode MCP est utilisée afin d'estimer la moyenne long terme de la vitesse du vent, on voudrait maintenant vérifier si la taille des blocs dans le bootstrap aura un effet important sur la variabilité (estimée à l'aide du bootstrap) de la moyenne long terme estimée. On se concentre donc sur la variance due à la droite (ou au plan) de régression et on risque maintenant de détecter une différence selon la taille des blocs utilisés dans notre bootstrap.

Redéfinissons d'abord la moyenne long-terme calculée avec les vitesses de vent mesurées moyennes aux heures de l'anémomètre 1 :

$$\bar{y} = \frac{1}{N+1} \sum_{i=0}^{N} y_i.$$

Si l'on ne possèdait pas les premières N_{valid} vitesses de vent collectées, on pourrait utiliser l'un des modèles linéaires afin d'obtenir des prévisions de la vitesse du vent pour $i \in E_{valid}$, qu'on note $\hat{y}_i$ et, du même coup, l'estimation de la vitesse moyenne long terme par la méthode MCP deviendrait :

$$\hat{\bar{y}} = \frac{1}{N+1} \left\{ \sum_{i \in E_{valid}} \hat{y}_i + \sum_{i \in E_{app}} y_i \right\}. \tag{5.4.10}$$

Maintenant, si l'on désire évaluer la variance de la moyenne long terme estimée $\hat{\bar{y}}$, on peut encore une fois utiliser le bootstrap. Pour chaque bootstrap b, il est possible d'estimer cette moyenne long terme en utilisant le même modèle de régression linéaire utilisé dans (5.4.10) et en prédisant la vitesse du vent pour $i \in E_{valid}$, puis en réutilisant les prévisions bootstrap trouvées à partir du modèle ($\widehat{\mathbf{Y}}^*$), de la façon suivante :

$$\hat{\bar{y}}_b^* = \frac{1}{N+1} \left\{ \sum_{i \in E_{valid}} \hat{y}_{b,i}^* + \sum_{i \in E_{app}} y_{b,i}^* \right\}.$$

pour b=1,...,1 000, où $\hat{y}_{b,i}^*$ est défini en (5.4.9) et $y_{b,i}^*$ est défini en (5.4.8).

Après avoir calculé pour les 1 000 bootstraps cette moyenne long terme estimée, on peut calculer la variance échantillonnale comme suit :

$$Var^*(\hat{\bar{y}}_b^*) = \frac{1}{999} \sum_{b=1}^{1000} \left(\hat{\bar{y}}_b^* - \hat{\bar{y}}^*\right)^2,$$

où

$$\hat{\bar{y}}^* = \frac{1}{1000} \sum_{b=1}^{1000} \hat{\bar{y}}_b^*$$

On fait ces calculs pour les prévisions estimées à partir des trois différents modèles. On obtient ainsi trois mesures de variance de la moyenne long terme estimée et ce pour une taille de bloc l donnée.

Afin de déterminer si la taille de bloc a un effet sur la variance de la moyenne long terme estimée, on présentera dans le tableau suivant le rapport des variances trouvées à partir du bootstrap, pour deux différentes tailles de blocs. Encore une fois, les tailles de blocs qui ont été considérées étaient 1, $N_{app}^{\frac{1}{3}}$ et 200.

TABLEAU 5.4. Rapports de la variance de l'estimateur de la moyenne bootstrap par bloc de longueur l par rapport à la variance pour un bloc de longueur 1

Site	N_{valid}	N_{app}	Rapport des tailles de bloc[1]	Données simulées	Données de station de référence[2]	Deux types de données[2]
1	55 184	17 521	$26/1$	7,80	11,11	7,10
			$200/1$	13,97	19,04	13,62
2	21 827	8 762	$26/1$	5,60	5,56	6,47
			$200/1$	9,84	20,09	10,47
3	19 605	17 521	$26/1$	8,66	9,53	6,33
			$200/1$	18,17	22,25	16,00
4	34 532	17 521	$26/1$	8,27	10,07	7,43
			$200/1$	18,29	14,67	13,70
5	5 998	2 008	$26/1$	4,44	4,47	4,16
			$200/1$	2,31	5,82	2,97
6	12 329	17 521	$26/1$	9,53	-	-
			$200/1$	26,13	-	-
7	25 949	17 521	$26/1$	6,70	9,55	6,09
			$200/1$	10,81	14,79	9,21
8	5 998	2 613	$26/1$	5,11	-	-
			$200/1$	7,60	-	-
9	28 994	17 521	$26/1$	6,02	-	-
			$200/1$	10,27	-	-
10	23 008	17 521	$26/1$	7,20	-	-
			$200/1$	19,10	-	-
11	3 673	17 521	$26/1$	6,49	-	-
			$200/1$	12,21	-	-
12	5 999	2 671	$26/1$	5,81	-	-
			$200/1$	10,30	-	-
13	15 115	17 521	$26/1$	7,51	-	-
			$200/1$	16,43	-	-
14	26 718	17 521	$26/1$	7,76	-	-
			$200/1$	14,49	-	-
15	9 096	17 521	$26/1$	7,71	-	-
			$200/1$	16,52	-	-
16	11 532	17 521	$26/1$	6,44	-	-
			$200/1$	11,16	-	-
17	5 887	2 989	$26/1$	5,96	6,99	5,57
			$200/1$	9,07	14,67	9,04
18	5 999	2 030	$26/1$	6,48	8,03	5,93
			$200/1$	15,16	13,91	14,24

[1]Ici, 26 réfère à la taille $N^{\frac{1}{3}}$ et peut donc valoir moins de 26 dans quelques cas. Voir le tableau 5.3 pour les tailles utilisées.

[2]Les cases vides représentent les sites où nous ne possédions pas de données de stations de référence.

(Suite)

Site	N_{valid}	N_{app}	Rapport des tailles de bloc[1]	Données simulées	Données de station de référence[2]	Deux types de données[2]
19	7 999	2 086	$26/1$	4,81	4,23	3,58
			$200/1$	5,76	5,78	4,85
20	7 999	1 771	$26/1$	5,58	5,86	4,14
			$200/1$	5,54	6,41	6,06
21	15 999	3 321	$26/1$	6,01	4,14	5,03
			$200/1$	10,75	6,59	7,40
22	5 999	2 585	$26/1$	7,29	-	-
			$200/1$	27,23	-	-
23	14 999	4 698	$26/1$	3,19	-	-
			$200/1$	5,79	-	-
24	11 999	4 271	$26/1$	5,49	10,52	6,24
			$200/1$	9,47	22,91	9,50
25	3 999	1 988	$26/1$	4,00	-	-
			$200/1$	9,87	-	-
26	1 599	788	$26/1$	4,39	-	-
			$200/1$	9,81	-	-
27	2 299	1 230	$26/1$	5,50	-	-
			$200/1$	8,80	-	-
28	5 999	2 719	$26/1$	7,71	-	-
			$200/1$	15,62	-	-
29	8 999	3 833	$26/1$	6,40	9,49	6,37
			$200/1$	14,61	26,30	15,84
30	2 741	17 521	$26/1$	7,71	10,32	7,99
			$200/1$	12,89	23,70	12,14
31	16 863	17 521	$26/1$	6,17	10,26	6,19
			$200/1$	11,09	16,23	8,20

[1]Ici, 26 réfère à la taille $N^{\frac{1}{3}}$ et peut donc valoir moins de 26 dans quelques cas. Voir le tableau 5.3 pour les tailles utilisées.

[2]Les cases vides représentent les sites où nous ne possédions pas de données de stations de référence.

5.4.2.6. *Discussion*

On peut maintenant voir que la taille des blocs de résidus a un effet sur la variabilité de la moyenne long terme estimée, contrairement à l'erreur quadratique moyenne de prévision calculée à partir du bootstrap. De façon générale, l'utilisation de blocs d'erreurs bootstrap de taille $N_{app}^{\frac{1}{3}}$ mène à des variances estimées entre 3,19 et 9,53 fois supérieures aux variances estimées à partir de

blocs de taille 1. En ce qui a trait aux blocs de taille 200, ils mènent à des variances estimées entre 5,54 et 27,23 fois supérieures aux variances correspondant aux blocs de taille 1. On remarque que le ratio $200/1$ est à peu près toujours deux fois plus grand que le ratio $26/1$. Les résultats indiquent donc qu'il est important de tenir compte du délai de corrélation entre les observations dans le temps, si l'on désire utiliser le bootstrap pour obtenir une estimation de la variance de la moyenne long terme de la vitesse du vent, parce qu'un changement dans la taille de bloc changera l'estimation de la variance de façon importante. Il serait intéressant de tester ici d'autres tailles de blocs que celle proposée par Davison et Hinkley (voir par exemple Politis et Romano (1995)) et de tenter de déterminer de quelque façon que ce soit la taille optimale pour l'estimation de la variance de la moyenne long terme estimée.

5.5. Conclusion sur les prévisions de la vitesse du vent

La méthode MCP permet d'obtenir des prévisions de la vitesse du vent dans le passé ainsi qu'une estimation de la vitesse moyenne du vent à long terme. Dans ce chapitre, on a tenté de justifier dans un premier temps l'utilisation des données méso-échelle à partir de la validation croisée. On a d'abord pu voir à partir de la validation croisée et du calcul de l'erreur relative de la moyenne long terme estimée qu'il était préférable d'utiliser les données méso-échelle que les données provenant d'une station de référence seules dans un modèle de régression linéaire simple. C'est ce que la compagnie *Hatch* voulait déterminer, puisqu'ils n'utilisent présentement que les données de station de référence et veulent savoir s'il vaut la peine de payer pour des données méso-échelle. Les résultats ont aussi montré que l'erreur due à la prévision de la vitesse du vent dans le passé serait encore davantage diminuée si l'on utilise tant les données méso-échelle que celles provenant d'une station de référence dans un modèle de régression linéaire multiple. Ainsi, les données de station de référence ne sont pas inutiles et peuvent être utilisées conjointement aux données méso-échelle pour améliorer les prévisions de la vitesse du vent.

Nous avons ensuite présenté une méthodologie, toujours pour estimer l'erreur de prévision mais dans le cas où l'on possédait seulement des données y appartenant à l'échantillon d'apprentissage (donc seulement les deux dernières années, par exemple), cas où on ne peut pas faire de validation croisée. Des blocs d'erreurs bootstrap ont permis de tenir compte de la dépendance entre les observations, et des prévisions bootstrap faites à partir de résidus en blocs de diverses tailles ont permis d'obtenir des estimations de l'erreur quadratique

moyenne de prévision qu'on a comparées à celles obtenues par validation croisée. Les résultats n'ont pas démontré une grande différence quant à la taille de bloc utilisée mais on a pu voir une différence entre les trois modèles de régression linéaire, le modèle avec tous les prédicteurs ainsi que celui utilisant seulement les données méso-échelle étant ceux qui mènent aux résultats les plus similaires par rapport à la validation croisée où toutes les données étaient utilisées. En gros, nous avons obtenu des différences absolues relatives de la racine de l'EQMP contenues entre 5,78% et 9,69% et c'est donc la différence à laquelle on peut s'attendre lorsqu'on évalue la racine de l'EQMP à partir du bootstrap par rapport à la validation croisée.

Finalement, l'étude de la variabilité de la moyenne long terme estimée à partir du bootstrap a montré que la taille des blocs dans le bootstrap influençait bel et bien l'estimation de la variance de la moyenne long terme de la vitesse du vent estimée. On ne connaît toujours pas la taille de bloc parfaite mais l'on sait que, vue la différence lorsqu'on fait varier la taille, l'hypothèse d'indépendance entre les observations ne tient pas. Effectivement, les variances pour des tailles de $N_{app}^{\frac{1}{3}}$ ou de 200 ne sont pas similaires à la variance obtenue avec une taille de 1, correspondant à une structure de dépendance de délai 0, ces premières variances étant beaucoup plus grandes, comme on peut s'y attendre s'il y a de la dépendance dans les données temporelles. Ces résultats vont de pair avec ceux trouvés dans le chapitre 4, à la section 4.4.2.

CONCLUSION

L'estimation de la production d'énergie annuelle produite par une turbine éolienne se fait en plusieurs étapes, chacune de ces étapes influençant grandement l'estimation de la production d'énergie annuelle, ce qui peut donc mener à d'importantes différences selon la méthodologie utilisée pour l'estimation.

Dans le chapitre 3, on a tenté de déterminer, dans un premier temps, laquelle des méthodes d'extrapolation de la vitesse du vent mènerait aux meilleures estimations à la hauteur d'une turbine éolienne classique. On a pu trouver, à partir de l'erreur quadratique moyenne d'extrapolation, qu'il n'y avait pas de grande différence entre le coefficient de cisaillement local ou global et la méthode du point de référence ou de la régression. En n'utilisant que deux hauteurs (donc les sites 1 à 30), on pouvait déjà prévoir qu'il n'y aurait aucune différence entre la méthode du point de référence et de la régression avec coefficient de cisaillement local. Or, pour le site 31, nous possédions les vitesses de vent à une autre hauteur, de sorte que la méthode de la régression se distinguait maintenant de celle du point de référence, même avec un coefficient de cisaillement local. On a trouvé que les méthodes optimales étaient maintenant celle du point de référence avec coefficient de cisaillement local et la hauteur de référence la plus près du plus haut point et celle de la régression avec coefficient de cisaillement local. En plus de tester la prévisions aux dix minutes, on a voulu évaluer l'estimation de la moyenne long terme de la vitesse du vent. Les meilleures méthodes qui en sont ressorties sont celles du point de référence ou de la régression avec un coefficient de cisaillement local. On ne peut pas dire que l'une des méthodes se démarque particulièrement, mais l'on remarque que l'utilisation d'un coefficient de cisaillement local semble préférable à celle d'un coefficient global.

Au chapitre 4, on désirait entre autres évaluer la modélisation de la vitesse du vent par une distribution *Weibull*. Un test d'adéquation nous a d'abord permis d'observer que pour chacun des 31 sites, les probabilités que la vitesse

prenne des valeurs dans les classes de 1 à 25 m/s ne proviendraient pas d'une loi *Weibull*. Nous avons ensuite voulu vérifier si un ajustement de *Weibull* à chaque année ou même à chaque mois serait préférable, puisque l'industrie utilise présentement une seule distribution globale pour modéliser celle de la vitesse du vent. Pour ce faire, on a évalué la variabilité des paramètres ajustés sur des distributions annuelles ou mensuelles et une comparaison de ces variabilités nous a permis de remarquer, dans un premier temps, que les estimateurs de la variance asymptotique et échantillonnale ne menaient pas du tout aux mêmes résultats, et dans un second temps, à partir de simulations bootstrap, que la dépendance entre les données n'était pas nécessairement la raison des différences, mais plutôt que l'ajustement de *Weibull* globales ou annuelles ne suffisait pas à capturer toute la variabilité de la distribution de la vitesse du vent dans le temps. Aux vues des résultats, on pense que des paramètres mensuels seraient préférables mais on ne peut en être sûr, entre autres parce que la dépendance entre les observations n'a pas été considérée dans nos simulations bootstrap. Finalement, dans ce chapitre, on a aussi voulu comparer l'estimation de la production d'énergie annuelle faite à partir de la distribution empirique des vitesses de vent ou de la *Weibull* ajustée sur les vitesses de vent, afin de voir à quel point la modélisation par la *Weibull* engendre une différence au niveau de l'estimation. On a trouvé, pour certains sites, des différences de plus de 6% entre les deux estimations. Cela porte donc à réfléchir à la façon dont la production d'énergie devrait dorénavant être estimée.

Dans le chapitre 5, on a utilisé la validation croisée et le bootstrap afin d'évaluer la variabilité des prévisions du vent dans le passé, puisqu'il s'agit d'une autre composante de l'estimation de la production d'énergie annuelle d'une turbine éolienne. La validation croisée nous a permis de trouver que les données méso-échelle, utilisées dans un modèle avec deux prédicteurs (celles-ci en plus des données de stations de référence), permettaient de diminuer l'erreur quadratique moyenne de prévision de la vitesse du vent dans le passé. Nous avons présenté, toujours dans ce chapitre, une méthodologie permettant d'évaluer l'erreur de prévision de la vitesse du vent dans le cas où nous ne possédons pas de données mesurées dans l'échantillon de validation. C'est le bootstrap par blocs qui nous a permis d'évaluer l'erreur de cette façon, en plus de nous permettre de considérer la structure de corrélation entre les vitesses de vent. Nous avons trouvé, pour des blocs de taille $l = N_{app}^{1/3}$, que le modèle à deux prédicteurs menait à des différences relatives absolues moyennes de $5{,}78\%$ par rapport aux résultats avec la validation croisée, alors que le modèle

avec données méso-échelle menait plutôt à une différence moyenne de $6{,}50\%$ et le modèle avec données de station de référence, une moyenne de $9{,}69\%$. Ce sont donc les différences dans les racines des erreurs quadratiques moyennes de prévision auxquelles on peut s'attendre en utilisant le bootstrap lorsque nous n'avons que quelques années de données disponibles. De plus, le fait de faire varier la longueur des blocs dans le bootstrap nous a permis de remarquer encore une fois qu'il existe bel et bien de la dépendance entre les vitesses de vent, puisque les diverses tailles de blocs menaient à des résultats très différents en termes de variance de la moyenne long terme estimée.

Pour finir, nous croyons qu'il aurait aussi pu être intéressant de voir les choses d'un autre point de vue, d'abord en utilisant dès le début des paramètres changeant dans le temps (surtout qu'il est plausible, vus les résultats au chapitre 4, que ce soit préférable par rapport à des paramètres globaux). Nous aurions donc pu explorer des approches bayésiennes où les paramètres sont aléatoires. Dans un deuxième temps, nous aurions aussi pu considérer un point de vue davantage axé sur les séries chronologiques. En effet, sauf vers la fin du chapitre 4 et dans le bootstrap par bloc au chapitre 5, nous avons considéré l'indépendance entre les vitesses de vent puisque cela demeurait plus simple que l'ajustement de modèles de séries chronologiques aux données. Dans de futures études, on pourrait donc ajuster des modèles de séries chronologiques aux données et tenter de déterminer le délai pour lequel les vitesses de vent sont dépendantes entre elles. Cela nous permettrait par la suite d'ajuster la méthodologie du chapitre 5 avec des blocs de bonne taille, ou encore de prédire la vitesse du vent passée à partir des modèles trouvés.

Somme toute, les analyses entreprises dans ce mémoire ont permis de trouver que l'utilisation d'un coefficient de cisaillement local plutôt que global pourrait améliorer l'extrapolation de la vitesse du vent, que la modélisation de la distribution des vitesses de vent par une seule *Weibull* ne tient pas compte de la variabilité des paramètres dans le temps et qu'il faut donc faire davantage attention à cette étape puisque des paramètres variant de façon mensuelle semblent déjà être plus près de la réalité, qu'il semble être préférable d'utiliser la distribution réelle des vitesses de vent pour estimer la production d'énergie plutôt que la distribution de *Weibull* ajustée et finalement, qu'il pourrait être avantageux pour les ingénieurs éoliens d'utiliser des méthodes tenant compte de la structure de dépendance entre les vitesses de vent plutôt que de considérer les vitesses de vent comme étant indépendantes entre elles lors de l'estimation de la production d'énergie éolienne annuelle.

BIBLIOGRAPHIE

Burton, T., Jenkins, N., Sharpe, D. et Bossanyi, E. (2011). *Wind Energy Handbook*, Wiley, Chichester.

Davison, A.C. et Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.

Drobinski, P. et Coulais, C. (2012). Is the Weibull distribution really suited for wind statistics modeling and wind power evaluation ? *arXiv :1211.3853.*

Efron, B. (1979). Bootstrap methods : Another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

Johnson, N.L., Kotz, S. et Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1*, Wiley, New York.

Künsch, H.R. (1989). The jacknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17, 1217-1241.

Peterson, E.W. et Hennessey Jr, J.P. (1978). On the use of power laws for estimates of wind power potential. *Journal of Applied Meteorology*, 17, 390-394.

Politis, D.N. et Romano, J.P. (1995). Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis*, 16, 67-103.

Rice, J.A. (2007). *Mathematical Statistics and Data Analysis, third edition*, Duxbury, Berkeley.

Searle, S.R. (1971). *Linear Models*, Wiley, New York.

`http://eolienne.f4jr.org/eolienne_etude_theorique` Consulté le 9 décembre 2013.

`http://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html` Consulté le 17 mars 2014.

`http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html` Consulté le 24 août 2014.