Prévision court terme de la production éolienne par Machine learning

Mamadou Dione

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAG004

Thèse de doctorat

ENSAE

IP PARIS

Prévision court terme de la production éolienne par Machine Learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et
de l'administration économique

École doctorale n°574 EDMH
Spécialité de doctorat : Mathématiques Appliquées

Thèse présentée et soutenue à Montpellier, le 10 juin 2020, par

MAMADOU DIONE

Composition du Jury :

Erwan Le Pennec Professeur associé, École polytechnique	Président
François Husson Professeur, Agrocampus Rennes	Rapporteur
Sylvain Sardy Professeur, Université de Genève	Rapporteur
Eric Matzner-Løber Professeur des Universités, ENSAE Paris	Directeur de thèse
Anne Ruiz-Gazen Professeur, Université Toulouse 1 Capitole	Examinatrice
Philippe Alexandre ENGIE Green France	Invité

Remerciements

Cette thèse est l'aboutissement d'un long parcours durant lequel j'ai eu le soutien de beaucoup de personnes. A toutes ces personnes je voudrais les remercier chaleureusement et leurs témoigner ma gratitude.

J'adresse mes sincères remerciements à mon Directeur de thèse Eric Matzner-Løber pour avoir accepté de travailler avec moi et d'encadrer mes travaux. Tes orientations avec pédagogie ont été sans faille durant toute la thèse. Je te suis reconnaissant de tes conseils, de ta patience, de ta disponibilité à chaque fois que le besoin s'est fait sentir et de ton hospitalité.

Je remercie ENGIE Green pour la confiance qu'ils m'ont accordée pour mener ce projet de thèse. Particulièrement, je remercie Philippe Alexandre le Directeur de la DEEI de m'avoir accueilli au sein du service étude, expertise et innovation. Merci pour l'intégration et les moyens mis à disposition pour bien mener mes travaux. Je remercie Nicolas Girard et Carolina Penin pour tous les échanges que nous avons eus durant cette thèse. Vos contributions ont beaucoup aidé à enrichir mon travail. Merci aussi à Alexandre Le Lay de m'avoir fait visiter les éoliennes et de toute ton attention durant le déroulement de ma thèse. Je remercie également Christian Koessler et Benoit Buffard pour les échanges et les discussions sur le sujet. Un grand merci à tous les collègues de la DEEI : Valérie, Emilien, Guillaume, Guilhem, Vincent, Bichet, Marco, Marie, Alix, Gireg, Eléa, Anthony, ...

Je tiens aussi à remercier solennellement les rapporteurs de la thèse François Husson et Sylvain Sardy ainsi que les autres membres du jury Erwan Le Pennec et Anne Ruiz-Gazen d'avoir acceptés de lire et d'examiner mes travaux.

Je remercie tous les chercheurs du laboratoire de statistique du CREST. Merci à tous les doctorants pour les conversations qu'on a eues à l'ENSAE. Merci aussi à mes collègues de bureau.

Je remercie mes parents, particulièrement mon Père et ma Mère. Vous m'avez éduqué dans le travail et le respect. A la mémoire de mon père et de ma femme Amy, je dédie ce travail. A mon père, tu t'es beaucoup investi durant tout mon parcours. Tu m'as beaucoup motivé dans mes études. Si j'ai passé

autant de temps dans les études, c'est que tu me disais toujours qu'il fallait apprendre, apprendre et apprendre. Je te serai toujours reconnaissant. Amy, Tu as été un soutien infaillible sur tous les plans. Je remercie ma mère de m'avoir toujours soutenu et accompagné. Merci à mes sœurs et à mes frères. Je remercie mon soutien au quotidien, merci pour nos projets et nos moments de bonheur. Merci Fanny.
A mes amis, je vous remercie beaucoup pour votre compréhension, pour toutes nos discussions quotidiennes et pour tous les bons moments que nous avons passés ensemble.

Je remercie Anne-Laure Fougères pour le soutien, les conseils lors des procédures pour venir poursuivre mes études en France et pour l'intégration. Merci aussi au LABEX MILYON pour la bourse du master qui m'a permis de se lancer dans l'apprentissage statistique. Merci à tous les professeurs qui ont eu à m'enseigner.

Table des matières

Contextes de la thèse

Contexte politique

Force vive de l'eau ou du vent, rayonnement solaire, géothermie, chaleur du bois et des autres ressources de la biomasse, sans oublier les carburants végétaux et la valorisation des déchets, les énergies renouvelables prennent de multiples formes. Leur développement constitue un enjeu fort dans un contexte de demande croissante d'énergie, d'épuisement potentiel des ressources fossiles et de la nécessité de réduction des émissions de gaz à effet de serre. L'Union Européenne a décidé, à l'occasion de la refonte de la directive sur les énergies renouvelables adoptée fin 2018, d'atteindre une part d'énergies renouvelables dans sa consommation finale brute d'énergie d'au moins 32% en 2030. Cet objectif est également celui que la France s'est fixé à l'horizon 2030, dans le cadre de la loi relative à la transition énergétique pour la croissance verte [Moreau and Glorieux-Freminet, 2019].

Les énergies renouvelables représentent 10.7% de la consommation d'énergie primaire et 16.3% de la consommation finale brute d'énergie en France en 2017. Ces parts sont en progression régulière depuis une dizaine d'années. La croissance importante de la production primaire d'énergies renouvelables depuis 2005 ($+62\%$) est principalement due à l'essor des biocarburants, des pompes à chaleur et de la filière éolienne [Moreau and Glorieux-Freminet, 2019]. D'où la place importante de l'énergie éolienne dans le développement présent et futur des énergies renouvelables. Le développement de l'énergie éolienne engendre des enjeux économiques notamment la réglementation sur sa vente, l'exploitation du réseau en temps réel, la stabilité du réseau électrique de même que les exigences et les coûts des services auxiliaires.

Le contexte économique

Contrairement au photovoltaïque soumis à appel d'offres depuis 2012, l'énergie éolienne était encore rémunérée par un tarif d'achat fixe. En effet,

pour développer la filière éolienne, l'État français avait mis en place en 2000 et jusqu'en 2015 un dispositif incitatif : contrat d'obligation d'achat. Dans le cadre de ces contrats EDF (Électricité De France) et, si les installations de production sont raccordées aux réseaux publics de distribution dans leurs zones de desserte, les entreprises locales de distribution, doivent acheter l'électricité produite à partir de l'énergie éolienne aux exploitants qui en font la demande, à un tarif d'achat fixé par arrêté [FEE, 2019].
Mais à partir de janvier 2016, le dispositif de soutien à l'éolien terrestre a évolué vers le dispositif de complément de rémunération mis en place par la loi relative à la transition énergétique pour la croissance verte. Dans le cadre de ces contrats, l'électricité produite par les installations est vendue directement par le producteur sur le marché de l'électricité, puis à se voir verser par l'État une prime basée sur un « target price », défini a priori en fonction des prix du marché.
Or la vente sur le marché impose d'annoncer à l'avance (environ 24h avant) la quantité d'électricité qui sera livrée en temps et en heure à EDF, grâce à une prévision court terme de la production. Ce nouveau système de rémunération commence à être appliqué à certains parcs éoliens. Il met ainsi le sujet de la prévision court terme au cœur de la valorisation de l'énergie éolienne.

Engie Green France

ENGIE Green est une filiale détenue à 100% par le groupe ENGIE. ENGIE Green est née de la fusion de La Compagnie du Vent au 15 décembre 2017 et de l'intégration des activités de développement, d'exploitation et de maintenance de Solairedirect en France. C'est le premier acteur de l'éolien terrestre et solaire du groupe ENGIE en France. ENGIE Green regroupe une expertise complète pour le développement, la conception, la construction, l'exploitation et la maintenance des sites éoliens et photovoltaïques. ENGIE Green en quelques chiffres clés (au 31 décembre 2017) :

- **1333** MW éoliens installés et exploités (91 parcs, 701 éoliennes)
- **86.6** MW éoliens exploités pour le compte de tiers (9 parcs, 46 éoliennes)
- **862** MW solaires installés et exploités (101 centrales)
- **3000** MW en développement

- près de 1 700 000 de personnes alimentées en électricité renouvelable par an.

- une équipe multidisciplinaire de plus de 400 collaborateurs basés sur les territoires.

FIGURE 1 – Parc éolien ENGIE Green

ENGIE Green exploite plus d'une centaine de parcs éoliens (exemple de parc éolien, figure 1) dans lesquels sont installés aussi des pylônes de mesure avec des anémomètres et des girouettes qui mesurent respectivement la vitesse et la direction du vent à des fréquences minutes ou dix minutes. Sur chaque nacelle des éoliennes composant un parc éolien, il y aussi des anémomètres et des girouettes qui mesurent la vitesse du vent et la direction du vent à la hauteur de l'éolienne. A l'aide d'un ordinateur placé sur chaque parc éolien, des paramètres liés à la production des éoliennes (production, vitesse du vent, état de l'éolienne par exemple) sont enregistrés et rapatriés dans des bases de données. Ces données servaient à faire des études de potentiel éolien, des

analyses de performance et de manière générale à suivre la production des parcs éoliens. Comme expliqué dans la section précédente, le besoin de prévisions de production s'est récemment exprimé au niveau des producteurs avec la rupture des contrats d'obligation d'achat.
L'objectif de cette thèse est d'analyser ces données avec des modèles statistiques et de proposer un modèle de prévision court terme (de 24 heures à environ 47 heures) de la production des parcs éoliens. Nous avons travaillé avec les données (de vent, de production, ...) provenant des parcs éoliens et des données de modèles météorologiques (vent, température, pression, ...).

La prévision de la production éolienne

Selon l'horizon de prévision, on distingue les trois catégories de prévision suivantes :

- Prévision immédiate à court terme (de l'instantané à environ 8 heures),
- Prévision à court terme (de 24 heures à environ 47 heures) qui est l'objet de cette thèse,
- Prévision à long terme.

Les applications de ces trois catégories de prévision sont résumées dans le tableau 1.

TABLE 1 – Time-horizon classification for wind forecasting, source [Wang et al., 2011]

Time-scale	Range	Applications
Immediate-short-term	8hours-ahead	- Real-time grid - Regulation actions
Short-term	Day-ahead	- Economic load dispatch planning - Load reasonable decisions - Operational security in electricity market
Long-term	Multiple-days-ahead	- Maintenance planning - Operation management - Optimal operating cost

La prévision éolienne peut être aussi classifiée selon qu'elle intègre les modèles numériques de prévision du temps (Numerical Weather Prediction model, NWP en anglais) ou pas [Giebel et al., 2011]. L'intégration des données de modèle NWP dépend de l'horizon de prévision. Pour un horizon de prévision inférieur à 3 à 6 heures environ, les séries chronologiques utilisant juste les données du SCADA (Supervisory Control And Data Aquisition, voir section 1.2.1) donnent des résultats satisfaisants. Après environ 3 à 6 heures, les modèles intégrant les données NWP surpassent les approches des séries chronologiques [Giebel et al., 2011]. Par conséquent, tous les modèles utilisés

par les services publics utilisent l'approche avec le NWP.
Il existe deux écoles de pensée différentes en matière de prévision à court terme avec NWP : l'approche physique et l'approche statistique.

- L'approche physique : les modèles physiques tentent d'utiliser les considérations physiques aussi longtemps que possible pour parvenir à la meilleure estimation possible de la vitesse du vent local avant d'utiliser les Model Output Statistics (MOS) ou différentes techniques statistiques relativement simples pour réduire l'erreur restante.
- L'approche statistique : les modèles statistiques dans leur forme pure tentent de trouver les relations entre une multitude de variables explicatives, y compris les résultats des modèles NWP, et les données de puissance mesurées, en utilisant généralement des techniques récursives. Souvent, des modèles de machine learning sont utilisés.

Les modèles qui s'appuient aussi sur la modélisation de chaque éolienne sur la base des équations de courbe de puissance :

$$P_w = 0.5\rho v^3,$$

où P_w la densité de la puissance, ρ la densité de l'air et v la composante horizontale de la vitesse du vent sont aussi considérés comme des modèles physiques [Costa et al., 2008].

Résumé des travaux en anglais

Summary of work in Chapter 1

The presentation of the data used for this study and the comparison of the performance of the two models GFS and ECMWF is the subject of Chapter 1. We used measured data from the wind farms and data from meteorological models, all related to wind energy production. These are mainly the U and V components of wind, wind speed and wind direction which are obtained from U and V, pressure, temperature, production, measured wind speeds, etc.

We evaluated the two models : GFS and ECMWF with the data from the pylon over the year 2017 :

- first by comparing the wind speed of each model to the realised wind speed on the wind farm. The ECMWF model made fewer wind forecast errors over the year 2017. We obtained similar results with temperature.
- Second by comparing the components (U, V). The first result is confirmed by the study of the bivariate distributions of the (U, V) components of the wind. After estimating the bivariate distributions of (U, V) of the GFS, ECMWF and measured (U, V) at each forecast horizon, we used the Kullback-Leibler divergence to show that the bivariate distribution of the ECMWF wind is "closer" to that of the pylon wind than that of the GFS wind.
- Third using a more global analysis by comparing the groups of variables GFS and ECMWF with actual measurements. The data for which actual measurements are available are U, V, temperature and pressure. Thus we have shown by the canonical correlation analysis that the ECMWF forecast table is more correlated with actual measurements. This confirms the results on the individual comparison between wind speeds and temperature.

We conclude following others that ECMWF forecasting are more relevant for wind farm predictions and we will use them during our work.

The work of this chapter was presented at the "17th Wind integration workshop : 17-19 October 2018, Stockholm (Sweden)".

Summary of work in Chapter 2

The work in this chapter focuses on short-term forecasting methods for wind generation. The major contribution is indirect modelling (Figure 3) which

has given better results than direct modelling (Figure 2). Indeed, we used production data, real wind measurements and ECMWF model data to predict 24 to 47 hour production using both direct and indirect approaches. The results were presented at ITISE 2018 (International conference on Time Series and Forecasting) and published in [Dione and Matzner-Løber, 2019].

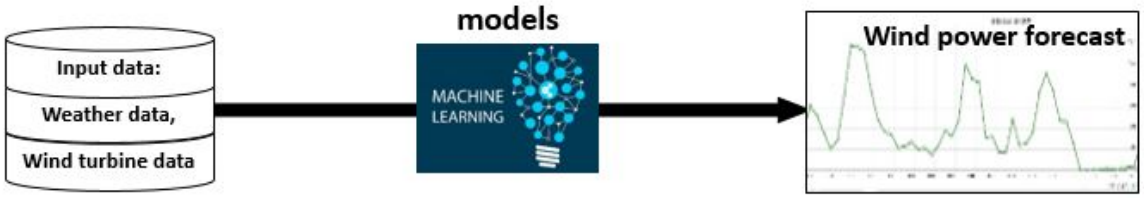


FIGURE 2 – Direct Approach to wind power forecasting.

FIGURE 3 – Indirect Approach to wind power forecasting

We used the ECMWF model forecasts at the nearest grid point, at the four nearest grid points and at the 16 nearest grid points of the wind farms (Figure 4).

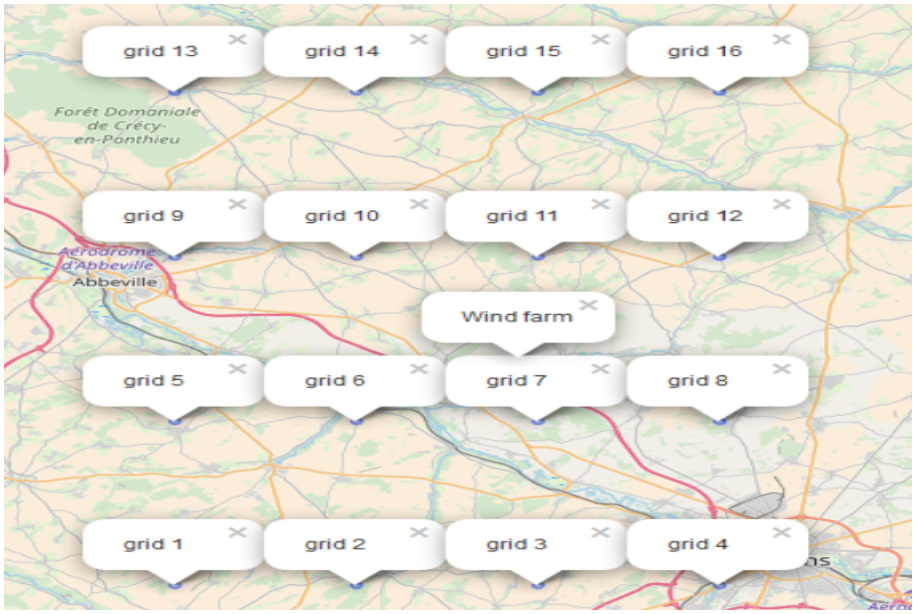


FIGURE 4 – ECMWF mesh around a wind farm.

Applying random forests, bagging, boosting, SVMs, GAMs, we obtain good performances forecast horizon for 24 for 47 hours with a mean absolute error of 6 to 10% as illustrated for wind farm 1, figure 5.

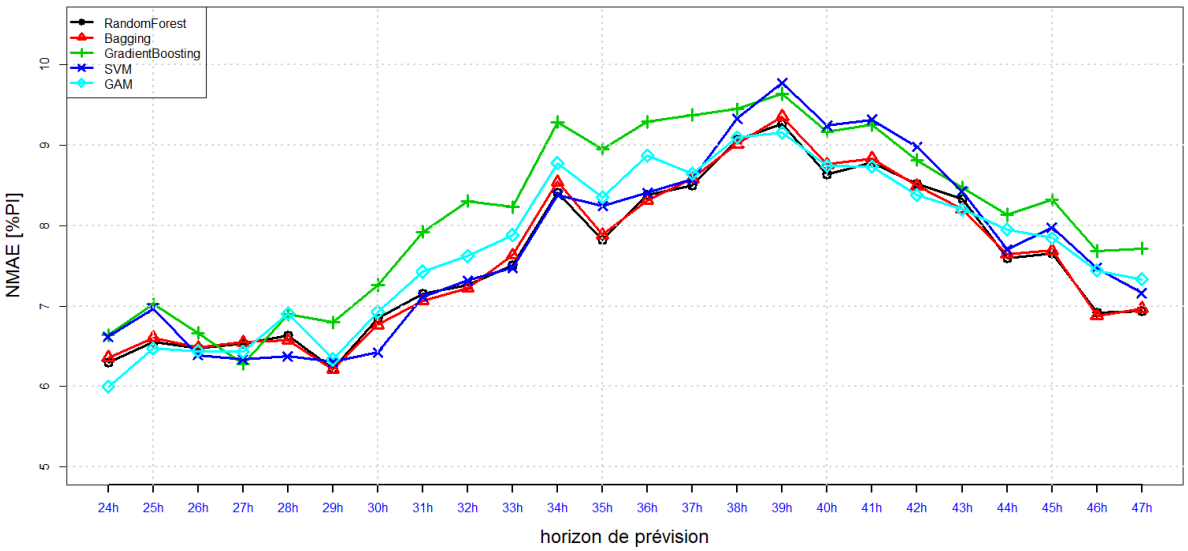


FIGURE 5 – NMAE per horizon, indirect approach with 16 grid points : Wind farm 1.

We have the following conclusions :

- if the real power curve is « smooth », the indirect model generally performs better than the direct model for each of the applied machine

learning algorithms ;

- the integration of several grid points at model input has shown that generally we have a reduction of the MAE but that with some machine learning algorithms, this result is not always true. A reduction in wind production prediction errors is not systematically obtained by taking into account several grid points ; it depends on the learning algorithm used ;
- An improvement of the results is possible by integrating business aspects such as wind dynamics with feature engineering. These aspects are the subject of the chapter 3.

Summary of work in Chapter 3

In order to improve wind production forecasts, it is important to take into account the business aspects, particularly meteorology. Based on meteorological forecasts, a certain number of phenomena can be modelled by post-engineering to serve as co-variables and potentially improve production forecasts. In the chapter 2 and in most of the studies on wind production forecasting, forecasts are made without taking into account correctly the space-time dependencies observed in the field. However, in recent years, the spatio-temporal structure has begun to be integrated into wind forecasting [Tastu et al., 2011].
Spatial and temporal modelling of weather forecasts has led to an improvement in forecast errors, as shown in the example in the table 2 of forecast errors for wind farm 1 for the year 2017. These results on the space-time dynamics of wind were published in [Dione and Matzner-Løber, 2019].

TABLE 2 – wind farm 1 : $NMAE$ indirect approach with and without taking into account space-time derivatives

Models	1 grid point	4 grid points	16 grid points
RF indirect	7.70	7.76	7.62
Integration of space-time derivatives			
RF indirect	7.45	7.57	7.44

However, the contribution of turbulence is very small compared to the contribution of space-time derivatives. Moreover, by combining turbulence and space-time derivatives there is no improvement in wind production forecasts. Indeed there is a strong correlation of wind speeds from one grid point to

another and thus the standard deviation that intervenes in the definition of the intensity of spatial turbulence is low in general (97% of the time less than 0.4). As a result, the impact of turbulence from the meteorological model on wind forecasting is negligible.

Until now, the choice of grid points has been made on the basis of geographical proximity to the wind farm whose production we wish to forecast. The question of the selection of the grid points has been asked in a statistical way. Given that at each grid point there are several meteorological variables, this problem amounts to a selection of a group of variables. [Gregorutti et al., 2015] have recently adapted the permuted significance measure of [Breiman, 2001] for groups of variables to select groups of variables in the context of random forests. Based on the variable group importance measure, the authors proposed a method for selecting multiple functional variables. We use this method in the case of weather variable group selection. The results showed that in general the closest grid points are more important. The distribution of the importance of grid points shows that five grid points are generally more important. These grid points are grid points 8, 7, 12, 11 and 10, which are among the grid points closest to the wind farm (see the grid around the wind farm, figure 4). Note that grid point 6 which is part of the nearest grid points is not in the group of the most important points. It is in this sense that the selection of grid points is useful instead of reasoning only in terms of proximity in the choice of grid points for the forecast of a wind farm.

Always with the aim of improving forecasts or anticipating the impact of meteorology, we were interested in the stability of weather forecasts. The question that arose is the following : knowing the forecasts for the next 24 hours and the forecasts for the last 24 hours (or even the last few days), can we anticipate the stability or dynamics of the wind ?

The results showed that when meteorology is stable (strong canonical correlation between past and future forecasts), production forecast errors tend to decrease, whereas unstable meteorology induces large production forecast errors in times of strong winds.

Summary of work in Chapter 4

Chapter 4 examines the problem of forecasting intervals with random forests. Three possible solutions exist in the literature : quantile regression forests ([Meinshausen, 2006]), estimation of the variance of random forests ([Sexton and Laake, 2009]) and estimation of the distribution of prediction er-

rors ([Lu and Hardin, 2017]). We cannot directly apply these methods to obtain a production forecast interval. We have a two-step model ; a random forest model for wind turbine height wind forecast and a real power curve model (estimated by a spline) to convert the wind forecast into a production forecast. For this reason, we decided to estimate the power curve uncertainty and random forest uncertainty to construct a production forecast interval. The results show that by combining these two uncertainties, we obtain wind production forecast intervals with a coverage probability very close to the nominal coverage probability. Figure 6 is an example of a prediction interval. These results was presented at the "13th International Conference on CFE-CMStatistics 2019".

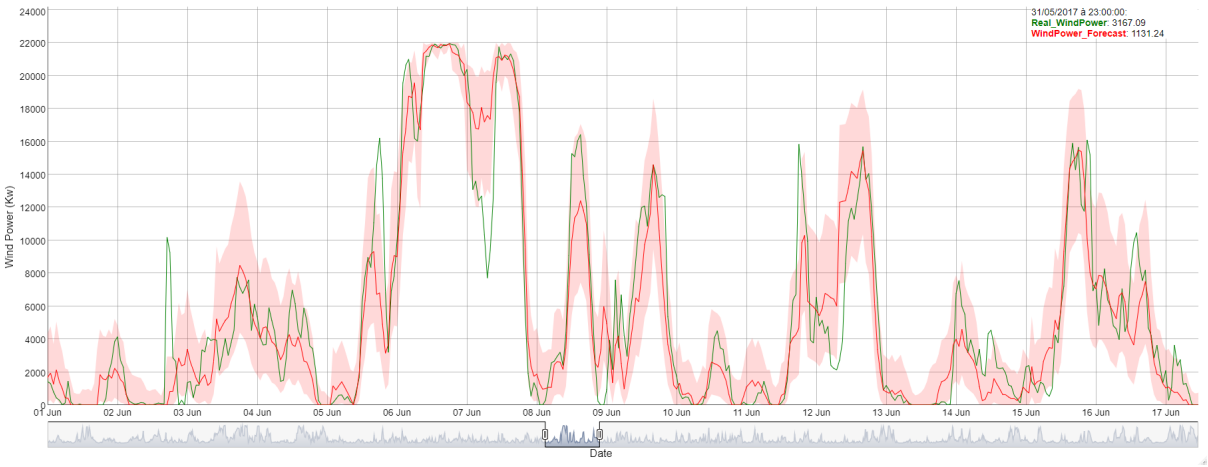


FIGURE 6 – Prediction interval $MSPE2 + Spline$.

Our forecast error evaluation methods (NMAE, NRMSE) compare pairs of time series, forecasts and observations and measure the vertical distance between these two time series. Another very important aspect to be taken into account in the evaluation of forecast errors is the horizontal offset or time distance between forecasts and observations. Indeed, a model can correctly forecast peak production but with a certain time lag. Knowledge of this type of error is useful for energy sales and industrial maintenance on wind farm.
The Temporal Distortion Index (TDI) measures the dissimilarity between time series ([Frías-Paredes et al., 2016] and [Frías-Paredes et al., 2017]). [Gastón et al., 2017] used the TDI to analyze solar irradiation forecasts. We used this index to take into account potential time lags between production forecasts and realizations. The results showed that the time misalignment appears to be largely responsible for the observed forecast error on the wind production forecast.

Summary of work in Chapter 5

Chapter 5 discusses the industrialization of the wind forecasting model. In the Darwin platform (unique digital platform to optimize ENGIE's renewable assets worldwide, figure 7), the Darwin Forecaster service is the industrial development of the work of this thesis. Thanks to this service, ENGIE has internal production forecasts which are mostly better than those of its forecast provider.

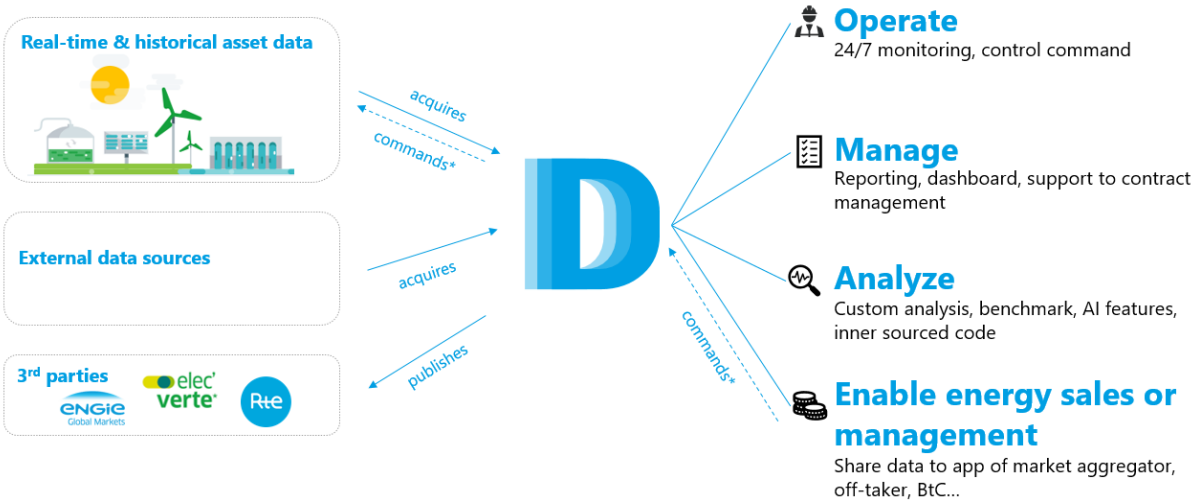


FIGURE 7 – Summary of Darwin's services. Source ENGIE.

The internal forecasts calculated in Darwin drastically lower the bill : the total costs for storage, data processing, training and maintenance of the machines as well as the smoothed cost for development should amount to 167 019 euros in 2020, three times less than for an external service (table 5.5). The net gain is therefore 396 731 euros in the first year. By 2021, Engie has already secured +9 GW of wind and solar projects. Beyond that, the Renewable Global Business Line (RGBL) projections foresee an annual capacity increase of +3 GW. In addition, if we decided to offer this service to external operators, we could generate additional revenues.

Introduction

La loi de transition énergétique définie par l'État a des implications précises sur les énergies renouvelables, en particulier sur son mécanisme de rémunération. Jusqu'en 2015, un contrat d'obligation d'achat permettait de vendre l'électricité d'origine éolienne selon un tarif fixe. A partir de 2015 avec la rupture des contrats d'obligation d'achat, il faudra vendre cette électricité sur le marché (selon des tarifs variables) avant d'obtenir un complément de rémunération destiné à diminuer le risque. Cette vente sur le marché requiert d'annoncer à l'avance (environ 24 heures à l'avance) la production qui sera livrée sur le réseau, donc de savoir prédire (à court terme) cette production. Pour cette raison, la prévision court terme du vent et de la production associée devient un enjeu important. C'est pourquoi ENGIE Green a décidé de lancer une thèse sur le sujet.

L'objectif de cette thèse est de prévoir la production P d'un parc éolien à partir de données météorologiques pour un horizon de 24 heures à 47 heures (prévision day-ahead). Ces données météorologiques que nous appellerons souvent prévisions météorologiques ne sont pas des mesures réelles en tant que telles mais des données issues de modèles de simulation numérique des organismes météorologiques. Au niveau industriel, la prévision court terme permet de répondre à deux besoins fondamentaux : la vente d'électricité sur le marché et la maintenance des parcs éoliens.

Les deux grands modèles de simulation numérique ECMWF et GFS sont très souvent utilisés dans la prévision court terme de la production. Ces modèles fournissent des prévisions de variables atmosphériques comme le vent (les composantes est-ouest U et nord-sud V du vent), la température, la pression sur l'ensemble du globe avec un maillage de 0.25°en latitude et en longitude (environ $28km \times 18km$) pour GFS et un maillage de 0.125°en latitude et en longitude (environ $13.8km \times 8.9km$) pour ECMWF. A partir des composantes U et V on peut obtenir la vitesse du vent W qui est la norme de U et V. La figure 8 montre le maillage du modèle ECMWF au tour d'un parc éolien en guise d'exemple. Chaque simulation (ou run) de ces modèles donne les prévisions

de variables atmosphériques jusqu'à un horizon de dix jours. Il existe quatre runs par jour (00h, 06h, 12h, et 18h) pour chaque modèle.

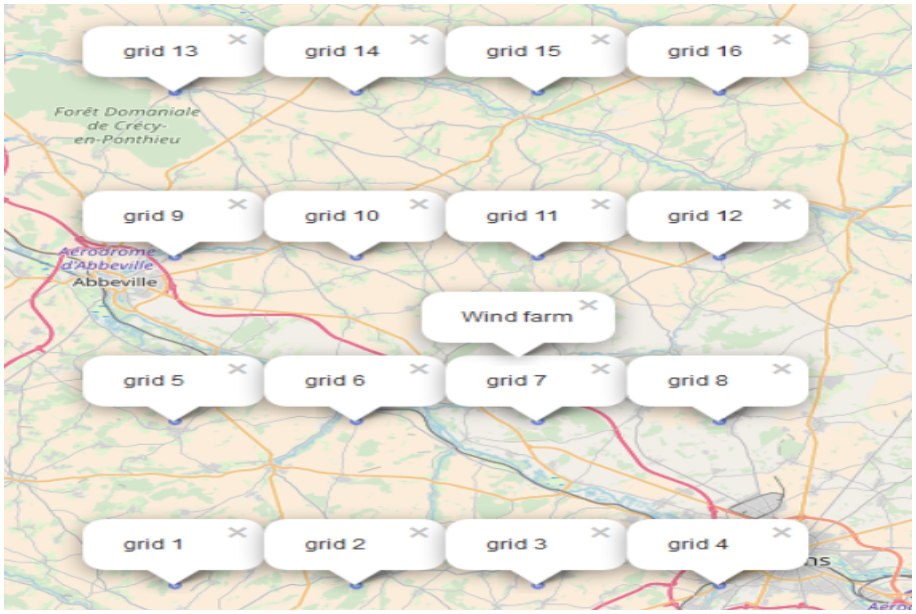


FIGURE 8 – Maillage ECMWF au tour d'un parc éolien.

Nous avons archivé toutes les données de prévision de ces modèles pour le RUN de minuit et pour un horizon de 47h. Il nous a semblé important de comparer la précision de ces données issues de ces deux modèles ECMWF et GFS sur les fermes éoliennes de l'étude. Ce travail de comparaison a été mené au chapitre 1. Nous avons comparé les prévisions via des indicateurs classiques d'erreur, mais aussi via l'analyse des densités du couple (U, V), et globalement via l'analyse canonique. Le modèle ECMWF donne des prévisions (ou données) plus précises.

Dans le chapitre 2 à partir des données de production, des anémomètres des éoliennes et des données issues du modèle ECMWF, nous avons prévu à un jour donnée, la production d'un parc pour le jour suivant. Nous avons appliqué des algorithmes de machine learning pour prévoir la production de parcs éoliens. Nous avons mis l'accent sur l'approche directe qui consiste à prévoir directement la production d'un parc et l'approche indirecte qui consiste à prévoir le vent à hauteur des éoliennes puis de le convertir en prévision de production à l'aide de la courbe de puissance réelle. Les résultats sont dans le même ordre de grandeur que les résultats publiés dans la littérature. Les résultats ont aussi montré que si la courbe de puissance réelle est très lisse, le modèle indirect est généralement plus performant que le modèle direct.

Cependant, afin de tenir compte de la dynamique du vent et comme nous

ne disposons pas d'indicateur comme par exemple l'absence ou la présence de rafales, nous avons créé de nouvelles variables. Le chapitre 3 est donc dédié en grande partie à l'intégration de la dynamique du vent afin d'améliorer les erreurs de prévision. Nous y avons aussi abordé la sélection de point de grille par forêt aléatoire.

Toute prévision est importante si nous connaissons l'intervalle de prévision associé à la valeur de la prévision. Par conséquent, nous avons étudié dans le chapitre 4 les intervalles de prévision en nous basant sur les forêts de régression quantile et l'estimation de la distribution des erreurs de prévision. Nous avons intégré les incertitudes sur la courbe de puissance car nous avons retenu le modèle indirect en deux étapes ; un modèle de forêt aléatoire pour la prévision des vitesses du vent à la hauteur des éoliennes et un modèle de courbe de puissance réelle (estimée par une spline) pour convertir le vent prévu en une prévision de production. Les résultats montrent qu'en combinant ces deux incertitudes, nous obtenons des intervalles de prévision de la production éolienne avec une probabilité de couverture très proche de la probabilité de couverture nominale.

Le chapitre 5 aborde le développement industriel des travaux de la thèse. Dans la plateforme Darwin (plateforme numérique unique pour optimiser les actifs renouvelables d'ENGIE dans le monde entier, figure 9), le service Darwin Forecaster est le développement industriel des travaux de cette thèse. Grâce à ce service, ENGIE dispose de prévisions de production internes.

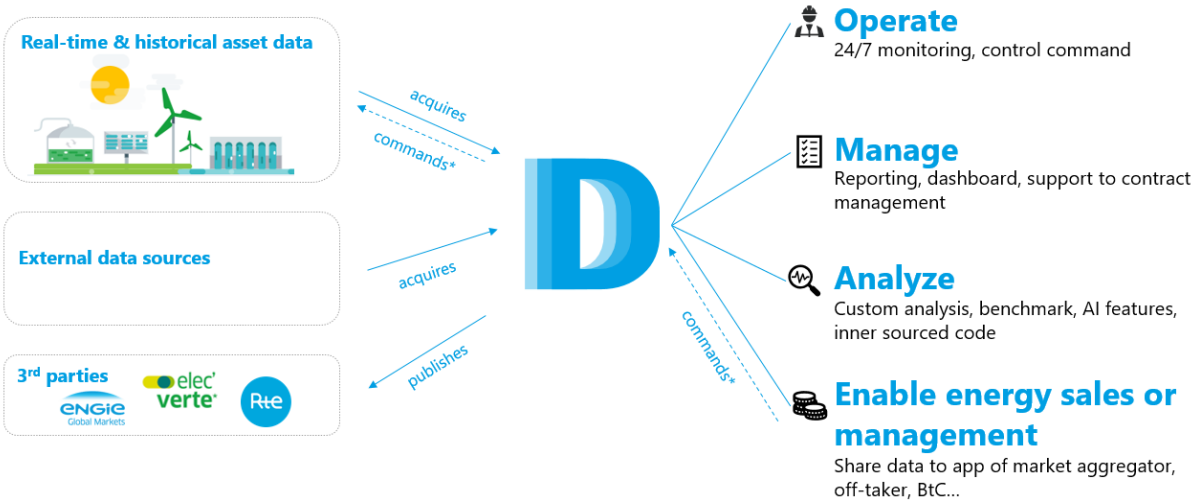


FIGURE 9 – Résumé des services de Darwin. Source ENGIE.

La stabilité du modèle de prévision dans le temps, la comparaison des prévisions du modèle interne avec les prévisions du fournisseur et les résultats

attendus ou observés sur l'industrialisation du modèle sont aussi développés dans le chapitre 5.

Chapitre 1

Présentation des données, comparaison des modèles météorologiques

Sommaire

1.1 Introduction

La Loi de Transition Énergétique définie par l'État Français a des implications précises sur les Énergies Renouvelables, en particulier sur son mécanisme de rémunération. Jusque-là, un contrat d'obligation d'achat permettait de vendre l'électricité d'origine éolienne selon un tarif fixe. Désormais, il faudra vendre cette électricité sur le Marché (selon des tarifs variables) avant d'obtenir un complément de rémunération destiné à diminuer le risque. Cette vente sur

le Marché requiert d'annoncer à l'avance (environ 24h avant) la production qui sera livrée sur le réseau et une surestimation ou une sous-estimation sera exposée à des pénalités. Il existe donc un besoin énorme de prévisions précises. A cet effet, nous allons nous servir de données météorologiques pour les transformer en prévisions de production éolienne par des modèles statistiques. Or, plusieurs sources météorologiques fournissent des variables atmosphériques (vent, température, pression, humidité, ...) à l'échelle mondiale ou locale. On dispose d'historiques de prévisions de variables atmosphériques provenant du modèle GFS (Global Forecast System 1.2.2) et du modèle ECMWF (European Centre for Medium-Range Weather Forecasts 1.2.3) que nous allons décrire puis comparer car ces deux modèles sont beaucoup utilisés dans la prévision éolienne. Les travaux de ce chapitre ont fait l'objet d'une présentation au "17th Wind integration workshop : 17-19 October 2018, Stockholm (Sweden)".
Les données sont présentées dans la section 1.2 et la comparaison des deux modèles de prévision GFS et ECMWF dans la section 1.3. L'évaluation des prévisions du vent des modèles GFS et ECMWF est faite dans la même section. La section 1.4 étudie par l'analyse canonique les relations canoniques des deux modèles météorologiques avec les mesures réelles. Une synthèse des résultats est effectuée en section 1.5.

1.2 Les Données

Les données proviennent de systèmes d'acquisition différents et fournissent des informations relatives à la production de l'énergie éolienne. La description englobe entre autres, les moyens d'acquisition, les fréquences et les durées de mise en disponibilité sur le réseau. On distingue les données issues de modèles météorologiques (GFS et ECMWF) et celles collectées au niveau des parcs éoliens (SCADA des éoliennes et des pylônes de mesures).

1.2.1 Les données provenant des parcs éoliens

Les données des éoliennes

Ces données sont collectées par le système de contrôle et d'acquisition de données (SCADA : Supervisory Control And Data Aquisition). Ce système est un logiciel destiné à la collecte de données en temps réel dans des sites distants en vue de contrôler les équipements et les conditions d'exploitation. A l'aide d'un ordinateur placé sur chaque parc éolien, une partie des variables (production, vitesse du vent, état de l'éolienne par exemple) provenant des éo-

liennes est enregistrée. Un anémomètre et une girouette placés sur la nacelle de chaque éolienne permettent de mesurer respectivement les vitesses et les directions du vent. Cependant des obligations contractuelles font des fois que les mesures des directions ne sont pas accessibles. Sur les parcs éoliens étudiés, le mécanisme de téléchargement se fait par une requête de l'ordinateur à chaque éolienne (à tour de rôle) et une réponse de ces dernières en fournissant les données (production, vitesse du vent, etc.) instantanées par minute ou la moyenne sur la minute (moyenne des 60 dernières valeurs relevées chaque seconde affectée à la soixantième seconde), selon la technologie des machines. Cette opération prend une vingtaine de secondes environ. Ces données fournies par les éoliennes sont enregistrées dans le disque dur de l'ordinateur sur le site.
Les données stockées sur l'ordinateur du site sont par la suite rapatriées dans une base de données qui interroge toutes les trois heures les ordinateurs sur les sites éoliens. Après rapatriement de ces données, elles sont lues à l'aide d'un robot qui les communique à un logiciel interne. Ce logiciel permet entre autres, le calcul de disponibilités par éolienne, par parc ou par zone pour une période de temps personnalisée, la détection des périodes d'arrêts, la catégorisation précise de chaque arrêt. La finalité des traitements est de savoir à chaque instant (chaque minute) si la machine est en fonctionnement ou en arrêt et en définir la raison (catégorisation). Le calcul de l'indice de fonctionnement décrit par la figure A.1 et le tableau A.3 (en annexe A) est un post-traitement qui permet d'affecter la valeur 1 à l'indice si l'éolienne est en fonctionnement et 0 si elle est en arrêt. Ce logiciel sert aussi d'outils de récupération des données. Avec la fusion en 2017 et suite à des changements internes, l'indice de fonctionnement n'est plus disponible dans les bases de données actuelles (2020). Le tableau A.1 décrit l'ensemble des variables provenant des éoliennes qui sont téléchargeables au final à partir du logiciel.

Les données des pylônes de mesures

Les données des pylônes (vitesse et direction du vent par exemple, annexe A.2) sont récupérées presque par le même mécanisme précédent, à la différence qu'elles sont enregistrées par moyenne de dix minutes, à des niveaux différents par rapport au sol (10m, 20m, 30m, ...) en fonction de la hauteur du pylône.

FIGURE 1.1 – Pylône de mesure

La moyenne dix minutes est la moyenne des 600 valeurs relevées à la seconde, qui est affectée à la dixième minute. Ces données sont enregistrées dans le disque dur du PC sur le parc qui, par le même mécanisme que celui décrit dans la section 1.2.1 les rapatrie vers une base de données puis dans un autre logiciel interne.

1.2.2 Les données du modèle GFS

Le Global Forecast System (GFS) est un modèle de prévisions météorologiques produit par les National Centers for Environmental Prediction (NCEP) dépendant du NWS (National Weather Service) des USA. Des centaines de variables atmosphériques et de terrain sont disponibles, depuis les températures, les vents, les précipitations jusqu'à l'humidité au sol et la concentration de l'ozone atmosphérique. Le monde entier est couvert par le modèle. Jusqu'au mois de mars 2015 GFS fournissait des prévisions tri-horaires avec une résolution de 0.5°en latitude et en longitude (environ $56km \times 36km$) avec un horizon de prévision allant jusqu'à dix jours.
En réponse à l'augmentation des ressources informatiques et à l'évolution de l'architecture informatique au NCEP, GFS a évolué en avril 2015 vers une résolution plus fine (de 0.5°à 0.25°) puis en octobre 2017 en fournissant des prévisions horaires avec toujours la même résolution 0.25°(environ $28km \times 18km$). Cette dernière évolution fournit aussi les prévisions de U et V à 80m et 100m en plus du niveau 10m du modèle GFS 0.25°tri-horaire. L'assimilation globale des données et les prévisions sont faites quatre fois par jour (00 : 00, 06 : 00, 12 : 00 et 18 : 00 UTC). Les données sont disponibles sur le site

http://nomads.ncep.noaa.gov environ 4h après chaque simulation (ou run) du modèle. Une partie (voir tableau 1.1) des données du modèle GFS 0.25 horaire issues du run de minuit est utilisée dans la suite. Cependant d'autres variables météorologiques sont disponibles sur le site du NCEP.

TABLE 1.1 – GFS data

Variables	Descriptions	Unités
U_10m	Composante est-ouest du vent à 10m du sol	m/s
V_10m	Composante nord-sud du vent à 10m du sol	m/s
U_80m	Composante est-ouest du vent à 80m du sol	m/s
V_80m	Composante nord-sud du vent à 80m du sol	m/s
U_100m	Composante est-ouest du vent à 100m du sol	m/s
V_100m	Composante nord-sud du vent à 100m du sol	m/s
T_2m	Température à 2m	Kelvin
Pr_80m	Pression à 80m	Pa
RH_2m	Humidité relative à 2m	%
GUST	Rafale de vent au sol	m/s

1.2.3 Les données du modèle ECMWF

ECMWF (European Center for Medium-Range Weather Forecasts) est une organisation intergouvernementale indépendante financée par 34 Etats européens. Ces données sont intégralement mises à la disposition des services météorologiques nationaux des Etats-membres. Le Centre propose également un catalogue de produits de prévision qui peuvent être achetés. Le modèle ECMWF fournit des prévisions horaires avec une résolution de 0.125°en latitude et en longitude (environ $13.8km \times 8.9km$) et quatre runs par jour (00h, 06h, 12h et 18h en UTC). Les prévisions de chaque run débutent 3h après l'heure du run pour un horizon de prévision pouvant aller jusqu'à dix jours. Ce modèle fournit entre autres les prévisions des composantes est-ouest U et nord-sud V du vent à 10m et 100m au-dessus du sol. On dispose uniquement d'une partie des variables météorologiques du modèle ECMWF qui sont liées à la production éolienne. Les variables utilisées dans ces travaux sont décrites dans le tableau 1.2.

TABLE 1.2 – ECMWF data

Variables	Descriptions	Unités
U_10m	Composante est-ouest du vent à 10m du sol	m/s
V_10m	Composante nord-sud du vent à 10m du sol	m/s
U_100m	Composante est-ouest du vent à 100m du sol	m/s
V_100m	Composante nord-sud du vent à 100m du sol	m/s
T_2m	Température à 2m du sol	deg C
Pr	Pression à la surface	Pa
TP	Précipitation à la surface	m

La figure 1.2 résume le mécanisme d'acquisition et d'acheminement des données.

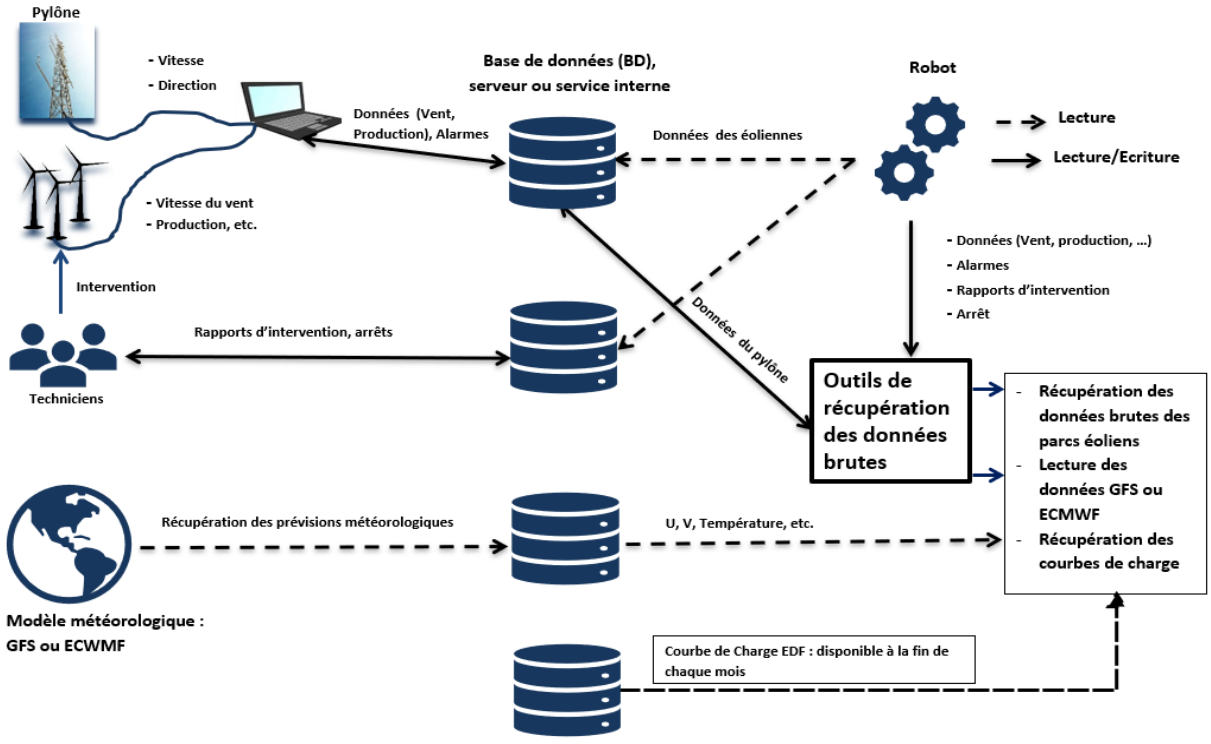


FIGURE 1.2 – Schéma récapitulatif de l'acheminement des données

Dans la suite de ce manuscrit, nous noterons U la composante est-ouest du vent, V la composante nord-sud du vent, Pr la pression et T la température. La vitesse du vent sera notée $W = \sqrt{U^2 + V^2}$.

1.3 Evaluation des prévisions du vent GFS et ECMWF

Les professionnels de l'énergie éolienne reçoivent des informations météorologiques issues de modèles météorologiques différents pour l'estimation ou la prévision de la ressource éolienne. Avant d'aborder la prévision éolienne, une étude comparative des performances de ces deux sources météorologiques en matière de prévision est faite dans cette partie. Cependant le débat sur la performance des deux modèles n'est pas nouveau. Plusieurs études comparatives des deux modèles existent dans la littérature.

1.3.1 Etat de l'art

Plusieurs articles ont abordé les performances et la fiabilité des données de ces deux modèles. Une analyse des performances du modèle GFS comparé à d'autres modèles météorologiques de 1984 à 2014 est faite dans le rapport sur l'examen des prévisions GFS en 2014 [Yang, 2014]. Les prévisions des modèles sont vérifiées par rapport à des analyses, qui sont produites par l'exécution de systèmes d'assimilation de données. La variable utilisée est la HGT (Hauteur géopotentielle ou Geopotential Height) pour une pression de 500hPa. La HGT est précieuse pour localiser des creux et des crêtes qui sont les équivalents de niveau supérieur des cyclones de surface et des anticyclones. Les auteurs ont utilisé de nombreux indicateurs de performance dont le coefficient de corrélation d'anomalie (CCA) qui calcule la corrélation entre les prévisions et les analyses (voir [Persson, 2015] pour plus de détails) pour comparer GFS à d'autres modèles. Leurs résultats montrent une amélioration globalement croissante des performances des deux modèles (GFS et ECMWF) entre 1984 et 2010. Ces performances restent plus ou moins stables à partir de 2010. Sur toute la période le modèle ECMWF a des performances meilleures que GFS. Leurs résultats révèlent également la dégradation des performances des prévisions avec l'horizon de prévision. [Wedam et al., 2009] ont étudié les précisions de quatre modèles dont GFS et ECMWF. Ils ont comparé les prévisions de la pression au niveau de la mer aux observations de pression sur des stations le long des côtes Est et Ouest des Etats Unis avec comme critère d'erreur les différences entre les prévisions de la pression interpolées et la pression observée sur les sites. Ils conclurent entre autres que le modèle ECMWF donne les meilleurs résultats sur toute la période d'étude et que les erreurs sont plus larges sur la côte Ouest que sur la côte Est.

[Kumar et al., 2017] ont analysé l'impact des modèles globaux sur le modèle de prévision méso-échelle WRF (Weather Research and Forecasting). En

effet, ces modèles globaux fournissent les conditions initiales (CI) et les conditions aux limites latérales (CLL) pour la méso-échelle. Ils ont utilisé journalièrement les modèles NCEP (National Centers for Environmental Prediction), GDAS (Global Data Assimilation System), NCMRWF (National Centre For Medium Range Weather Forcasting) et ECMWF pour générer les CI et les CLL pour le modèle WRF (version 3.4). Leurs résultats ont démontré que les prévisions initialisées à partir de ECMWF sont plus proches des observations in situ. Une analyse des quatre cycles (00H, 06H, 12H et 18H) du GFS a montré que les cycles 06H et 18H sont moins performants que les cycles 00H et 12H dans l'hémisphère nord [Fanglin, 2015]. Il n'y a pas de différence significative entre les quatre cycles dans l'hémisphère sud. Une campagne de mesure de la pluviométrie équatoriale au niveau de l'océan indien (novembre et décembre 2011) et du vent zonal a montré que les deux modèles GFS et ECMWF ont de bonnes prévisions à l'horizon 1 à 2 jours et qu'ECMWF a des performances nettement meilleures que GFS à grande échelle à 5-15 jours [Kerns and Chen, 2014].
Nous nous sommes intéressés aux performances des deux modèles en matière de prévision du vent, de température et de la pression.

1.3.2 Indicateurs de performance

Notations

t	instant d'évaluation ou de la prévision d'origine
h	horizon ou pas de temps de prévision
h_{max}	horizon maximum de prévision
N	taille de l'échantillon
Y_t	mesure réelle
$\hat{Y}_{t+h\|t}$	prévision au temps $t+h$ faite au temps d'origine t
$e_{t+h\|t}$	erreur de prévision $e_{t+h\|t} = \hat{Y}_{t+h\|t} - Y_{t+h}$

Plusieurs indicateurs sont utilisés dans la littérature [Gensler et al., 2016] pour évaluer les erreurs de modèles de prévision. Puisqu'on évalue des prévisions déterministes, on s'est restreint à quelques scores d'erreurs dans le cas de prévisions déterministes.

Les erreurs de mesures de base

Pour évaluer la qualité de prévision, un certain nombre d'erreurs de prévision déterministes uniques sont agrégées dans un score global. Le tableau 1.3 résume ces erreurs. Il existe cependant des scores d'erreurs plus sophistiqués qui sont basés sur l'une de ces erreurs. Chaque score peut être calculé séparément pour chaque pas de temps de prévision h, ou sous forme de score global résumé pour tous les pas de temps de prévision. La formule reste la même dans ce cas, bien que, naturellement, tous les points pertinents pour l'évaluation doivent ensuite être inclus.
La moyenne des erreurs aussi appelée biais dans certaines publications sur la prévision éolienne :

$$\bar{e_h} = \frac{1}{N_h} \sum_{i=1}^{N_h} e_i.$$

Cette mesure a la propriété d'équilibrer les erreurs positives et négatives. Par conséquent, elle indique uniquement en moyenne si un modèle surestime ou sous-estime une prévision.
Le mean absolute error (MAE) est défini par :

$$MAE_h = \frac{1}{N_h} \sum_{i=1}^{N_h} |e_i|.$$

Le MAE résume l'erreur absolue de chaque prévision. Elle prend donc en compte les erreurs de manière linéaire. Si la différence minimale globale des valeurs des erreurs doit être déterminée, le MAE est le score approprié.
Le mean squared error (MSE) est défini par :

$$MSE_h = \frac{1}{N_h} \sum_{i=1}^{N_h} e_i^2.$$

Contrairement au score MAE, ce score prend en compte les erreurs de manière quadratique. Ainsi, les erreurs élevées sont davantage pénalisées, alors que les erreurs faibles ont une influence moindre sur le score global. Si un modèle de prévision doit éviter des erreurs extrêmes, le score MSE est la mesure d'erreur la plus appropriée. Cependant, le score MSE est un score au carré, la valeur a peu de relations avec les différences réelles. Par conséquent, ce score est principalement utilisé à des fins d'optimisation lors de la formation du modèle de prévision ([Gensler et al., 2016]).
Le root mean squared error (RMSE) est défini par :

$$RMSE_h = \sqrt{\frac{1}{N_h} \sum_{i=1}^{N_h} e_i^2}.$$

Le RMSE a la même signification qualitative que le MSE. Cependant, lorsque la racine carrée de la valeur MSE est calculée, la valeur est représentée dans l'unité physique d'origine, ce qui facilite la mise en relation avec une valeur prévue.

TABLE 1.3 – Résumé des mesures de base d'erreurs déterministes

Mesure d'erreur	Formule	Objectif
$\bar{e_h}$	$\bar{e_h} = \frac{1}{N_h} \sum_{i=1}^{N_h} e_i$	Indique si un algorithme surestime ou sous-estime
MAE	$MAE_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \lvert e_i \rvert$	Mesure d'erreur absolue linéaire. Pondération proportionnelle des erreurs
MSE	$MSE_h = \frac{1}{N_h} \sum_{i=1}^{N_h} e_i^2$	Erreur quadratique. Plus grande pondération des grosses erreurs
RMSE	$RMSE_h = \sqrt{\frac{1}{N_h} \sum_{i=1}^{N_h} e_i^2}$	La racine carrée du MSE a l'unité physique d'origine de la prévision

Les techniques de normalisation

Il existe une multitude de types de normalisation des erreurs, chacun ayant un objectif précis.

1 La méthode de normalisation la plus simple consiste à diviser par la valeur maximale mesurée Y_{max} (ou la puissance installée du parc lorsqu'il s'agit de prévisions de production éolienne). En utilisant cette forme de normalisation, une comparaison sans échelle de la qualité de la prévision pour différents sites est possible.

2 Une autre façon de normaliser consiste à diviser l'erreur par la mesure réelle Y_t (si les mesures Y_t ne prennent pas des valeurs nulles). Cette forme de normalisation réalise une erreur relative au sens d'une erreur en pourcentage. Elle donne plus de poids dans un scénario de mesure réelle faible que dans un scénario de mesure réelle élevée.

3 L'erreur peut être normalisée en prenant en compte l'écart absolu entre la mesure réelle Y_t et la moyenne des mesures $\bar{Y}$: $\frac{e_{t+h|t}}{|Y_t - \bar{Y}|}$.
Cette forme de normalisation pénalise les erreurs proches de la moyenne

des mesures, tandis que les erreurs aux extrémités ont moins d'influence sur le score d'erreur global.

4 L'erreur peut être normalisée par rapport aux caractéristiques dynamiques de la mesure actuelle, l'erreur normalisée est par conséquent donnée par : $\frac{e_{t+h|t}}{|Y_t - Y_{t-1}|}$.
En général, un problème de prévision est plus difficile lorsque la dynamique de la situation météorologique (et donc de la série chronologique) est élevée. Cette forme de normalisation vise à pénaliser les erreurs dans les situations à variabilité dynamique faible plus élevée, tandis que les situations avec une forte variabilité sont pondérées plus bas. Ce mode de normalisation réduit l'impact des situations météorologiques difficiles.

Les erreurs dérivées des erreurs de base

Il existe un certain nombre de scores d'erreur qui combinent l'un des scores d'erreur de base (voir section 1.3.2) et une technique de normalisation (voir section 1.3.2). Globalement, il en existe deux (voir tableau 1.4) et qui permettent entre autres de comparer les erreurs de prévision d'un modèle sur plusieurs parcs éoliens dans les mêmes proportions. La liste des erreurs qui dérivent de ces combinaisons n'est pas exhaustive (voir [Gensler et al., 2016] pour plus de combinaisons).

TABLE 1.4 – Erreurs dérivées des erreurs de bases et une technique de normalisation.

Mesure d'erreur	Formule
NMAE	$NMAE_h = \frac{MAE_h}{Y_{max}}$
NRMSE	$NRMSE_h = \frac{RMSE_h}{Y_{max}}$

Évaluation des écarts, corrélation et comparaison des modèles

La mesure standard pour l'évaluation de la déviation est la déviation standard. Cependant, comme il est déjà traité avec des erreurs, le terme écart type d'erreurs (SDE : Standard Deviation of Errors) est introduit dans [Madsen et al., 2005] qui est néanmoins très similaire au calcul classique de

l'écart type :

$$SDE_h = \sqrt{\frac{\sum_1^{N_h}(e_i - \bar{e})^2}{N_h - (q+1)}},$$

où q est le nombre de paramètres estimés et $\bar{e}$ la moyenne des erreurs. Par conséquent $q = 0$ pour les données de test. Le critère SDE est une estimation de l'écart type de la distribution d'erreurs et seule l'erreur aléatoire contribue au critère SDE. Une autre technique d'évaluation de la qualité des prévisions est le coefficient de corrélation.
Nous allons utiliser comme indicateurs statistiques la moyenne des erreurs, le MAE, le RMSE et le SED.

1.3.3 Performances GFS et ECMWF

Pour évaluer la qualité des prévisions des modèles GFS et ECMWF, on considère les quatre points d'intersection des maillages des deux modèles les plus proches du pylône de mesure (voir figure 1.3).

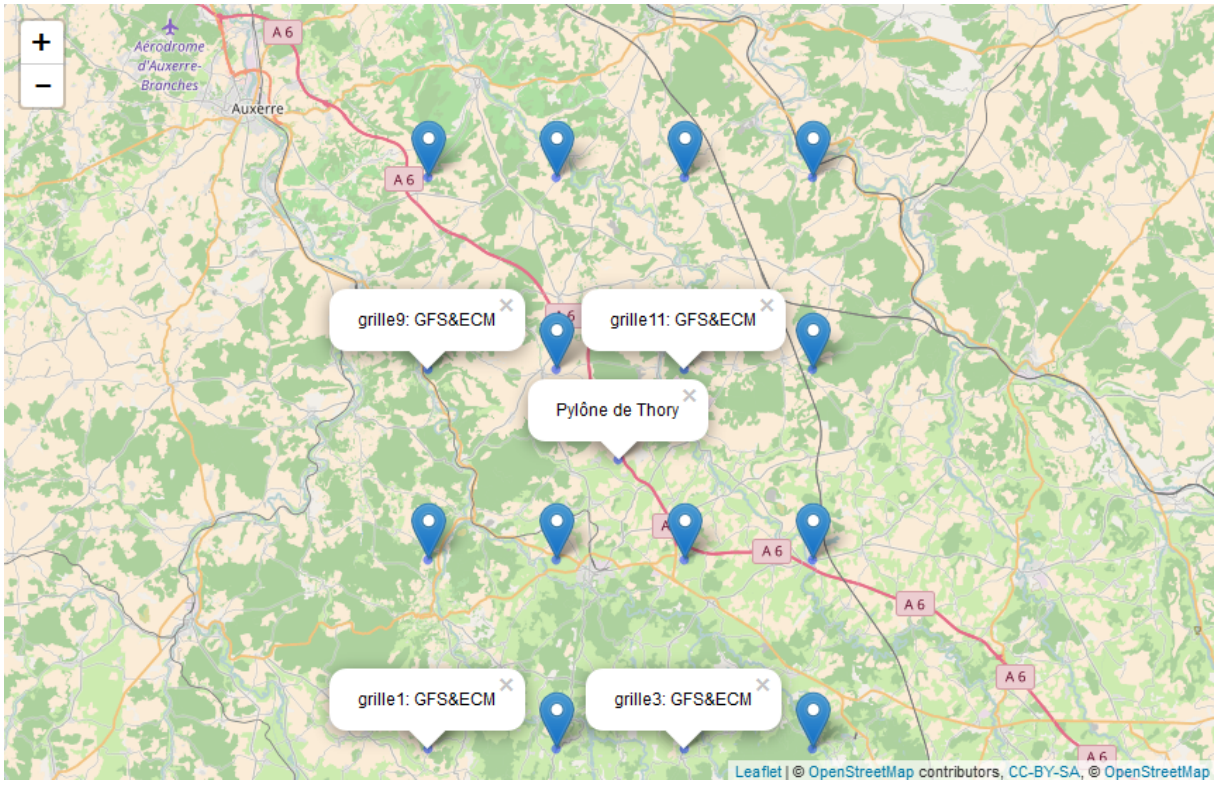


FIGURE 1.3 – Position du pylône et des quatre points d'intersection des deux maillages GFS et ECMWF les plus proches du pylône

On s'intéresse aux variables de vent (U, V et W), de température T et de pression Pr. D'un coté nous avons les mesures réalisées au niveau du mât et

de l'autre les prévisions issues de GFS ou ECMWF. On considère les prévisions météorologiques (GFS et ECMWF) du vent à 100m de hauteur qu'on compare avec les mesures réelles du vent à la même hauteur. Pour la température, les prévisions GFS et ECMWF sont à 2m de hauteur et les mesures ont été obtenues à 5m de hauteur. Même s'il y a une différence de 3m, la température reste relativement stable à cette hauteur. Nous utilisons un historique de prévisions de GFS 0.25 horaire et ECMWF aux mêmes points de grille (les quatre points de grille dans la figure 1.3). Ces prévisions sont issues du run de 00h de chaque modèle avec un horizon de prévision de 47h. Pour le run de 00h, les prévisions commencent à 00h+3h jusqu'à 00h+47h. Nous avons donc pour chaque jour de 2017, pour chaque variable, 45 valeurs horaires.

Qualité des prévisions de la vitesse du vent

Nous commençons par une étude globale en analysant toutes les erreurs obtenues entre la vitesse mesurée au pylône et les prévisions.

TABLE 1.5 – Résumé des mesures d'erreurs

Vitesse du vent	$\bar{e}$	MAE	RMSE	SED
GFS grille 1	0.06	1.41	1.88	1.88
ECMWF grille 1	-0.60	1.38	1.79	1.69
GFS grille 2	0.13	1.42	1.90	1.89
ECMWF grille 2	-0.48	1.31	1.71	1.64
GFS grille 3	0.08	1.38	1.83	1.83
ECMWF grille 3	-0.18	1.27	1.66	1.65
GFS grille 4	0.12	1.38	1.84	1.83
ECMWF grille 4	-0.19	1.23	1.62	1.61

On note dans le tableau 1.5 que le modèle ECMWF a tendance à avoir une moyenne des écarts qui est négative contrairement au modèle GFS. Ce qui montre que globalement le modèle ECMWF tend à sous-estimer le vent à $100m$. En analysant les résultats des prévisions des deux modèles on note que sur les quatre points de grille les erreurs de prévision sont uniformes. Les prévisions météorologiques du vent d'un point de grille à un autre sont fortement corrélées avec pratiquement les mêmes proportions d'erreurs de prévision. En

observant le MAE et le $RMSE$ qui sont plus significatifs pour comparer les prévisions, on note que le modèle ECMWF est légèrement meilleur en matière de prévision du vent. Les erreurs de prévision du modèle ECMWF sont moins dispersées que celles du modèle GFS.

Nous avons affiné notre analyse en évaluant la performance de chaque modèle à chaque horizon de prévision. On s'est restreint à l'utilisation du MAE comme indicateur statistique puisque nous avons abouti pratiquement aux mêmes conclusions que le RMSE dans la section précédente. On a calculé le MAE de chaque modèle à chaque horizon de prévision. Pour simplifier, nous avons affiché les erreurs par horizon de prévision au point de grille le plus proche du site (figure 1.4). Nous avons observé la même tendance sur les trois autres points de grille.

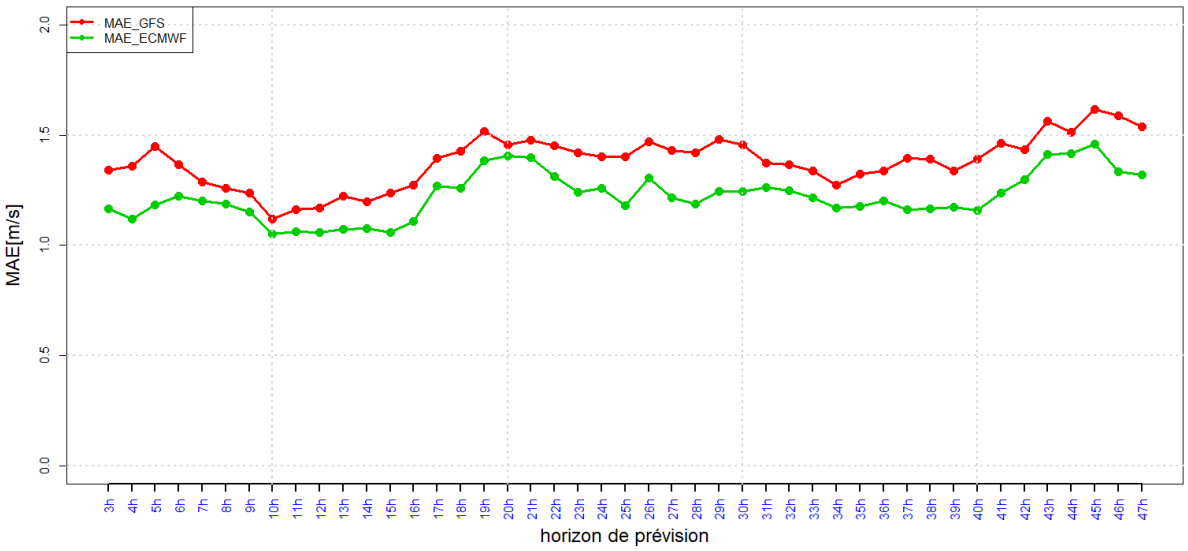


FIGURE 1.4 – MAE par horizon de prévision de vitesse du vent GFS (en rouge) et ECMWF (en vert) au point de grille le plus proche.

Le résultat obtenu montre que le modèle ECMWF fait moins d'erreurs que le modèle GFS pratiquement à chaque horizon. Ce qui vient confirmer la conclusion de l'analyse globale faite dans la section 1.3.3. On remarque aussi une légère augmentation du MAE avec l'horizon des prévisions : plus l'horizon de prévision est lointain, plus les erreurs de prévision augmentent. Les erreurs de prévision ont aussi tendance à augmenter entre 16h et minuit.

1.3.4 Comparaison des distributions bivariées (U, V)

On considère les composantes est-ouest du vent U et nord-sud du vent V de GFS et de ECMWF au point de grille le plus proche et celles du vent du pylône. Les prévisions de chaque modèle sont dépendantes d'une heure à l'autre, nous avons estimé les distributions par horizon de prévision (ce qui correspond à 45 analyses). A un horizon h donné, nous avons les 365 données quotidiennes de U et V. Nous commençons par une estimation à noyau des densités bivariées [Sim, 1996] de (U, V). L'estimation à noyau de la densité bivariée est une approche non paramétrique pour estimer la densité d'une variable aléatoire. Elle a déjà été utilisée par [Zhang et al., 2013] pour estimer la densité bivariée des vitesses et des directions du vent. De manière générale, pour un échantillon multivarié $X_1, X_2, \ldots, X_n$ de dimension d, issu d'une densité f inconnue, l'estimateur à noyau de la densité multivariée, est défini par :

$$\hat{f}(\mathbf{x}, \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i),$$

où, $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$ et $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T$, $i = 1, 2, \ldots, n$, $K(\mathbf{x})$ est un noyau qui est une fonction de densité de probabilité symétrique, $\mathbf{H}$ est une matrice de "fenêtre" (bandwidth en anglais) qui est symétrique, définie positive et $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K\left(\mathbf{H}^{-1/2}\mathbf{x}\right)$, ici $\mathbf{X}_i = (U_i, V_i)$ sont les données de tous les jours de l'année pour un horizon donné. L'horizon varie de 3h à 47h. Nous avons pris un noyau Gaussien, $K(\mathbf{x}) = (2\pi)^{-d/2} exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})$. L'avantage de l'estimation à noyau de la distribution bivariée est qu'elle permet de représenter des données de vent multimodales, ce qui est rare dans la littérature sur la distribution du vent ([Zhang et al., 2013]). Les figures 1.5(a) et 1.6(a) représentent la distribution bivariée du pylône de mesure à 13h. La distribution bivariée du vent GFS au point de grille le plus proche à 13h est représentée par les figures 1.5(b) et 1.6(b) et celle ECMWF (au même point de grille) à 13h par les figures 1.5(c) et 1.6(c). Pour simplifier nous n'avons pas intégré les images des distributions bivariées à tous les horizon (il n'y a pas assez de différences visuelles). Il est intéressant de noter que la distribution de probabilité estimée est multimodale. D'ailleurs, d'après [Zhang et al., 2013] la distribution de probabilité estimée des données de vent terrestre est de nature multimodale. La figure 1.6 montre que le vent est plus dense autour de $U \approx 5\,m/s$ et $V \approx 1\,m/s$: on retrouve une information connue sur le site où on sait que le vent dominant est sud-ouest. On note aussi que la distribution bivariée ECMWF semble être plus "proche" de celle du pylône.

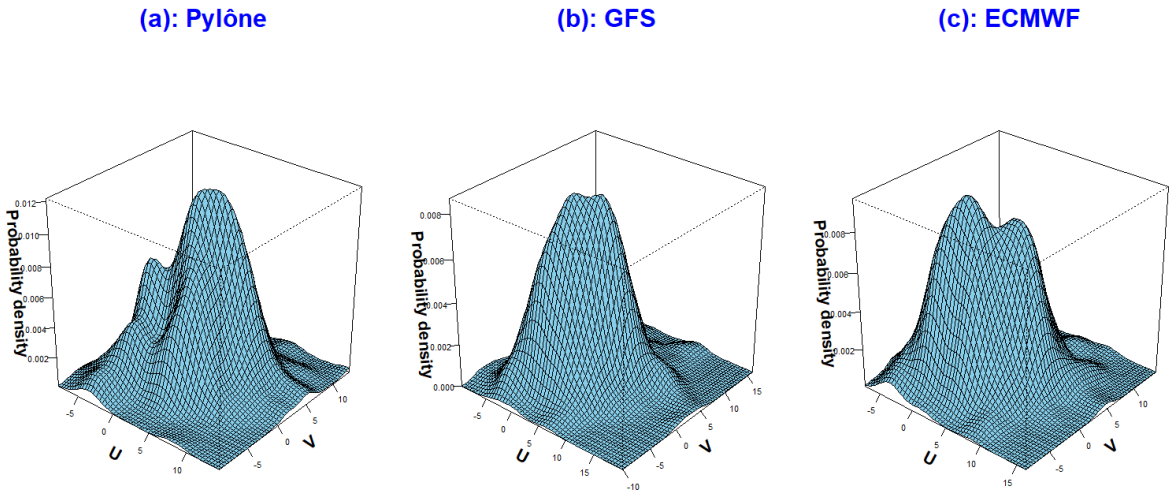


FIGURE 1.5 – Densités bivariées du vent à 13h

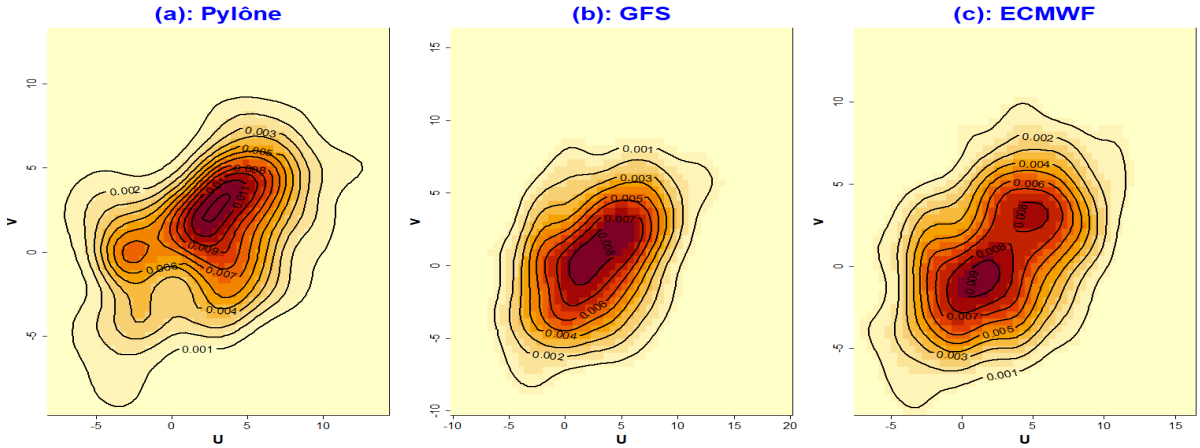


FIGURE 1.6 – Distributions bivariées du vent à 13h

Pour comparer les distributions bivariées du vent GFS et ECMWF à celle du pylône de mesure on a utilisé la divergence de Kullback-Leibler ([Kullback and Leibler, 1951]). Elle permet de mesurer l'écart entre deux densités de fonction p et q. La divergence de Kullback-Leibler (KL) est définie dans le cas continu par :

$$D_{KL}(p,q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx,$$

et dans le cas discret par :

$$D_{KL}(p,q) = \sum_i p(i) \log \frac{p(i)}{q(i)}.$$

Elle n'est pas une distance au sens topologique du terme. En effet, l'inégalité triangulaire n'est pas vérifiée ainsi que la propriété de symétrie. La divergence de Kullback-Leibler est positive ou nulle et plus elle est proche de zéro, plus q est "proche" de p.

Une fois les densités bivariées (U, V) de GFS, d'ECMWF et du pylône estimées, on a calculé la divergence de Kullback-Leibler entre les distributions bivariées pylône et GFS puis la divergence de Kullback-Leibler entre pylône et ECMWF pour tous les horizons de prévision de 3h à 47h. D'après la figure 1.7, il y a certains horizons de prévision où la distribution bivariée du vent GFS est plus "proche" de celle du pylône mais globalement, la distribution bivariée du vent ECMWF semble être plus "proche" de celle du pylône de mesure.

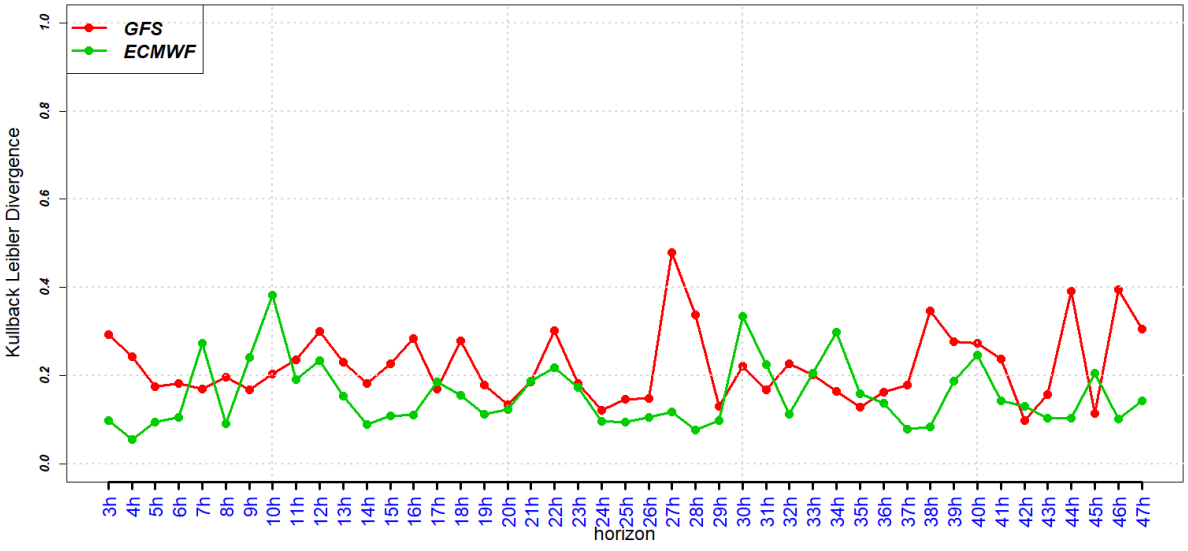


FIGURE 1.7 – Divergence de Kullback-Leibler par horizon de prévision pour GFS et ECMWF par rapport au pylône

Même si du point de vue de la comparaison avec la réalité, le modèle ECMWF produit moins d'erreurs de prévision des vitesses du vent, nous nous sommes intéressés dans la section suivante aux ressemblances entre les deux groupes de variables des deux modèles GFS et ECMWF par rapport aux variables du pylône de mesure. L'analyse de corrélation canonique permet d'étudier les relations entre deux tableaux de données.

1.4 Analyse de corrélation canonique

L'analyse de corrélation canonique (ACC) est une méthode proposée par [Hotelling, 1936] pour étudier les relations pouvant exister entre deux groupes de variables (deux tableaux de données). Elle a pour intérêt de mesurer et caractériser les corrélations entre deux groupes de variables mesurées sur les mêmes individus. L'analyse de corrélation canonique a typiquement deux objectifs :

- réduction de données : expliquer la covariance entre deux ensembles de variables en utilisant un petit nombre de combinaisons linéaires
- interprétation des données : trouver les caractéristiques (c'est-à-dire les variables canoniques) qui sont importantes pour expliquer la covariance entre des ensembles de variables.

1.4.1 Description et interprétation de la méthode

On considère deux groupes de variables $\mathbf{X}_1$ et $\mathbf{X}_2$ mesurés sur les $n > 1$ mêmes individus et avec respectivement $p > 1$ et $q > 1$ variables, tels que les matrices de variances de chaque groupe existent. Notons $\Sigma_{\mathbf{X}_1}$ et $\Sigma_{\mathbf{X}_2}$ les matrices de covariance respectives de $\mathbf{X}_1$ et $\mathbf{X}_2$ puis $\mu_{\mathbf{X}_1}$ et $\mu_{\mathbf{X}_2}$ les moyennes de $\mathbf{X}_1$ et $\mathbf{X}_2$ respectivement. Soit $X = (\mathbf{X}_1, \mathbf{X}_2)$ de moyenne $\mu_X = (\mu_{\mathbf{X}_1}, \mu_{\mathbf{X}_2})$ et de matrice de variance

$$\Sigma_X = \begin{pmatrix} \Sigma_{\mathbf{X}_1} & \Sigma_{\mathbf{X}_1\mathbf{X}_2} \\ \\ \Sigma_{\mathbf{X}_2\mathbf{X}_1} & \Sigma_{\mathbf{X}_2} \end{pmatrix}.$$

L'analyse de corrélation canonique consiste à trouver deux variables que nous notons $\Phi = \mathbf{X}_1 a$ et $\Psi = \mathbf{X}_2 b$ (qui sont des combinaisons linéaires des variables de $\mathbf{X}_1$ et de $\mathbf{X}_2$ respectivement) telles que la corrélation entre eux soit maximale, où $a \in \mathbb{R}^p$ et $b \in \mathbb{R}^q$. La formulation mathématique consiste à trouver les deux combinaisons linéaires telles que :

$$\{a, b\} = arg.max_{a,b}\{Cor(\mathbf{X}_1 a, \mathbf{X}_2 b)\} \tag{1.1}$$

sous les contraintes que

$$Var(\mathbf{X}_1 a) = Var(\mathbf{X}_2 b) = 1. \tag{1.2}$$

Ces contraintes permettent d'assurer l'unicité des vecteurs a et b. Après avoir déterminé la première corrélation ρ_1 (appelée première corrélation canonique) entre Φ_1 et Ψ_1 (les premières variables canoniques), on peut trouver les autres corrélations canoniques par récurrence en maximisant :

$$\rho_i = Cor(\Phi_i, \Psi_i)$$

sous les contraintes $Var(\Phi_i) = Var(\Psi_i) = 1$ et des contraintes additionnelles $Cor(\Phi_s, \Phi_j) = Cor(\Psi_s, \Psi_j) = 0$ pour tout $j \in \{1, 2, \ldots, (s-1)\}$. On définit ainsi une suite de s couples de variables canoniques et une suite de corrélations canoniques décroissantes $\rho_1 \geq \rho_2 \geq, \ldots, \geq \rho_s$ avec $s = \min(p, q)$.
Sous la contrainte (1.2), les coefficients (appelés aussi facteurs) a et b solution du problème vérifient les deux équations suivantes :

$$\Sigma_{\mathbf{X}_1}^{-1}\Sigma_{\mathbf{X}_1\mathbf{X}_2}\Sigma_{\mathbf{X}_2}^{-1}\Sigma_{\mathbf{X}_2\mathbf{X}_1}a = \rho^2 a \textit{ et } \Sigma_{\mathbf{X}_2}^{-1}\Sigma_{\mathbf{X}_2\mathbf{X}_1}\Sigma_{\mathbf{X}_1}^{-1}\Sigma_{\mathbf{X}_1\mathbf{X}_2}b = \rho^2 b,$$

C'est-à-dire que a (respectivement b) est un vecteur propre normé de $\Sigma_{\mathbf{X}_1}^{-1}\Sigma_{\mathbf{X}_1\mathbf{X}_2}\Sigma_{\mathbf{X}_2}^{-1}\Sigma_{\mathbf{X}_2\mathbf{X}_1}$ (respectivement de $\Sigma_{\mathbf{X}_2}^{-1}\Sigma_{\mathbf{X}_2\mathbf{X}_1}\Sigma_{\mathbf{X}_1}^{-1}\Sigma_{\mathbf{X}_1\mathbf{X}_2}$) associé à la plus grande valeur propre commune aux deux matrices et qui est ρ^2.
Une autre manière consiste à aborder l'analyse de corrélation canonique sous forme de projeté orthogonal. En effet si $\mathbf{X}_1$ et $\mathbf{X}_2$ sont centrés (sinon on les centre), l'interprétation géométrique du problème 1.1 consiste à rechercher des directions de $E_{\mathbf{X}_1}$ et de $E_{\mathbf{X}_2}$ les plus proches possibles (d'angle minimal), c'est-à-dire telles que $\mathbf{X}_1 a \approx \mathbf{X}_2 b$, où $E_{\mathbf{X}_1}$ est l'espace formé par les variables de $\mathbf{X}_1$ et $E_{\mathbf{X}_2}$ l'espace formé par les variables de $\mathbf{X}_2$. Soient $P_{\mathbf{X}_1} = \mathbf{X}_1({}^t\mathbf{X}_1\mathbf{X}_1)^{-1}\,{}^t\mathbf{X}_1$ et $P_{\mathbf{X}_2} = \mathbf{X}_2({}^t\mathbf{X}_2\mathbf{X}_2)^{-1}\,{}^t\mathbf{X}_2$ les projetés orthogonaux sur $E_{\mathbf{X}_1}$ et $E_{\mathbf{X}_2}$ respectivement. Φ et Ψ sont respectivement aussi les vecteurs propres associés $P_{X_1}P_{X_2}$ et $P_{X_2}P_{X_1}$:

$$P_{X_1}P_{X_2}\Phi = \rho^2\Phi \textit{ et } P_{X_2}P_{X_1}\Psi = \rho^2\Psi$$

L'interprétation des résultats (corrélations canoniques et variables canoniques) en analyse canonique n'est pas aussi simple que dans les autres méthodes d'analyse des données. Pour ce qui est du choix des composantes, une méthode simple consiste à couper à une chute des corrélations tandis que des tests statistiques permettent aussi de choisir le nombre de corrélations canoniques synonymes du nombre de dimensions à garder. Si les coefficients de corrélation canonique sont ρ_k, $(k = 1, \ldots, s)$, le test de l'hypothèse classique pour savoir quelle est la dimension de représentation :

$$H_0^k : \rho_1 \neq, \ldots, \rho_k \neq 0, \rho_{k+1} = \ldots = \rho_s = 0$$

utilise la statistique :

$$-\{(n-1) - k - (p+q=1)/2 + \prod_{i=1}^{k}\rho_i^{-2}\}Ln\left(\prod_{i=k+1}^{s}(1-\rho^2)\right) \;{}_{(}H_0^k)\chi^2_{(p-k)(q_k)}.$$

La procédure consiste donc à tester séquentiellement les hypothèses $H_0^1, \ldots, H_0^k, \ldots$; on décide de la dimension dès qu'une hypothèse H_0^k ne peut

pas être rejetée ([Bellanger et al., 2006]). Nous ne sommes pas dans une optique de réduction de la dimension c'est pourquoi nous n'avons pas abordé cette partie dans la suite. Nous nous sommes plus focalisés sur l'analyse des relations entre les deux tableaux de données. Nous avons utilisé la somme des valeurs propres pour interpréter les résultats : plus la somme des valeurs propres est élevée, plus les corrélations entre les deux tableaux de données sont fortes.

1.4.2 Comparaison GFS et ECMWF par l'ACC

Dans cette partie, nous avons cherché à comparer non pas individuellement mais de manière globale, un ensemble de variables des modèles GFS et ECMWF par rapport à leurs mesures réelles disponibles au niveau du pylône. Nous avons pour chaque heure les mesures du pylône $\mathbf{X}_1 = (U, V, Pr, T)$ et nous avons aussi les prévisions des modèles météorologiques que nous noterons $\mathbf{X}_2^{ECMWF}$ et $\mathbf{X}_2^{GFS}$. Nous avons appliqué l'ACC entre $\mathbf{X}_1$ et $\mathbf{X}_2^{ECMWF}$ et entre $\mathbf{X}_1$ et $\mathbf{X}_2^{GFS}$. Ensuite nous avons calculé la sommes des valeurs propres découlant de l'ACC pour chaque horizon (figure 1.8).

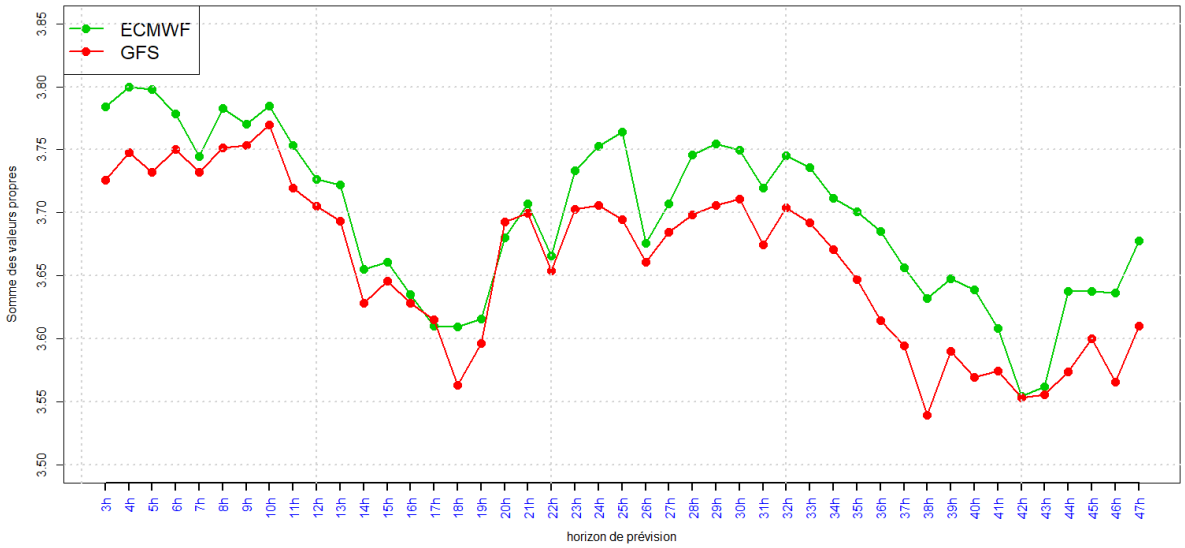


FIGURE 1.8 – Somme des valeurs propres par horizon de prévision pour GFS et ECMWF par rapport au pylône.

Si la somme des valeurs propres entre le groupe de variables du pylône et le groupe de variables GFS ou la sommes des valeurs propres entre le groupes des variables du pylône et celui de ECMWF est égale à 4 (dimension de chaque

tableau), le modèle GFS ou le modèle ECMWF est en corrélation parfaite avec le pylône.

Les résultats (figure 1.8) montrent que les prévisions GFS X_2^{GFS} et ECMWF X_2^{ECMWF} sont plus corrélées aux mesures du pylône X_1 dans la nuit que dans le journée. En effet, le vent est plus stable dans la nuit (aux environs de 22h à 08h) et plus instables dans la journée (de 9h à 22h environ). Les prévisions ECMWF sont légèrement plus corrélées aux mesures réelles du pylône que les prévisions GFS. Ce qui confirme les résultats obtenus par la comparaison individuelle des variables des deux modèles par rapport aux mesures réelles du pylône dans la section 1.3.

1.5 Conclusion

Dans cette partie nous avons évalué les performances des modèles météorologiques GFS et ECMWF. L'évaluation des performances des prévisions des modèles GFS et ECMWF a montré que le modèle ECMWF offre des prévisions du vent légèrement meilleures que le modèle GFS. L'analyse a également montré que les erreurs de prévision des modèles météorologiques augmentent avec l'horizon de prévision. En dernière partie de ce chapitre l'analyse canonique a permis de montrer que les variables météorologiques du modèle ECMWF sont plus corrélées aux variables du pylône.
Dans le chapitre suivant nous avons utilisé les prévisions du modèle ECMWF comme input pour la prévision court terme de la production des éoliennes.

Chapitre 2

prévision court terme de la production par Machine Learning

Sommaire

2.1 Introduction

L'énergie éolienne va devoir s'adapter à un nouveau mode de valorisation. Pour les producteurs, il faudra nécessairement anticiper la vente d'électricité sur le marché. Il existe donc un besoin énorme de prévisions précises. Dans cette étude nous proposons quelques modèles statistiques de prévisions du productible de parcs éoliens. La prévision de la production des parcs éoliens n'est cependant pas nouvelle ([Juban et al., 2008], [Tastu et al., 2011],

[Kusiak et al., 2008], [Fugon et al., 2008] et [Costa et al., 2008]). Elle est établie depuis des décennies au moyen de modèles et de techniques différents. Ce chapitre est basé sur notre publication [Dione and Matzner-Løber, 2019]. Nous nous concentrerons sur des algorithmes d'apprentissage automatique pour prévoir la production de parcs éoliens à l'aide de données météorologiques. À court terme (de 24 à 47 heures), il existe deux approches différentes dans la littérature : l'approche directe qui consiste à prévoir directement la production à partir des entrées (principalement des variables météorologiques) et l'approche indirecte qui prévoit d'abord le vent sur le site à partir des données météorologiques puis le transforme en production. C'est une méthode en une étape contre une en deux étapes qui peuvent être résumées dans les images 2.1 et 2.2.

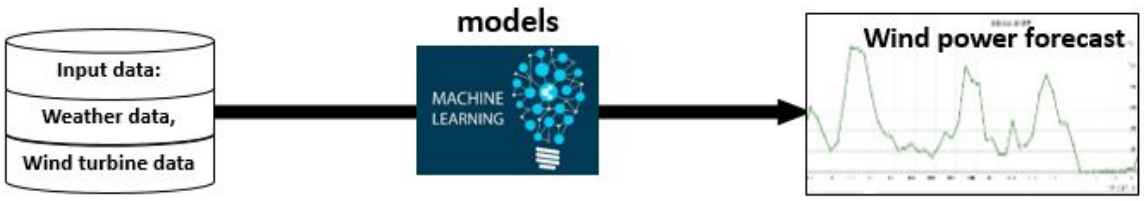


FIGURE 2.1 – Approche directe de la prévision de la production éolienne

FIGURE 2.2 – Approche indirecte de la prévision de la production éolienne

La première citée est la plus simple par ce qu'elle est constituée d'un seul modèle basé sur un apprentissage direct. Cependant elle n'exploite pas les mesures réelles des vitesses du vent sur les sites. L'approche consiste à prédire la production en se basant uniquement sur les prévisions météorologiques. L'approche indirecte de son coté exploite les mesures réelles du vent qui apporte une information en plus et permet d'améliorer les prévisions météorologiques du vent avant de les convertir en prévisions de production. Cependant elle a

des limites notamment si les mesures du vent par les anémomètres sont très variables (avec beaucoup d'incertitudes par example en cas de dysfonctionnement ou de présence de neige ou de gel sur les anémomètres). Dans ce cas la courbe de puissance réelle qui résulte des mesures réelles du vent et de la production réelle est très incertaine et moins lisse. Par conséquent la deuxième phase de la traduction du vent prédit en prévision de production apporte plus d'incertitudes qu'elle n'améliore les prévisions de la production. Dans ce cas l'approche indirecte peut être moins avantageuse.

Dans la section 2.2, l'état de l'art sur la prévision court terme dans le domaine de l'éolien est fait. La section 2.3 donne les détails sur les données utilisées et la section 2.4 résume la méthode de modélisation et les principes des modèles appliqués dans notre étude. Les résultats sont discutés dans la section 2.5 et enfin une conclusion est dégagée dans la section 2.6.

2.2 Etat de l'art

Le développement de l'énergie éolienne ces dernières années a entraîné un challenge dans la gestion du marché de l'électricité à cause du caractère très variable du vent. Pour certains professionnels, une telle variabilité peut augmenter les coûts globaux de l'énergie produite et limiter ainsi les avantages de l'utilisation d'une telle ressource énergétique [Juban et al., 2008]. En effet si la production livrée sur le marché est sur-estimée ou sous-estimée par rapport à la demande, les industriels sont exposés à des pénalités ou des gains sur la vente d'électricité. Les outils de prévision permettent de réduire l'incertitude sur la vente d'électricité en fournissant des prévisions fiables (avec le moins d'écart possible par rapport à la production réelle qui sera livrée). Ainsi on retrouve dans la littérature deux types de modélisation pour la prévision du productible éolien : la modélisation déterministe et la modélisation probabiliste [Kariniotakis and Giebel, 2017, Giebel and Kariniotakis, 2009]. Dans la modélisation déterministe, on distingue les modèles physiques et les modèles statistiques [Baïle, 2010]. Les modèles physiques s'appuient sur la modélisation de chaque éolienne sur la base des équations [Costa et al., 2008] de courbe de puissance :

$$P_w = 0.5\rho v^3,$$

où P_w la densité de la puissance, ρ la densité de l'air et v la composante horizontale de la vitesse du vent.

D'autres techniques statistiques s'appuient sur un apprentissage direct à partir des données. On y retrouve les modèles statistiques et les mé-

thodes de machine learning notamment utilisées dans la prévision court terme [Fugon et al., 2008]. Entre autres, les modèles de régression paramétrique, les support vector machine (SVM) pour la régression, les arbres de régression (CART : Classification And Regression Tree) et les forêts aléatoires sont souvent utilisés. Les sources d'entrées des modèles de prévisions éoliennes peuvent varier en fonction des horizons de prévision. Selon que l'horizon de prévision soit de quelques minutes, quelques heures ou quelques jours, certains suggèrent des modèles basés sur la production temps réel, des modèles basés sur l'imagerie satellite et au sol ou des modèles basés sur les prévisions météorologiques [Najac, 2012]. Il existe aussi d'autres modèles physiques en plus des modèles statistiques pour la descente d'échelle du vent.

L'article [Kusiak et al., 2008] aborde la prévision court terme de la production éolienne par une approche de data mining en considérant les données du modèle RUC (Rapid Update Cycle) et du modèle NAM (North American Mesoscale) sur seize points de grille les plus proches d'un parc. D'abord, les auteurs font une sélection du nombre de points de grille à considérer par la méthode du boosting tree algorithm, ensuite ils appliquent une ACP pour réduire la dimension des variables et enfin ils appliquent des modèles de SVM-reg, multilayer perceptron network (MLP), radial basis function (RBF), arbres de régression et de forêts aléatoires. Le modèle MLP a donné les meilleurs résultats avec un MAE entre 9.8 et 11.46% en utilisant comme entrée les données RUC (horizon 12h) et un MAE entre 5.93 et 10.57% pour les données d'entrée NAM. La prévision du vent par MLP puis de la production par K-NN ne leurs a pas donnés une amélioration des prévisions.

Une étude prédictive de la production éolienne de trois parcs de complexités différentes en France avec un modèle linéaire, la persistance (modèle de référence) et des modèles non linéaires (Réseaux de neurones, forêts aléatoires, et SVM) est faite dans [Fugon et al., 2008]. Les auteurs ont utilisé comme entrées des données du modèle ARPEGE de météo France, sur une période de 18 mois avec un horizon de prévision de 60h. La performance de leurs modèles a été évaluée par le Mean absolute Error (MAE) :

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}_i - Y_i|$$

et le Root Mean Square Error (RMSE) :

$$RMSE(\hat{Y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}.$$

Pour obtenir le taux d'erreur en terme de puissance installée, la Normalized Mean Absolute Error (NMAE) :

$$NMAE = 100 * \frac{MAE}{PI}\%,$$

ou le Normalized Root Mean Square Error (NRMSE) :

$$NMAE = 100 * \frac{MAE}{PI}\%$$

sont aussi utilisés comme indicateurs statistiques. Les forêts aléatoires ont donné les meilleurs résultats avec un $NRMSE$ entre 9 et 15% environ pour les horizons de prévision allant de 3h à 60h. Une étude de différentes approches probabilistes est faite dans [Juban et al., 2008], suivie d'une évaluation des performances déterministes (moyenne des distributions) avec comme critère d'évaluation le NMAE. Leurs résultats montrent un NMAE entre 9 et 15% environ pour un horizon de prévision de $60h$ avec une augmentation globale en fonction des horizons de prévision. La prévision de la densité de la production éolienne en utilisant les prévisions d'ensemble du modèle ECMWF est abordée dans [Taylor et al., 2009]. Dans cet article, les auteurs estiment la moyenne et la variance du carré de la vitesse par des modèles AR-GARCH ou ARFI-GARCH, puis après simulation de 10000 vitesses du vent au carré suivant une loi gaussienne, ils appliquent une courbe de puissance théorique pour obtenir la production en prenant la moyenne de la conversion des 10000 vitesses. Le MAE et le logarithme du maximum de vraisemblance sont utilisés pour évaluer respectivement les prévisions de la production et les prévisions des distributions de probabilités. Ils obtiennent sur cinq parcs (50.25 MW), un MAE moyen entre 50 et 150kW du premier au dixième jour. Il est important de souligner que dans leur étude, l'aspect spatial n'a pas été pris en compte.

Nous utilisons l'approche statistique dans la suite avec des méthodes de machine learning. Le NMAE est utilisé comme indicateur de performance pour se positionner par rapport à l'état de l'art. La description des données utilisées est faite dans la section suivante.

2.3 Les données

Les données utilisées sont décrites dans la section 1.2 du chapitre 1. Nous avons fait un modèle de prévision pour chaque parc éolien. Pour cela nous avons besoin d'une méthode de calcul de la production horaire du parc. La production totale du parc à un instant t découle d'une part, de la moyenne

des 60 relevées de la production par minute précédant t pour chaque éolienne (pour passer de la fréquence minute en fréquence horaire) et d'autre part, de la somme des productions horaires des éoliennes. Nous avons vu dans le chapitre 1 qu'à une mesure de production correspondait un indice de fonctionnement indiquant si l'éolienne était en fonctionnement normale ou pas. Ainsi, nous avons défini deux types de production horaire du parc.

- Type de production 1 : on ne tient compte que des périodes où il n'y a eu aucun dysfonctionnement ou arrêt d'aucune éolienne durant les 60 minutes précédant l'instant t. Dans ce cas la production horaire d'une éolienne E que nous avons notée $ProdH_1^E$ à l'instant t est donnée par :

$$ProdH_1^E(t) = \frac{1}{60} \sum_{i=t-59}^{t} ProdMin^E(i),$$

où $ProdMin^E$ est la production par minute de l'éolienne E. La production totale du parc notée $ProdH_1^{Parc}$ est la somme des productions des éoliennes.

$$ProdH_1^{Parc}(t) = \sum_{E=1}^{NT} ProdH_1^E(t),$$

où NT est le nombre total d'éoliennes. Cette situation exclut beaucoup de données car il arrive souvent que les éoliennes ne fonctionnent pas en même temps et sans arrêt pendant 60 minutes. C'est pourquoi nous avons défini un deuxième type de production.

- Type de production 2 : on ne tient compte que des périodes où au moins une éolienne n'a eu d'arrêt ou de dysfonctionnement durant les 60 minutes précédant l'instant t. On calcule la moyenne horaire de production des éoliennes ayant fonctionné durant toutes les 60 minutes comme précédemment :

$$ProdH_1^E(t) = \frac{1}{60} \sum_{i=t-59}^{t} ProdMin^E(i).$$

On note Ne ce nombre d'éoliennes. La production horaire du parc est :

$$ProdH_2^{parc}(t) = \left(\frac{1}{Ne} \sum_{E=1}^{Ne} ProdH_1^E(t) \right) \times NT.$$

Si toutes les éoliennes ont fonctionné durant les 60 minutes, le type de production 2 est égale au type de production 1 ($ProdH_2^{parc}(t) =$

$ProdH_1^{parc}(t)$). Si au moins une éolienne n'a pas fonctionné toutes les 60 minutes, $ProdH_2^{parc}(t)$ est une production corrigée du parc en utilisant uniquement les éoliennes ayant fonctionné durant les 60 minutes. Il peut arriver qu'aucune éolienne n'a fonctionné sans arrêt durant les 60 minutes. Nous avons décidé d'exclure cette situation dans l'analyse car elle correspond à beaucoup de dysfonctionnements et peut fausser l'analyse des performances des prévisions.

Nous avons décidé dans la suite du manuscrit d'utiliser le type de production 2 qui sera noté P sauf autre indication. En effet le type de production 1 exclut beaucoup de données.

Les données météorologiques au point de grille i sont notées $X(i)$ et mesurent U_i, V_i, W_i, Dir_i, T_i, Pr_i et TP_i. La figure 2.3 montre les 16 points de grille du maillage ECMWF les plus proche d'un parc éolien.

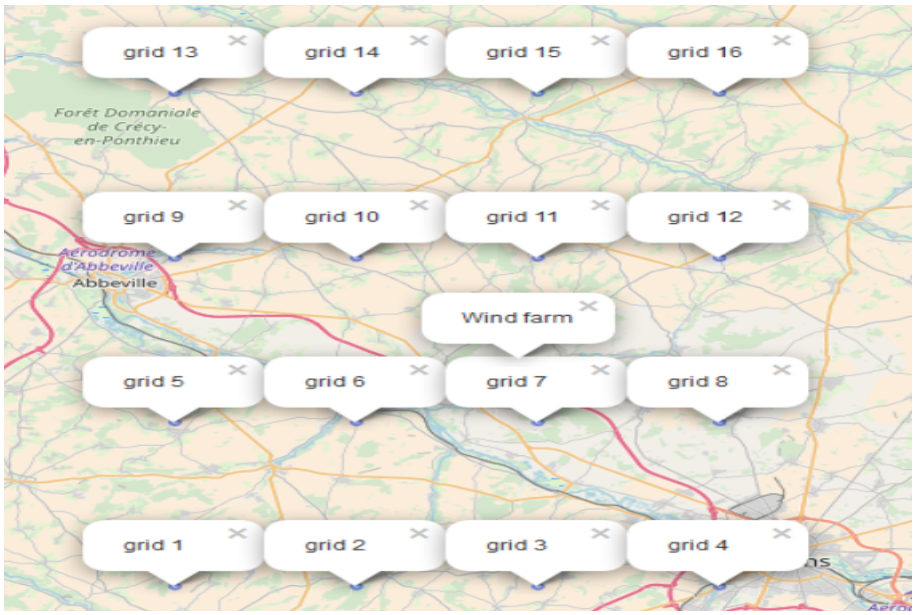


FIGURE 2.3 – Maillage ECMWF au tour d'un parc.

Pour chaque approche, trois situations sont analysées pour étudier l'impact de l'aspect spatial :

- utiliser comme entrée X, les données météorologiques au point de grille le plus proche d'un parc ($X = X(7)$ pour le parc 1),
- aux 4 points de grille les plus proches $X = (X(6), X(7), X(10), X(11))$
- aux 16 points de grille les plus proches d'un parc $X = (X(1), \cdots, X(16))$.

Les données sont divisées en deux échantillons : un échantillon de 2 ans pour l'apprentissage (2015 et 2016) un échantillon d'une année (2017) pour

tester les modèles.

Approche directe :

- apprentissage (2015 et 2016) :
  - prévision directe : $P = h(X) + \zeta$, apprendre le modèle de prévision de la production du parc et obtenir l'estimateur (ou le modèle de prévision) h,
- test (2017) : chaque jour, utiliser les données météorologiques X (horizon 24h à 47h) pour prévoir la production $\hat{P} = \hat{h}(X)$.

Approche indirecte :

- Apprentissage (2015 et 2016) :
  - estimation de la courbe de puissance réelle : $P = g(W_{nacelle}) + \eta$ par spline de lissage (voir section 2.4.7) pour obtenir $\hat{g}$ qui est l'estimation de la fonction g inconnue,
  - prévision indirecte : $W_{nacelle} = h(X) + \zeta$, apprendre le modèle de prévision de la vitesse du vent sur le parc et obtenir l'estimateur (ou le modèle de prévision) h,
- test (2017) : chaque jour, utiliser les données météorologiques X (horizon 24h à 47h) pour prévoir le vent sur le site $\hat{W}_{nacelle} = \hat{h}(X)$ et transformer ce vent en prévision de production $\hat{P} = \hat{g}(\hat{W}_{nacelle})$.

2.4 Méthodologie et algorithme de machine learning

À partir des années 1980, le développement de l'informatique a facilité la mise en œuvre de méthodes non linéaires. Des arbres de classification et de régression ont été introduits par [Stone et al., 1984]. Breiman a ensuite introduit les forêts aléatoires en 2001 [Breiman, 2001]. En général, le cadre statistique est le suivant : nous voulons prédire une variable Y à partir de p variables explicatives $X = (X_1, X_2, \ldots, X_p)$ qui sont des vecteurs. De manière générale, le modèle statistique s'écrit : $Y = f(X) + \epsilon$, où f est une fonction (Modèle) à estimer en minimisant les erreurs ϵ. Dans cette étude, Y sera soit la production du parc éolien P, soit la vitesse du vent au sommet de la nacelle $W_{nacelle}$. Un ensemble de techniques de modélisation et de méthodes d'apprentissage statistique est présenté dans cette partie. Nous allons d'abord faire la différence entre l'approche directe et l'approche indirecte.

Nous avons appliqué dans ce chapitre quelques modèles couramment utilisés dans la littérature.

2.4.1 Persistance

Le modèle de persistance est un modèle de référence très utilisée dans le cadre de la prévision de l'énergie éolienne. Elle consiste simplement à utiliser la dernière observation comme prévision pour tous les horizons [Fugon et al., 2008]. Plus précisément, $\hat{P}_{t_0+h} = P_{t_0}$ avec $\hat{P}$ représentant les prévisions, P les mesures, t_0 le temps initial des prévisions et h l'horizon de prévision.

2.4.2 Bagging

Les arbres de décision (voir [Stone et al., 1984] pour les détails) souffrent d'une grande variance. C'est-à-dire si on sépare aléatoirement les données d'apprentissage en deux parties, et on ajuste un arbre de décision sur chaque ensemble, les résultats obtenus pourraient être très différents. Une manière naturelle de réduire la variance est d'avoir plusieurs échantillons de la population, de construire un modèle prédictif sur chaque échantillon séparément et faire la moyenne des prévisions. En d'autres termes, on calcule $\hat{f}^1(x)$, $\hat{f}^2(x)$, ..., $\hat{f}^B(x)$ en utilisant B ensembles d'apprentissage et on fait la moyenne,

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x),$$

pour obtenir une variance plus faible. En pratique on ne dispose pas de plusieurs échantillons. Cependant on peut obtenir plusieurs échantillons B par bootstrap de manière à calculer $\hat{f}^{*b}(x)$ et de faire la moyenne des prévisions pour obtenir,

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

ceci est appelé le bagging. Le principe reste le même dans le cadre d'une régression. A la place des arbres de décision on construira des arbres de régression dans le cadre de l'utilisation du bagging pour la régression. L'idée du Bagging, et qu'en appliquant la règle de base sur différents échantillons bootstrap, on en modifie les prévisions, et donc on construit à terme une collection de prédicteurs variés. L'étape d'agrégation permet alors d'obtenir un prédicteur performant (Voir [Genuer, 2010] pour plus de détails).

2.4.3 Boosting

Le Boosting, introduit par [Freund and Schapire, 1996] en 1996 est une des méthodes d'ensemble les plus performantes à ce jour ([Genuer, 2010]). C'est une approche générale qui peut-être appliquée à plusieurs méthodes d'apprentissage statistique pour la régression ou la classification. Le boosting procède de manière similaire au bagging, excepté du fait que les arbres sont construits séquentiellement : chaque arbre est construit en utilisant les informations de l'arbre précédent. Le principe du boosting est de tirer un premier échantillon bootstrap $B_n^{\Theta 1}$, où chaque observation a une probabilité $1/n$ d'être tirée. La variable aléatoire $\Theta 1$ représente alors ce tirage aléatoire. Puis d'appliquer une règle de base (arbre de régression dans ce cas d'étude) pour obtenir un premier prédicteur $\hat{f}(., B_n^{\Theta 1})$. Ensuite, l'erreur de $\hat{f}(., B_n^{\Theta 1})$ sur l'échantillon d'apprentissage est calculée. On tire un deuxième échantillon bootstrap dont la loi du tirage des observations n'est plus uniforme. La probabilité pour une observation d'être tirée dépend de la prévision $\hat{f}(., B_n^{\Theta 1})$ sur cette observation. Le principe est d'augmenter la probabilité de tirer une observation mal prédite, et de diminuer celle de tirer une observation bien prédite. On applique la règle de base sur ce nouvel échantillon puis on réitère le processus. La collection de prédicteurs obtenus est alors agrégée en faisant une moyenne pondérée.

2.4.4 Les forêts aléatoires

Depuis son introduction par L. Breiman [Breiman, 2001] en 2001, beaucoup de publications ont abordé la théorie des forêts aléatoires avec des applications dans plusieurs domaines. L'aspect théorique des forêts aléatoires a été bien abordé dans [Genuer, 2010]. Le principe des forêts aléatoires est tout d'abord de générer plusieurs échantillons bootstrap $\mathcal{B}_n^{\Theta_1}, \ldots, \mathcal{B}_n^{\Theta_B}$. Ensuite, sur chaque échantillon $\mathcal{B}_n^{\Theta_l}$, une variante de CART (Classification And Regression Tree) est expliquée. En d'autre terme, un arbre est construit de façon suivante. Le découpage d'un nœud se fait par un tirage aléatoire de m variables et la recherche de la meilleure coupure suivant les m variables sélectionnées. De plus, l'arbre construit est complètement développé (arbre maximal) et non élagué. L'élagage consiste à chercher le meilleur sous-arbre élagué de l'arbre maximal (meilleur au sens de l'erreur de généralisation). La collection d'arbres obtenus est enfin agrégée (moyenne en régression, vote majoritaire en classification) pour donner le prédicteur de forêts aléatoires. Le tirage, à chaque nœud, des m variables se fait, sans remise, et uniformément parmi toutes les variables (chaque variable a une probabilité $1/p$ d'être

choisie). Le nombre m ($m \leqslant p$) est fixé au début de la construction de la forêt et est donc identique pour tous les arbres. C'est un paramètre très important de la méthode.
Pour les Random Forests, il y a donc deux sources d'aléas pour générer la collection des prédicteurs individuels : l'aléa dû au bootstrap et l'aléa du choix des variables pour découper chaque noeud d'un arbre. Ainsi, on perturbe à la fois l'échantillon sur lequel on lance la règle de base, et le coeur de la construction de la règle de base. Ce tirage aléatoire de variables pour découper un nœud avait déjà été utilisé par Amit and Geman (1997) dans des problèmes de reconnaissance d'image. Leur méthode a beaucoup influencé Leo Breiman dans sa mise au point des Random Forests.

2.4.5 Support Vector Machine Regression

Les Support Vector Machines sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination, c'est-à-dire la prévision d'une variable qualitative binaire [Cortes and Vapnik, 1995]. Ils ont ensuite été généralisés à la prévision d'une variable quantitative. Nous examinons d'abord le cas de la régression linéaire. Soit un échantillon (Y, X) où X est un ensemble multivarié de N observations avec des valeurs de réponse observées Y. L'objectif est de trouver une fonction $f(X) = X'\beta + b$ qui s'écarte de Y d'une valeur non supérieure à ε pour chaque point d'entraînement X, et qui est en même temps aussi proche que possible. La formulation en problème d'optimisation convexe est de minimiser :

$$J(\beta) = \frac{1}{2}\beta'\beta$$

sous réserve que tous les résidus ayant une valeur inférieure à ε ; ou, sous forme d'équation :

$$\forall\, i \,:\, |Y_i - f(X_i))| \leq \varepsilon.$$

S'il n'y a pas une telle fonction $f(.)$ qui satisfait ces contraintes pour tous les points, les variables d'écart ξ_i et ξ_i^* sont introduites pour chaque point. L'inclusion de variables d'écart mène à la fonction objectif :

$$J(\beta) = \frac{1}{2}\beta'\beta + C\sum_{i=1}^{N}(\xi_i + \xi_i^*),$$

sous contraintes :

$$\forall\, i \,:\, Y_i - f(X_i) \leq \varepsilon + \xi_i,$$

$$\forall\, i \,:\, f(X_i) - Y_i \leq \varepsilon + \xi_i^*,$$

$$\forall\, i \,:\, \xi_i \,, \xi_i^* \geq 0.$$

C est une constante positive qui contrôle la pénalité infligée pour les observations se situant en dehors de la marge epsilon (ε) et aide à prévenir les sur-ajustements (régularisations).
Le problème d'optimisation décrit précédemment est plus simple à résoudre en calcul dans sa formulation double de Lagrange en introduisant des multiplicateurs non négatifs α_i et α_i^* pour chaque observation X_i.
La formule duale est de minimiser :

$$L(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) X_i' X_j + \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} Y_i (\alpha_i + \alpha_i^*),$$

sous contraintes :

$$\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0,$$

$$\forall\, i \,:\, 0 \leqslant \alpha_i \leqslant C,$$

$$\forall\, i \,:\, 0 \leqslant \alpha_i^* \leqslant C.$$

Le paramètre β peut-être complètement décrit comme une combinaison linéaire des observations d'apprentissage en utilisant l'équation $\beta = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) X_i$.
La fonction utilisée pour prédire de nouvelles valeurs ne dépend que des vecteurs de support :

$$f(X_j) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)(X_i' X_j) + b.$$

Les conditions de complémentarité Karush-Kuhn-Tucker (KKT) sont des contraintes d'optimisation requises pour obtenir des solutions optimales. Pour la régression SVM linéaire, ces conditions sont les suivantes :

$$\forall\, i \,:\, \alpha_i(\varepsilon + \xi_i - Y_i + X_n^{'}\beta + b) = 0$$

$$\forall\, i \,:\, \alpha_i^*(\varepsilon + \xi_i^* + Y_i - X_n^{'}\beta - b) = 0$$

$$\forall\, i \,:\, \xi_i(C - \alpha_i) = 0$$

$$\forall\, i \,:\, \xi_i^*(C - \alpha_i^*) = 0.$$

Ces conditions indiquent que toutes les observations strictement à l'intérieur du tube epsilon ont des multiplicateurs de Lagrange $\alpha = 0$ et $\alpha^* = 0$. Si α_i ou α_i^* n'est pas nul, l'observation correspondante est appelée vecteur de support. Pour obtenir une régression SVM non linéaire, on peut remplacer le produit $X_i^{'}X_j$ avec une fonction du noyau non linéaire $G(X_i^{'}, X_j)$ (par exemple un noyau Gaussien $G(X_i, X_j) = \exp(- \parallel X_i - X_j \parallel^2)$, ou noyau polynomial $G(X_i, X_j) = (1 - X_i^{'}X_j)^q$, où $q \in \{2, 3, ...\}$).

2.4.6 Modèles additifs généralisés

Les modèles additifs généralisés ont été introduits par Hastie et Tibshirani en 1986 [Hasti and Tibshirani, 1986]. Ils constituent une généralisation de la régression multiple. En régression linéaire, on calcule un ajustement linéaire des moindres carrés pour un ensemble de variables $X = (X_1, \cdots, X_p)$ pour prédire une variable Y. L'équation de régression linéaire peut être formulée comme suit :

$$Y = b_0 + b_1X_1 + \cdots + b_pX_p + \varepsilon. \tag{2.1}$$

Les modèles additifs généralisés remplacent la forme linéaire ($\sum \beta_j X_j$) avec une somme de fonctions lisses $\sum s_j(X_j)$. Les $s_j(.)$ sont des fonctions inconnues qui peuvent être estimées par n'importe quel lisseur (moindres carrés, splines, ...) avec une procédure locale de scoring (voir [Hasti and Tibshirani, 1986] pour plus de détails).

2.4.7 Estimation sur les bases de splines

Dans cette partie nous faisons un résumé de la notion des fonctions de splines que nous utiliserons pour calibrer le modèle d'estimation de la courbe de puissance réelle dans l'approche indirecte. Le terme "fonctions de splines" tire ses origines des travaux de Whittaker (1923) sur les méthodes de graduation de données avant d'être introduit en 1946 par Schoenberg. Cependant c'est au début des années 1960 que la théorie des splines s'est développée.

Leur utilisation comme méthode de régression non paramétrique est attribuée à Wahba (1975), qui a démontré leurs propriétés statistiques. Notons également que ses travaux en collaboration (Kimerldorf et Wahba (1971), Wahba et Wold (1975) et Craven et Wahba (1979)) ont favorisé le développement des splines de lissage. Pour obtenir des estimateurs qui sont des polynômes par morceaux et qui ont les propriétés de régularité, on utilise les bases de splines.

Splines linéaires et cubiques

Considérons K nœuds $x_1, x_2, \ldots, x_K$ tels que $0 < x_1 < x_2 < x_3 < \cdots < x_K$ et,

$$f(x) = \beta_0 + \beta_1 x + \beta_2(x - x_1)_+ + \beta_3(x - x_2)_+ + \beta_4(x - x_3)_+ + \ldots + \beta_{K+1}(x - x_K)_+.$$

Les nœuds $x_1, x_2, \ldots, x_K$ déterminent les intervalles de la partition.

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x \\ &= \beta_0 + \beta_1 x + \beta_2(x - x_1)_+ \text{ si } x_1 \leqslant x \leqslant x_2 \\ &= \beta_0 + \beta_1 x + \beta_2(x - x_1)_+ + \beta_3(x - x_2)_+ \text{ si } x_2 \leqslant x \leqslant x_3 \end{aligned}$$

La fonction f est continue, si on veut imposer plus de régularité (par exemple de classe $\mathcal{C}^2$), on utilise des splines cubiques :

$$f(x) = \beta_0 + \beta_1 x^3 + \beta_2(x - x_1)_+^3 + \beta_3(x - x_2)_+^3 + \beta_4(x - x_3)_+^3 + \ldots + \beta_{K+1}(x - x_K)_+^3$$

La fonction $(x - x_i)^3$ s'annule ainsi que ses dérivées d'ordre 1 et 2 en x, donc f est de classe $\mathcal{C}^2$. Pour éviter les problèmes de bords, on impose souvent des contraintes supplémentaires aux splines cubiques, notamment la linéarité de la fonction sur les deux intervalles correspondant aux extrémités. On se place sur $[0, 1]$. Soit $\xi_0 < \xi_1 < \ldots < \xi_K < 1$.

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3.$$

On impose $f''(0) = f^{(3)}(0) = 0, \quad f''(\xi_K) = f^{(3)}(\xi_K) = 0$. On en déduit :

$$\beta_2 = \beta_3 = 0, \quad \sum_{k=1}^{K} \theta_k(\xi_K - \xi_k) = 0, \quad \sum_{k=1}^{K} \theta_k = 0,$$

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k \left[(x - \xi_k)_+^3 - (x - \xi_K)_+^3\right] && (2.2) \\ &= \beta_0 + \beta_1 x + \sum_{k=1}^{K-1} \theta_k (\xi_K - \xi_k) \left[\frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{(\xi_K - \xi_k)}\right]. && (2.3) \end{aligned}$$

On pose $\gamma_k = \theta_k(\xi_K - \xi_k)$ et $d_k = \dfrac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{(\xi_K - \xi_k)}$, $\sum_{k=1}^{K-1} \gamma_k = 0$ donc

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \gamma_k (d_k(x) - d_{K-1}(x)).$$

On obtient la base de splines naturelles :

$$N_1(x) = 1,\ N_2(x) = x,\ \forall 1 \leqslant k \leqslant K-2,\ N_{k+2}(x) = d_k(x) - d_{K-1}(x).$$

On doit choisir la position et le nombre de nœuds. Le choix des nœuds est parfois difficile et délicat. Une manière de s'affranchir de ce choix consiste à considérer n nœuds correspondant aux valeurs de l'échantillon observé.

Méthodes de régularisation

On se place dans un modèle de régression : $Y_i = f(X_i) + \epsilon_i,\ 1 \leqslant i \leqslant n$.
On minimise parmi les fonctions f splines naturelles de nœuds en les X_i ($f(x) = \sum_{k=1}^{n} \theta_k N_k(x)$) le critère pénalisé :

$$C(f, \lambda) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 \left(f''(t)\right)^2 dt,$$

où $\lambda > 0$. Le premier terme mesure la distance entre le modèle et les données observées alors que le second pénalise la courbure de la fonction. λ est un paramètre de lissage, il établit un équilibre entre ces deux termes. Notons que dans le cas $\lambda = 0$, f peut être n'importe quelle fonction qui interpole les données et dans le cas $\lambda = \infty$ on se ramène à une régression linéaire simple (la dérivée seconde étant forcée à zéro). Un résultat remarquable qui justifie l'introduction des fonctions splines est le suivant : lorsque λ est fini non nul, ce critère défini sur un espace de Sobolev de fonctions pour lesquelles le second terme est défini, admet une solution explicite, fini-dimensionnelle : une spline cubique naturelle à n nœuds localisés aux valeurs de l'échantillon. En notant $\Omega_{l,k} = \int_0^1 N_k''(x) N_l''(x) dx$ et $N_{i,j} = N_j(X_i)$, le critère à minimiser est

$$C(\theta, \lambda) = \| Y - H\theta \|^2 + \lambda \theta^* \Omega \theta.$$

La solution est :

$$\hat{\theta} = (N^* N + \lambda \Omega)^{-1} N^* Y$$

et

$$\hat{f}(x) = \sum_{k=1}^{n} \hat{\theta}_k N_k(x).$$

Enfin, une validation croisée permet de choisir le paramètre de lissage λ.

2.5 Résultats et discussions

Dans cette partie nous avons appliqué les algorithmes de machine learning décrits dans la section 2.4 pour la prévision de quatre parcs éoliens (parc 1, parc 2, parc 3 et parc 4).
Pour chaque modèle l'accent est mis sur l'approche directe et indirecte et l'utilisation des données météorologiques sur plusieurs points de grille. L'horizon de prévision est de $24h$ à $47h$. Pour rappel l'apprentissage est fait sur deux ans (2015 et 2016) et le test sur une année (2017). Le $NMAE$ est utilisé comme indicateur de performance.

2.5.1 Approche directe et approche indirecte

TABLE 2.1 – $NMAE$ parc 1

methods	approach	1 grid point	4 grid points	16 grid points
RF	direct	8.56	8.65	8.38
	indirect	7.70	7.76	**7.62**
Bagging	direct	8.32	8.24	8.15
	indirect	7.92	7.77	**7.70**
Boosting	direct	8.34	8.35	8.35
	indirect	**8.07**	8.10	8.23
SVM	direct	8.17	8.50	8.74
	indirect	**7.88**	7.99	8.06
GAM	direct	7.98	8.74	9.75
	indirect	**7.65**	8.09	8.85
Persistance	20.50			

TABLE 2.2 – $NMAE$ parc 2

methods	approach	1 grid point	4 grid points	16 grid points
RF	direct	8.92	8.84	8.75
	indirect	8.56	8.42	**8.21**
Bagging	direct	9.09	8.87	8.80
	indirect	8.66	8.48	**8.26**
Boosting	direct	9.18	9.25	9.15
	indirect	8.74	8.76	8.61
SVM	direct	8.87	8.75	8.47
	indirect	8.60	8.48	**8.21**
GAM	direct	9.11	9.11	8.93
	indirect	8.54	8.43	**8.31**
Persistance	20.96			

TABLE 2.3 – $NMAE$ parc 3

methods	approach	1 grid point	4 grid points	16 grid points
RF	direct	8.64	8.50	8.34
	indirect	8.60	8.50	**8.28**
Bagging	direct	8.75	8.54	8.38
	indirect	8.73	8.54	**8.32**
Boosting	direct	8.66	8.57	8.38
	indirect	8.63	8.56	**8.38**
SVM	direct	8.34	8.27	**8.12**
	indirect	8.39	8.32	8.17
GAM	direct	8.65	8.41	8.34
	indirect	8.53	8.40	**8.32**
Persistance	21.86			

TABLE 2.4 – $NMAE$ parc 4

methods	approach	1 grid point	4 grid points	16 grid points
RF	direct	7.19	7.16	6.99
	indirect	6.90	6.77	**6.62**
Bagging	direct	7.28	7.21	7.04
	indirect	6.90	6.77	**6.65**
Boosting	direct	7.73	7.61	7.48
	indirect	7.27	7.11	7.04
SVM	direct	6.87	6.88	6.81
	indirect	6.77	6.74	**6.65**
GAM	direct	7.25	7.03	6.93
	indirect	6.90	6.69	**6.51**
Persistance	19.14			

Les résultats empiriques obtenus montrent que dans les deux approches (directe et indirecte) les algorithmes de machine learning ont à peu près les mêmes performances avec un léger avantage des forêts aléatoires, des SVMs et des GAMs. Tous les modèles statistiques donnent de meilleures prévisions que la persistance. L'approche indirecte fournie de meilleures prévisions que l'approche directe sur les parcs éoliens 1, 2 et 4 avec tous les algorithmes utilisés. Sur le parc éolien 3 l'approche indirecte fait aussi bien que l'approche directe. Empiriquement les résultats montrent que l'approche indirecte donne de meilleures prévisions que l'approche directe.

Dans l'approche indirecte on a deux étapes : une étape d'estimation du vent sur le site et une étape de conversion de ce vent en prévision de production. Dans la première étape, on utilise les données météorologiques à des hauteurs différentes, tableau 1.2 (au point de grille le plus proche, aux quatre points de grille les plus proches ou aux seize points de grille les plus proches du site) pour prévoir le vent à la hauteur de la nacelle. Si on compare ce vent prévu à hauteur de la nacelle avec le vent nacelle alors on a un MAE $0.90\,m/s$ sur la période d'apprentissage alors que si on utilisait le vent prévu à 100m on aurait un MAE de $1.28\,m/s$. Cette première étape permet de mieux prévoir le vent nacelle.

Dans la deuxième étape de cette approche indirecte, on utilise la courbe de puissance réelle qui résulte d'une estimation par spline à l'aide du vent réel et de le production réelle qui généralement est très lisse avec moins d'incertitudes.

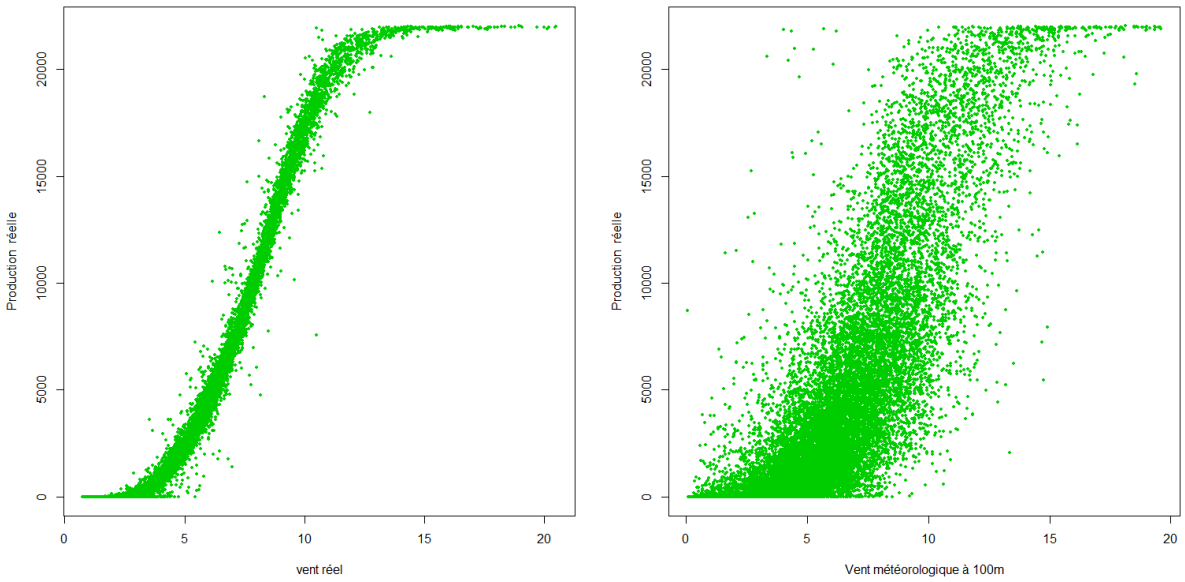


FIGURE 2.4 – Courbe de puissance avec le vent nacelle à gauche et courbe de puissance avec le vent météorologique à droite

En guise d'illustration, sur la figure 2.4 on note beaucoup d'incertitude sur la courbe de puissance avec le vent météorologique (courbe de puissance à droite) contrairement à la courbe de puissance avec le vent nacelle (courbe de puissance à gauche) qui est beaucoup plus lisse. Ceci n'est qu'une illustration dans le cas univarié mais elle met en évidence l'apport de la fonction de transfert dans la deuxième étape de l'approche indirecte. En définitive, nous estimons que l'amélioration des prévisions du vent météorologique et l'utilisation des mesures réelles du vent sur le site (avec la courbe de puissance réelle) expliquent le résultat empirique de la meilleure performance de l'approche indirecte par rapport à l'approche directe.
On note également qu'avec les forêts aléatoires la prise en compte de plusieurs points de grille (notamment les seize points de grille) réduit les erreurs de prévision de la production éolienne. On a pratiquement la même conclusion avec le Bagging qui est un cas particulier des forêts aléatoires de même que le Boosting. Sur les parcs éoliens 2, 3 et 4 on observe également que l'utilisation de plusieurs points de grille réduit les $NMAE$ en utilisant les GAMs ou les SVMs. De manière générale l'intégration de plusieurs points de grille avec l'approche indirecte a donné les erreurs de prévision les plus faibles. Les performances des prévisions dépendent d'un site à l'autre. La production des parcs éoliens 1 et 4 est plus facile à prédire que celle des parcs 2 et 3. Dans la section suivante on s'est intéressé aux erreurs de prévision par horizon.

2.5.2 Performance des modèles par horizon de prévision

Après une première analyse des résultats sur la performance globale des des méthodes étudiées on s'intéresse à la performance des prévisions selon l'horizon. A cet effet on s'est restreint à l'approche indirecte avec les prévisions météorologiques aux 16 points de grille les plus proches de chaque parc. Le modèle de persistance n'est pas analysé dans cette section (ses performances étant de loin inférieures à celles des modèles de machine learning). Pour chaque horizon de prévision, nous calculons les erreurs de prévision pour chaque parc (figures 2.5, 2.6, 2.7 et 2.8).

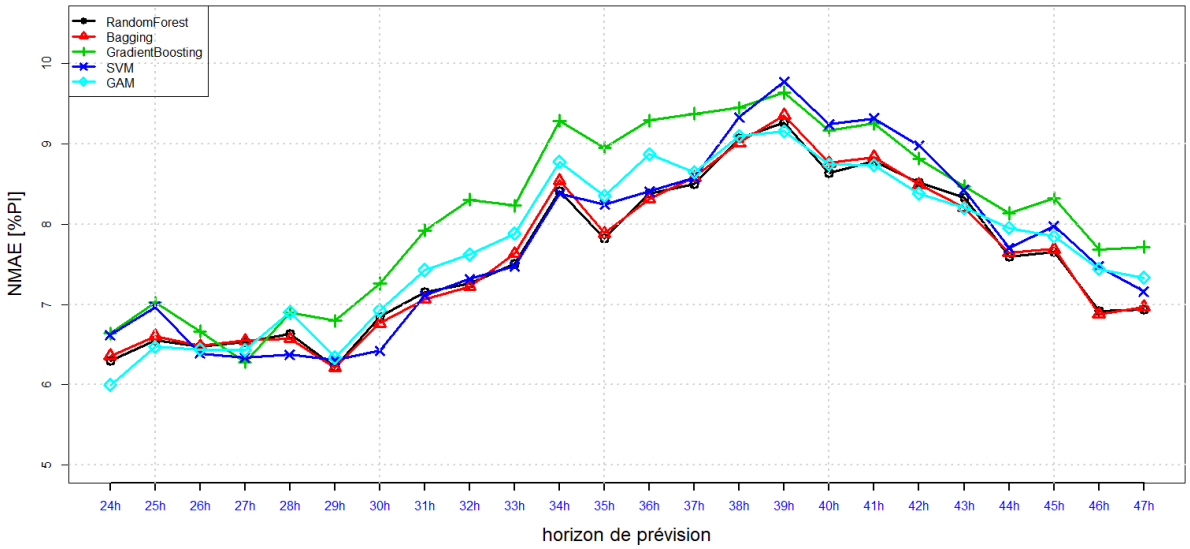


FIGURE 2.5 – NMAE par horizon, approche indirecte avec 16 points de grille : Parc éolien 1

Les erreurs de prévision sont plus faibles quand les prévisions sont moins éloignées. On retrouve ici les résultats vus en analyse canonique (section 1.4.2) où la corrélation entre les données météorologiques et le pylône étaient plus fortes dans la nuit que dans la journée (figure 1.8).

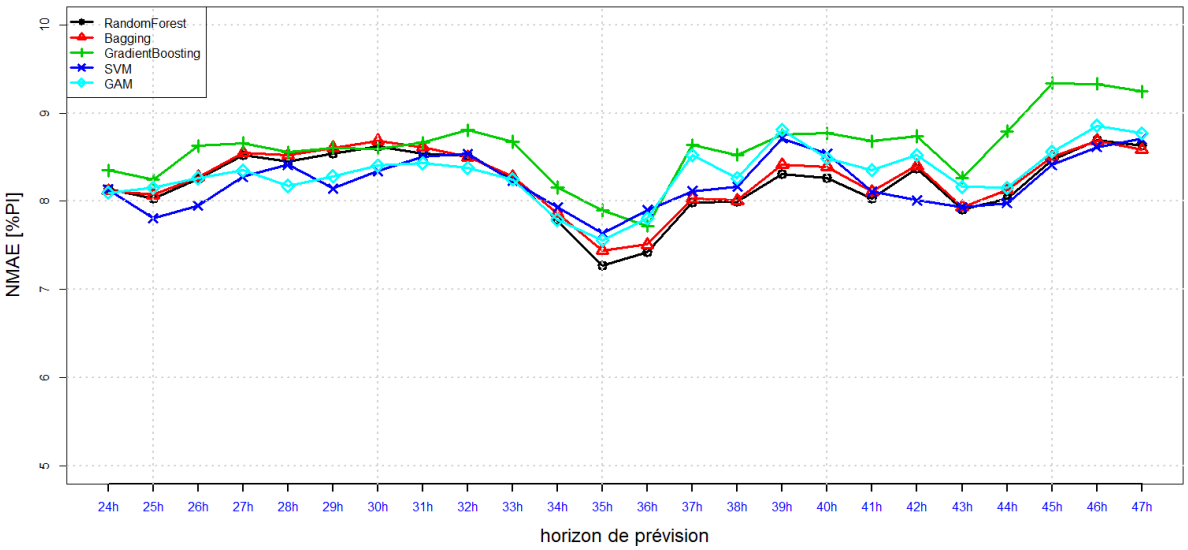


FIGURE 2.6 – NMAE par horizon, approche indirecte avec 16 points de grille : Parc éolien 2

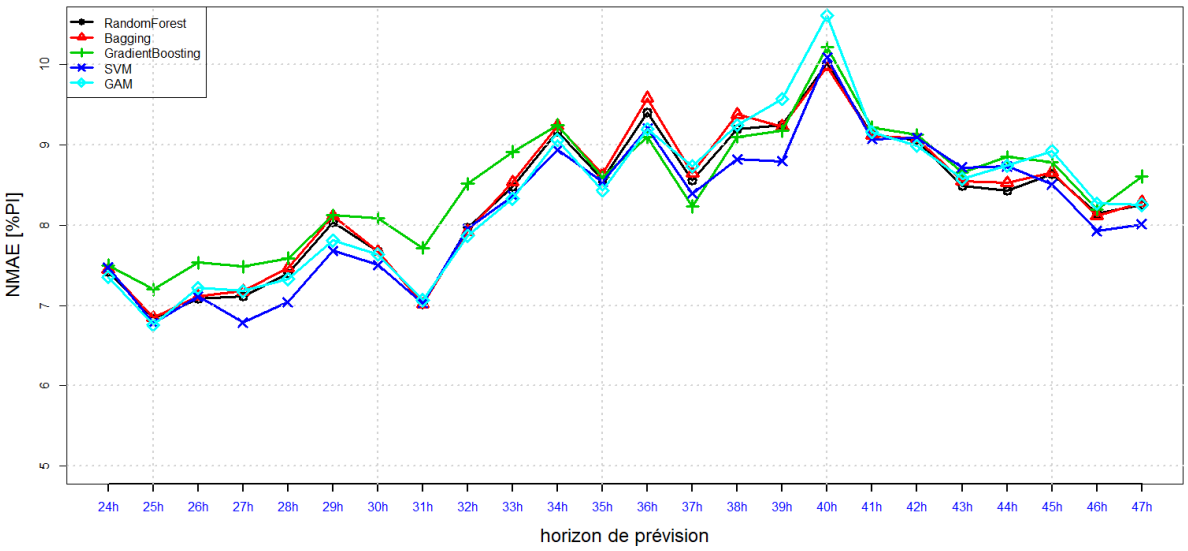


FIGURE 2.7 – NMAE par horizon, approche indirecte avec 16 points de grille : Parc éolien 3

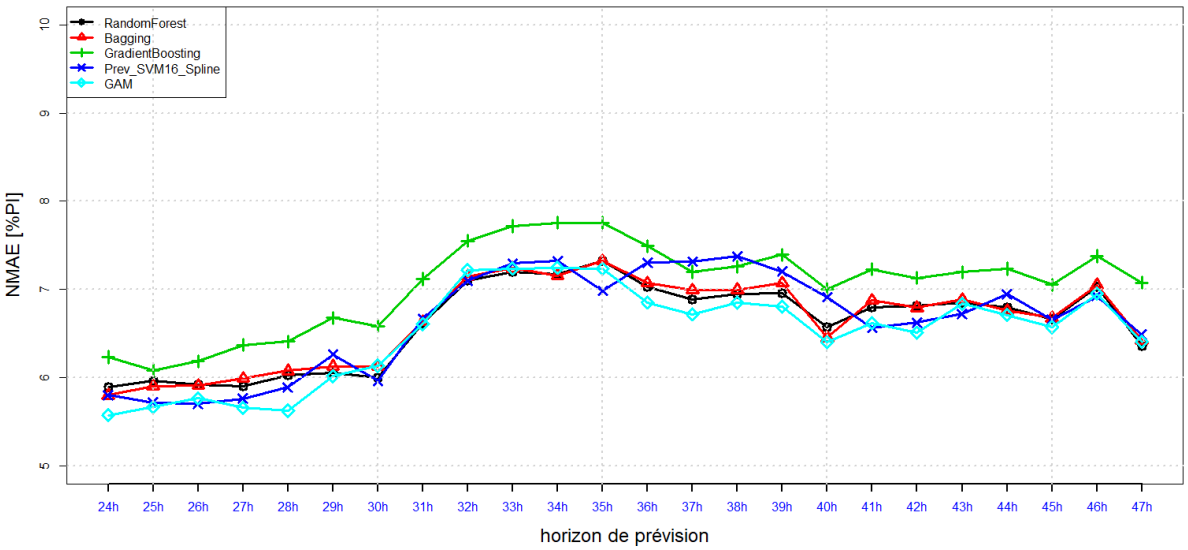


FIGURE 2.8 – NMAE par horizon, approche indirecte avec 16 points de grille : Parc éolien 4

En observant les erreurs par horizon de prévision, on note une légère augmentation des erreurs de prévision avec l'horizon de prévision. Si l'horizon de prévision augmente, les prévisions météorologiques sont moins fiables et les erreurs de prévision de la production éolienne sont plus importantes. On note une hausse des erreurs de prévision au milieu de la journée. En effet, les prévisions météorologiques sont en général plus stables dans la nuit. Pour tous les quatre parcs (figures 2.5, 2.6, 2.7 et 2.8), il n'y a pas un modèle qui est meilleur sur tous les horizons. Les Forêts aléatoires, les GAMs et les SVMs restent légèrement plus performants.

2.6 Conclusion

Dans ce chapitre on a abordé la prévision court-terme de la production éolienne par machine learning. L'intégration de plusieurs points de grille a permis d'améliorer par moment les erreurs de prévision. Les résultats montrent également que la performance des prévisions dépend de la complexité des parcs. La prévision de certains parcs éoliens est moins incertaine que celle d'autres parcs. La contribution majeure de ce chapitre est la performance de l'approche indirecte par rapport à l'approche directe. Les résultats de ce chapitre montrent que si la courbe de puissance réelle est « lisse », l'estimation du vent sur le site puis la transformation de ce vent en prévision de production

est en général plus performante que la prévision directe de la production éolienne. Avec une moyenne des erreurs absolues de 6, 7 à 8% sur un horizon de prévision de 24 à 47 heures, les méthodes de machine learning donnent des résultats très concluants de la prévision éolienne. Une amélioration des résultats est possible en intégrant les aspects métiers tels que la dynamique du vent avec du feature engineering.

Chapitre 3

Modélisation spatio-temporelle et sélection de points de grille pour la prévisions

Sommaire

3.1 Introduction

Dans un souci d'amélioration des prévisions de la production éolienne, il est important de prendre en compte les aspects métiers notamment de la météorologie. A partir des prévisions météorologiques un certain nombre de phénomènes peuvent être modélisés par post-engineering pour servir de co-variables et potentiellement améliorer les prévisions de la production. Dans le chapitre 2 et dans la plupart des études de prévision de la production éolienne, les prévisions sont faites sans tenir compte correctement des dépendances spatio-temporelles observées dans le domaine. Cependant depuis quelques

années, la structure spatio-temporelle commence à être intégrée dans la prévision éolienne [Tastu et al., 2011]. En prenant comme référence les travaux du chapitre 2, nous nous sommes intéressés d'abord à l'apport de la dynamique spatio-temporelle du vent 3.2 et à l'impact la turbulence sur la prévision de la production éolienne 3.3. Ensuite nous avons étudié la sélection de points de grille 3.4 en se basant sur une mesure d'importance des points de grille.

Les résultats sur la dynamique spatio-temporelle du vent ont été publié dans [Dione and Matzner-Løber, 2019].

3.2 L'impact des dérivées spatio-temporelles du vent

La modélisation de la dynamique spatio-temporelle du vent dans la prévision de la production éolienne n'est pas nouvelle. Plusieurs techniques de modélisation existent dans la littérature. L'analyse et la modélisation spatio-temporelle pour la prévision court terme ont été déjà étudiées dans l'article [Tastu et al., 2011]. Avec des prévisions de la production de 22 parcs éoliens situés au Danemark, les auteurs ont démontré une structure spatio-temporelle des erreurs de prévision de la production, montré l'impact de la vitesse et de la direction sur la nature et la forme de la structure et proposé un modèle pour capturer cette structure. Après un clustering sur les 22 parcs, les ACFs (Auto-Correlation Function) et CCFs (Cross-Correlation Function) ont permis d'étudier la dépendance intra-groupe et intergroupe respectivement pour démontrer la corrélation temporelle et spatiale. Les effets de la vitesse et de la direction du vent ont été examinés en faisant une analyse par secteur (quatre secteurs) et par intervalle de vitesse (l'intervalle $[0, 25]$ divisé en cinq intervalles). Ils conclurent que l'information spatio-temporelle a fait passer leur RMSE (%Pnom) de 11.667% à 6.49%.

Des chaînes de Markov sont utilisées dans l'article [He et al., 2014] pour une analyse spatio-temporelle de la prévision très court terme (horizon 10 minutes) de la production des éoliennes. Les auteurs ont comparé la prévision court terme de distributions avec des approches de modèles autorégressifs d'ordre élevé et les prévisions ponctuelles avec la persistance. Leurs tests numériques ont démontré une performance améliorée avec les chaines de Markov développées par l'analyse spatio-temporelle avec un NMAE de 6.98% pour un horizon de prévision de dix minutes. L'article [Ghaderi et al., 2017] aborde la prévision des vitesses du vent sur plusieurs stations météorologiques (57 au total) dans la côte Est des USA par deep-learning en intégrant l'information spatiale et temporelle. Pour chaque horizon de prévision (de 1h à 6h),

un modèle (AR(2), AR(3), deep-learning, etc) est ajusté en utilisant l'information présente (au temps t) et passée (au temps $t-l$, $t-l+1$, ..., $t-1$). La méthode de deep-learning en utilisant l'information sur toutes les stations a donné le meilleur résultat avec un RMSE de $1.62\ m/s$ contre $2.76\ m/s$ pour les modèles AR. L'article [Lenzi et al., 2017] étudie les prévisions probabilistes de la production éolienne de 349 parcs au Danemark. Les auteurs considèrent la production d'un parc comme un processus indexé par le temps et l'espace qu'ils normalisent par la production totale du parc. Ils appliquent une transformation log-normale à ce processus normalisé pour obtenir un processus de loi normale. A partir de ce processus de loi normale, trois modèles sont considérés : un modèle temporel, un modèle spatio-temporel et un modèle combinant le modèle temporel et le modèle spatio-temporel. Le RMSE et le CPRS (Continuous Ranked Probabilty Score) ont été utilisés comme indicateurs de performance. Deux types d'approches sont considérés : un modèle pour un parc et une agrégation des prévisions sur plusieurs parcs. Les auteurs obtiennent un RMSE entre 3.5% et 12% (horizon de 15 minutes à 5h) avec la modélisation par parc et un RMSE entre 1.8% et 10% avec l'agrégation des prévisions.

3.2.1 Modélisation des variations spatio-temporelles

Nous avons calculé et intégré en entrée de modèle les variations temporelles des différentes variables météorologiques. En effet si le modèle était linéaire, les variations seraient directement pris en compte par le modèle. Cependant le modèle est non linéaire c'est la raison pour laquelle nous utilisons la dynamique spatio-temporelle du vent en entrée de modèle. Considérons par exemple la composante est-ouest du vent à 10 mètres, au point de grille i, en un instant t que nous avons notée $U\_10m_{(i,t)}$. Les variations temporelles sont définies par $U\_10m_{(i,t)} - U\_10m_{(i,t-1)}$. Nous avons fait la même chose pour toutes les variables météorologiques. Le principe est résumé sur la figure 3.1 où $X_{(i,t)}$ représente les données météorologiques en un point de grille i à l'instant t. La généralisation avec les autres points de grille se fait de la même manière.

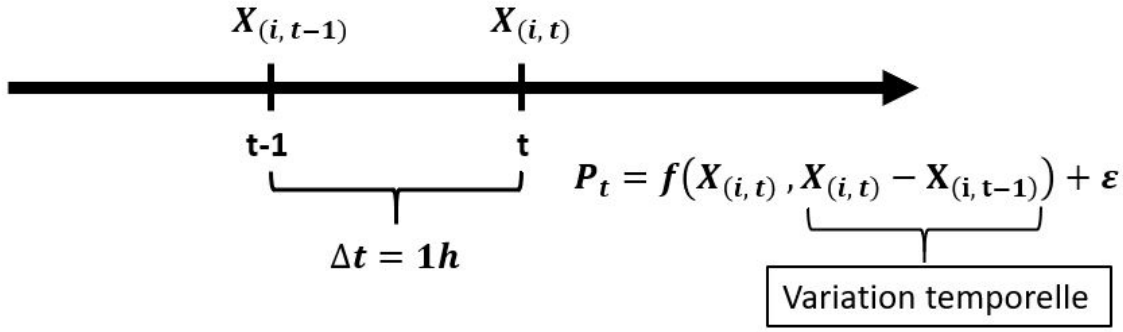


FIGURE 3.1 – Modélisation de la variation temporelle

Pour modéliser la dynamique spatiale du vent à chaque point de grille en un instant t, nous avons calculé pour les composantes du vent $(U,\ V,\ W\ et\ Dir)$ les variations verticales et horizontales entre ce point et les deux autres points sur la grille situés en haut sur l'axe vertical et à gauche sur l'axe horizontal au même instant t. Par exemple pour U_10m au point de grille 6 à l'instant t on a $(U\_10m_{(6,t)} - U\_10m_{(7,t)}, U\_10m_{(6,t)} - U\_10m_{(10,t)})$ comme variations spatiales. La figure 3.2 illustre cet exemple. Le calcul sur les autres points de grille se généralise de la même manière.

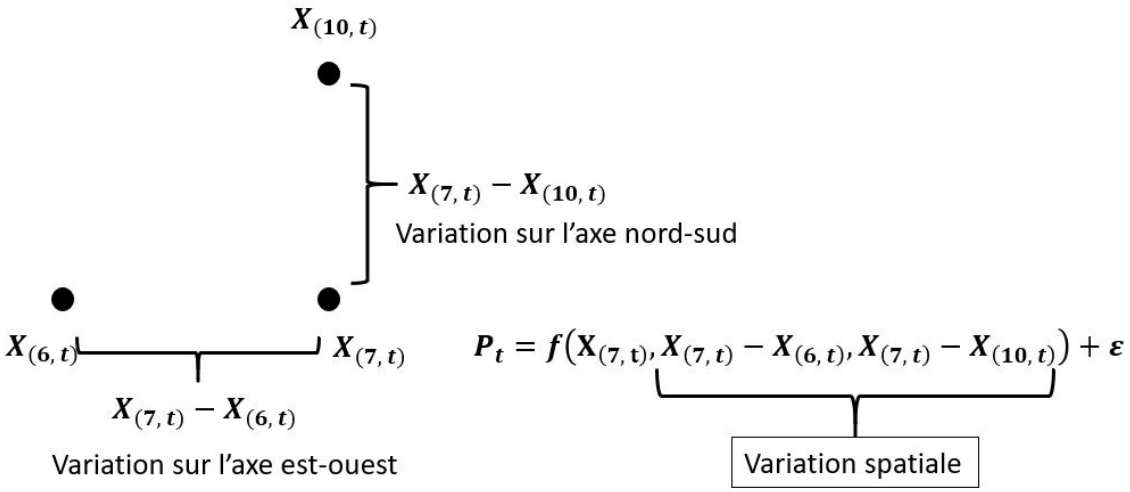


FIGURE 3.2 – Modélisation de la variation spatiale

Nous avons combiné les variations spatiales et temporelles en entrée de modèle pour étudier l'impact des variations spatio-temporelles sur la prévision de la production éolienne.

3.2.2 Résultats

Pour l'application des dérivées spatio-temporelles nous présentons les résultats des forêts aléatoires avec le modèle indirect. En effet nous avons observé les mêmes résultats avec les SVMs et Les GAMs (qui étaient les meilleurs modèles avec les forêts aléatoires dans le chapitre 2). La restriction de l'application avec l'approche indirecte se justifie par le fait qu'elle est la meilleure approche d'après les résultats du chapitre 2 et donc c'est l'approche retenue pour la modélisation. Mais les conclusions de cette partie sont extensibles à l'approche directe. On a considéré les mêmes parcs éoliens du chapitre 2. Pour chaque parc, on a comparé la modélisation avec les dérivées spatio-temporelles à la modélisation sans les dérivées spatio-temporelles (résultats du chapitre 2).

TABLE 3.1 – $NMAE$ approche indirecte avec (en rouge) et sans (en bleu) la prise en compte des dérivées spatio-temporelles

Parcs éoliens	Modèles	1 point de grille	4 points de grille	16 points de grille
Parc éolien 1	RF indirect	7.70	7.76	7.62
	Intégration des dérivées spatio-temporelles			
	RF indirect	7.45	7.57	7.44
Parc éolien 2	RF indirect	8.56	8.42	8.21
	Intégration des dérivées spatio-temporelles			
	RF indirect	7.99	7.96	7.90
Parc éolien 3	RF indirect	8.60	8.50	8.28
	Intégration des dérivées spatio-temporelles			
	RF indirect	8.29	8.18	8.10
Parc éolien 4	RF indirect	8.56	8.65	8.38
	Intégration des dérivées spatio-temporelles			
	RF indirect	8.21	8.28	8.21

Dans les quatre sites éoliens, on a obtenu une réduction des erreurs de prévision avec les dérivées spatio-temporelles dans le cas de la modélisation avec un point de grille, quatre points de grille ou seize points de grille (voir tableau 3.1). On peut observer aussi qu'avec les dérivées spatio-temporelles, en utilisant un seul point de grille (et indirectement deux autres points pour calculer les dérivées spatiales), on atteint quasiment les mêmes performances qu'avec les seize points. La réduction de l'erreur de prévision est très faible en passant d'un point de grille à seize points de grille avec l'intégration des dérivées spatio-temporelles. En définitive, empiriquement, l'intégration de dérivées

spatio-temporelles dans la prévision éolienne permet de réduire les erreurs de prévision ($NMAE$) de l'ordre de 3%. Dans la section suivante nous abordons la prévision de la production éolienne et la turbulence du vent.

3.3 Turbulence du vent et prévision éolienne

La turbulence se caractérise par un champ de vitesses dont les directions, les sens, les vitesses des particules qui le composent ne présentent aucune similarité : à un intervalle de temps au même endroit, ou bien en se décalant dans l'espace au même instant on ne retrouve pas de symétries de la vitesse d'une particule. Dans l'énergie éolienne, la turbulence est évaluée par son intensité. L'intensité de la turbulence (IT) est calculée en divisant l'écart type d'une mesure de la vitesse du vent sur des intervalles de 10 minutes par la vitesse moyenne du vent :

$$IT = \frac{\sigma_W}{\bar{W}},$$

où σ_W est l'écart type de la vitesse du vent par rapport à la vitesse moyenne du vent et $\bar{W}$ la vitesse moyenne du vent [Göçmen and Giebel, 2016]. Cependant les modèles météorologiques fournissent en général des prévisions de fréquence horaire rendant impossible le calcul de la turbulence telle que définie précédemment. Mais il existe une autre définition basée sur les vitesses du vent à différents endroits au même instant t. Dans ce cas, pour chaque point de grille, σ_W est l'écart type des vitesses du vent sur les seize points de grille et $\bar{W}$ la moyenne estimée du vent par le modèle météorologique au point de grille en question. C'est cette définition qui nous permet de calculer la turbulence en se servant du vent aux seize points de grille et de l'utiliser en entrée supplémentaire pour la prévision de la production éolienne. Nous avons appliqué le modèle indirect avec les forêts aléatoires. On s'est restreint par simplicité au parc éolien 1. Les résultats étant les mêmes sur les autres sites éoliens.

TABLE 3.2 – $NMAE$ parc éolien 1 : impact de la turbulence sur la prévision éolienne

	1 point de grille	4 points de grille	16 points de grille
RF indirect	7.70	7.76	7.62
spatio-temp	7.45	7.57	7.44
spatio-temp et turbulence	7.49	7.48	7.45
Turbulence	7.67	7.75	7.64

L'apport de la turbulence est très faible comparé à l'apport des dérivées

spatio-temporelles. De plus en combinant la turbulence et les dérivées spatio-temporelles il n'y a pas d'amélioration des prévisions de la production éolienne. En effet, il y a une forte corrélation des vitesses du vent d'un point de grille à un autre et donc l'écart type qui intervient dans la définition de l'intensité de la turbulence spatiale est faible en général (97% du temps inférieur à 0.4). De ce fait, l'impact de la turbulence issue du modèle météorologique sur la prévision éolienne est négligeable.

3.4 Sélection de points de grille ou de groupes de variables

Dans le domaine de l'apprentissage statistique il est parfois important d'identifier les co-variables les plus pertinentes pour expliquer un phénomène. Cette identification est plus connue sous le nom de sélection de variables. Les objectifs de la sélection de variables sont multiples. D'abord elle permet la réduction de la dimension et améliore la connaissance de causalité entre les variables explicatives et le phénomène (ou la variable d'intérêt) à prédire, ensuite elle permet l'interprétabilité et la reproductibilité des résultats et enfin dans certains cas elle améliore la qualité de la prévision. La sélection de variables a fait l'objet de plusieurs études. En régression linéaire, la méthode Lasso [Tibshirani, 1996] est largement utilisée. D'autres procédures de sélection de variables existent également pour les méthodes non linéaires. La mesure d'importance par permutation introduite par Breiman [Breiman, 2001] est aussi utilisée pour la sélection de variables (voir [Genuer et al., 2010], [Gregorutti et al., 2014]). Récemment la sélection de variables a été généralisée en sélection de groupe de variables dans des situations où des groupes de variables peuvent être clairement identifiés [Gregorutti et al., 2015]. D'ailleurs il a été montré qu'il est parfois intéressant de sélectionner des groupes de variables plutôt que de sélectionner des variables individuellement [He and Yu, 2010]. En effet, l'interprétation du modèle peut être améliorée ainsi que l'exactitude de la prévision en regroupant les variables en fonction d'une connaissance à priori des données. En fin de compte, le regroupement des variables peut être considéré comme une solution pour stabiliser les méthodes de sélection des variables [Gregorutti et al., 2015]. Dans ce sens, le groupe Lasso a été développé pour traiter des groupes de variables, voir par exemple [Yuan and Lin, 2006]. La sélection de variables groupées a également été proposée pour les méthodes à noyau [Zhang et al., 2008] et les réseaux de neu-

rones [Chakraborty and Pal, 2008]. Les auteurs de [Gregorutti et al., 2015] ont récemment adapté la mesure d'importance par permutation de Breiman pour les groupes de variables de manière à sélectionner des groupes de variables dans le contexte des forêts aléatoires. En se basant sur la mesure d'importance de groupe de variables, les auteurs ont proposé une méthode de sélection de variables fonctionnelles multiples. Nous utilisons cette méthode dans le cas de sélection de groupe de variables météorologiques.
Dans cette partie on s'intéresse à la sélection de groupe de variables avec les forêts aléatoires. On assimile un point de de grille à un groupe de variables. Les variables en un point de grille étant l'ensemble des variables météorologiques à ce point. L'objectif est de sélectionner les points de grille les plus pertinents pour optimiser les points de grille à prendre en considération et mieux prédire la production sur un site éolien.

3.4.1 Mesure d'importance de groupe de variables

Pour quantifier l'importance des groupes de variables, les auteurs de l'article [Gregorutti et al., 2015] ont généralisé la méthode de permutation de Breiman. Considérons un échantillon $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Comme nous l'avons expliqué dans le chapitre 2, une forêt aléatoire est l'agrégation d'arbres $(\hat{f}_1, \ldots, \hat{f}_M)$ construites à partir de M échantillons bootstrap $D_n^1, \ldots, D_n^M$ de $\mathcal{D}_n$. Pour chaque $m \in \{1, \ldots, M\}$, soit $\bar{D}_n^m := D_n \setminus D_n^m$ l'échantillon contenant les observations n'appartenant pas l'échantillon bootstrap D_n^m. Le risque de $\hat{f}_m$ dans l'échantillon $\bar{D}_n^m$ est définie par :

$$\hat{R}(\hat{f}_m, \bar{D}_n^m) = \frac{1}{|\bar{D}_n^m|} \sum_{i:(X_i,Y_i)\in\bar{D}_n^m} \left(Y_i - \hat{f}_m(X_i)\right)^2.$$

Soit $\bar{D}_n^{mj}$ la version permutée de $\bar{D}_n^m$ obtenue par permutation aléatoire de la variable X_j dans chaque échantillon $\bar{D}_n^m$. La mesure d'importance par permutation de la variable X_j est donnée par :

$$\hat{I}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{D}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{D}_n^m)\right].$$

Cette méthode a été élargie dans [Gregorutti et al., 2015] pour estimer la mesure d'importance par permutation de groupe de variables. On considère $J = (j_1, \ldots, j_k)$ un k-uplet d'indices croissants dans $\{1, \ldots, p\}$, avec $k \leq p$. Pour tout $m \in \{1, \ldots, M\}$, soit $\bar{D}_n^{mJ}$ la version permutée de $\bar{D}_n^m$ obtenue par permutation aléatoire du groupe de variables X_J dans chaque échantillon $\bar{D}_n^{mJ}$. La même permutation est utilisée pour chaque variable X_j du groupe

de variables X_J. De ce fait, la loi jointe (empirique) de X_J reste inchangée par la permutation, tandis que le lien entre X_J et la variable d'intérêt Y et les autres prédicteurs est rompu. La mesure d'importance du groupe de variables $X_J = (X_{j1}, \ldots, X_{jk})$ est définie par :

$$\hat{I}(X_J) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}\left(\hat{f}_m, \bar{D}_n^{mJ}\right) - \hat{R}\left(\hat{f}_m, \bar{D}_n^{m}\right) \right].$$

Si les groupes n'ont pas la même dimension, la version normalisée de cette mesure d'importance peut-être utilisée :

$$\hat{I}_{nor}\left(X_J\right) := \frac{1}{|J|} \hat{I}\left(X_J\right).$$

3.4.2 Algorithme de sélection de groupe de variables

La procédure de sélection est basée sur l'algorithme RFE (Recursive Feature Elimination) proposé par les auteurs de [Guyon et al., 2002] dans le contexte des support vector machines. Dans l'article [Gregorutti et al., 2015], les auteurs ont proposé une version de l'algorithme RFE avec les forêts aléatoires qui est guidée par l'importance de groupe de variables. La procédure est résumée dans l'algorithme 1. Cette approche d'élimination backward produit une collection de sous-ensembles imbriqués de groupes. Les groupes sélectionnés sont obtenus en minimisant l'erreur de validation calculée à l'étape 2 de l'algorithme.

Algorithme 1 : Recursive Feature Elimination

1 : ajuster un modèle de forêt aléatoire

2 : calculer l'erreur à l'aide d'un échantillon de validation

3 : calculer la mesure d'importance des groupes de variables

4 : éliminer le groupe de variables le moins important

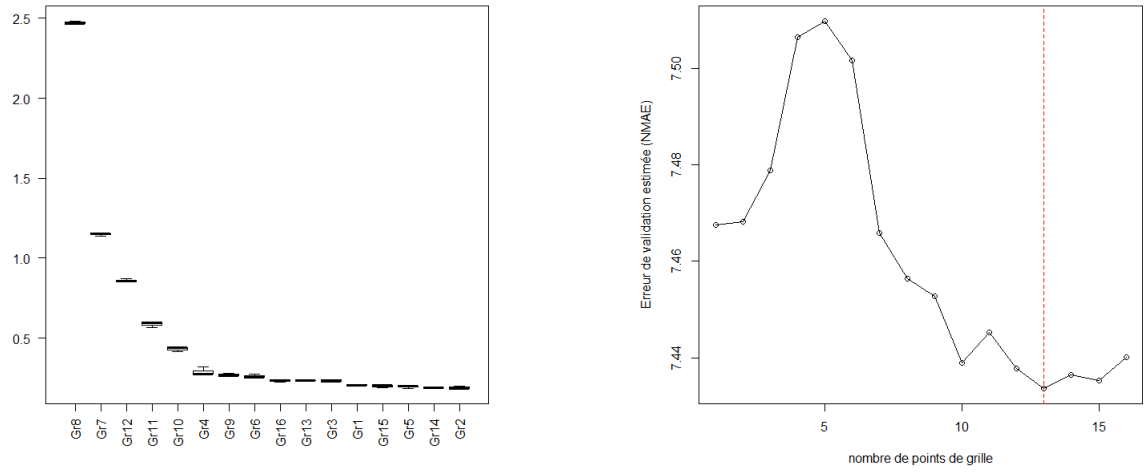
5 : répéter les étapes 1-4 jusqu'à ce qu'il ne reste plus aucun groupe

Il a également été montré dans l'article [Gregorutti et al., 2015] que, lorsque les prédicteurs sont corrélés, l'algorithme RFE offre de meilleures performances que la stratégie «non récursive» (NRFE) qui calcule l'importance des variables groupées une seule fois et ne recalcule pas l'importance à chaque étape de l'algorithme. Pour cette raison nous appliquons cette version

récursive puisque nous avons certains prédicteurs qui sont fortement corrélés (exemple les vitesses du vent, les températures d'un point de grille à un autre).

3.4.3 Résultats

Nous avons regroupé les variables météorologiques de chaque point de grille pour former les groupes. L'algorithme 1 est ensuite appliqué. Le même schéma d'apprentissage que dans le chapitre 2 est utilisé avec un échantillon d'apprentissage de deux ans et un échantillon de validation d'une année. La procédure est répétée plusieurs fois pour réduire la variabilité et le modèle final est sélectionné en minimisant l'erreur de prévision ($NMAE$). La distribution de l'importance des points de grille (figure 3.3 (a)) montre que cinq points de grille sont en général plus importants. Ces points de grille sont les points de grille 8, 7, 12, 11 et 10 qui font partie des points de grille les plus proches du parc 1 éolien (voir le maillage autour du parc, figure 2.3). A noter que le point de grille 6 qui fait partie des points les plus proches n'est pas dans le groupe des points les plus importants. C'est dans ce sens que la sélection des points de grille est utile au lieu de raisonner uniquement en termes de proximité dans le choix des points de grille pour la prévision d'un parc éolien. En observant la moyenne des $NMAE$ sur les itérations de l'algorithme 1 (figure 3.3 (b)), on constate que 13 points de grille minimisent l'erreur de prévision. Cependant à partir de dix points de grille le risque (l'erreur de prévision) a tendance à se stabiliser. Il faut aussi noter que la variation de l'erreur de prévision d'un groupe de variables (un point de grille) à seize groupes de variables (seize points de grille) n'est pas très importante. S'agissant de la fréquence de sélection des points de grille, les points les plus importants font toujours partie du groupe de points de grille qui minimisent l'erreur de validation et que les points de grille les moins importants ont des fréquences de sélection plus faibles. Mais de manière générale, chaque point de grille est sélectionné au moins une fois. La pertinence de la sélection de grille est que pour un parc éolien donné au lieu de sélectionner des points de grille par proximité, on peut choisir un ensemble de points de grille et sélectionner un sous-groupe de points de grille pour lequel les données météorologiques à ces points optimisent le modèle de prévision du parc en question.



(a) (b)

FIGURE 3.3 – Sélection de points de grille. A gauche (a) : importance des points de grille. A droite (b) : NMAE en fonction du nombre de points de grille.

3.5 Stabilité de la dynamique des prévisions

Nous nous sommes intéressés à la dynamique des prévisions d'un jour à l'autre. La question que nous nous sommes posée est la suivante : connaissant les prévisions sur les prochaines 24 heures et les prévisions sur les dernières 24 heures (voire les derniers jours), peut-on anticiper sur la stabilité ou la dynamique du vent ? Une première idée est de suivre les trajectoires du vent. Nous estimons que si le vent qui arrive sur un parc éolien en un instant t a subi moins de changement de direction (moins de variations de U et V), on peut s'attendre à une météorologie stable et donc moins d'erreurs de prévision sur les prochaines 24 heures. C'est ce qui est illustré sur la figure 3.4.

FIGURE 3.4 – Trajectoires du vent à 10m (en rouge) et à 100m (en bleu), d'une durée de 24 heures qui arrive sur un site éolien le 06/12/2017 à minuit.

Cette figure 3.4 montre la trajectoire du vent à 10m (en rouge) et la trajectoire du vent à 100m (en blue) entre le 05/12/2017 à minuit et le 06/12/2017 à minuit où le mouvement du vent est plus stable. Les erreurs de prévision enregistrées le 06/12/2017 sont relativement faibles ($NMAE \approx 7\%$). Par contre si le vent qui arrive sur le parc a subit plusieurs changements de direction (correspondant à des trajectoires du vent plus curvilignes), on est face à de multiples variations de U et V et donc une météorologie potentiellement plus instable. La figure 3.5 illustre cet analyse et correspond à la trajectoire du vent à 10m (en rouge) et à 100m (en blue) entre le 29/04/2017 à minuit le 30/04/2017 à minuit. Ce jour (30/04/2017) nous avons constaté de grosses erreurs de prévision ($NMAE \approx 15\%$).

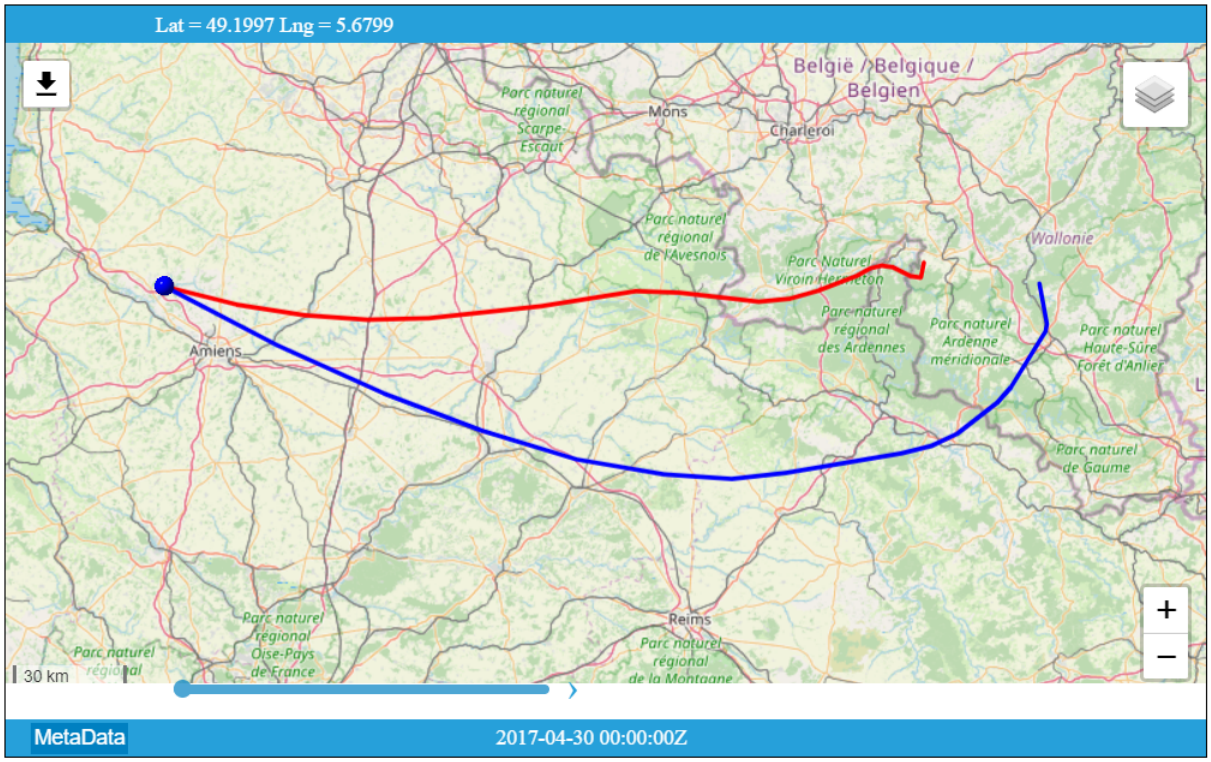


FIGURE 3.5 – Trajectoires du vent à 10m (en rouge) et à 100m (en bleu), d'une durée de 24 heures qui arrive sur un site éolien le 30/04/2017 à minuit.

La connaissance de ces trajectoires du vent pourrait permettre de suivre le direction du vent et d'anticiper sur la dynamique des prévisions et permettrait d'anticiper sur la stabilité ou non des prévisions météorologiques. Cependant, l'accès à ces données de façon industrielle n'est pas possible. Nous avons donc analysé les évolutions des prévisions. L'idée est que, si le run à un jour donné donne les prévisions météorologiques du jour suivant (notons X_2 ce tableau des prévisions) sachant que le run au jour précédent nous avait donné les prévisions au jour en question (notons X_1 ce tableau des prévisions), une forte corrélation entre ces deux tableaux de prévisions météorologiques correspond à une dynamique stable des prévisions météorologiques. Et puisqu'on sait que les variations des prévisions sont souvent sources d'erreurs, on peut s'attendre à ce que des prévisions qui suivent la même dynamique (avec une corrélation forte) correspondent à des périodes avec moins d'incertitudes. Pour répondre à ce besoin nous avons appliqué l'analyse de corrélation canonique. L'analyse de corrélation canonique (ACC) comme définie dans la section 1.4 est une méthode proposée par [Hotelling, 1936] pour étudier les relations pouvant exister entre deux groupes de variables (deux tableaux de données). Elle a pour intérêt de mesurer et caractériser les corrélations entre deux

groupes de variables mesurées sur les mêmes individus. Nous cherchons à caractériser les corrélations entre les deux tableaux de prévision de deux jours successifs en calculant à chaque fois la somme des valeurs propres comme dans la section 1.4.2. Nous considérons les prévisions météorologiques au point de grille le plus proche du parc éolien 1 avec 11 variables (les composantes du vent (U, V) à 10m et à 100m, les directions du vent à 10m et à 100m, les vitesses du vent à 10m et à 100m, la température à 2m, la pression à la surface et les précipitations totales à la surface) sur l'année 2017. Si pour un jour donné, la somme des valeurs propres découlant de l'analyse canonique est égale à 11, on a une corrélation parfaite entre ce jour et le jour précédent. Par contre une faible somme des valeurs propres va correspondre à une corrélation faible des prévisions entre ces deux jours et donc à une dynamique différente des prévisions.

Sur l'année 2017, la sommes des valeurs propres découlant de l'analyse de corrélation canonique entre les prévisions de deux jours successifs varie entre 6.5 et 9.8. On a une distribution uni-modale. Une distribution bi-modale pourrait correspondre aux jours où les prévisions sont stables et aux autres jours où elles ne le sont pas. Nous allons donc analyser les jours avec une faible somme des valeurs propres et les jours avec une forte somme des valeurs propres. Pour mieux interpréter la somme des valeurs propres par jour, nous avons calculé pour chaque jour, le NMAE des prévisions de production (on est sur une période de test dont on dispose de prévisions de la production) et la moyenne journalière des vitesses du vent. Nous nous sommes intéressés d'abord aux jours avec une faible somme des valeurs propres (moins de corrélation avec le jour précédent). Nous avons pris comme valeur critique le quantile 10%. Ainsi, tous les jours avec une somme des valeurs propres inférieure à cette valeur sont considérés comme étant moins corrélés avec le jour précédent. La figure 3.6 montre le NMAE ($NMAEParJour$) et la moyenne des vitesses du vent ($MoyenneDuVent$) pour ces jours de corrélation faible. On peut noter que lorsqu'il y a un vent relativement fort ($\geq 8m/s$) et que la somme des valeurs propres est faible, on a de grosses erreurs de prévision dues à une météorologie instable. Par contre, une somme des valeurs propres avec un vent faible ne correspond pas forcément à de grosses erreurs de prévision. En effet, si le vent est faible, que le météorologie se trompe ou pas, on a pas une répercussion sur les prévisions de production (avec un vent de moins de 5 m/s environ la production est toujours quasiment nulle).

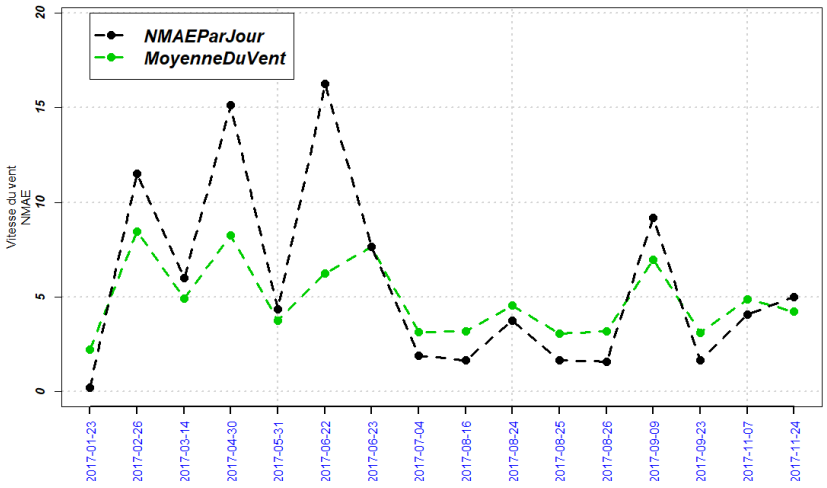


FIGURE 3.6 – NMAE ($NMAEParJour$) et la moyenne des vitesses du vent ($MoyenneDuVent$) pour les jours à faible corrélation canonique.

La deuxième partie consiste à analyser les jours où la somme des valeurs propres est élevée (supérieure au quantile 90%). On observe (figure 3.7) qu'en général, lorsque la somme des valeurs propres est élevée (correspondant aux jours avec des prévisions stables), les erreurs de prévision de la production sont plus stables et relativement plus faibles.

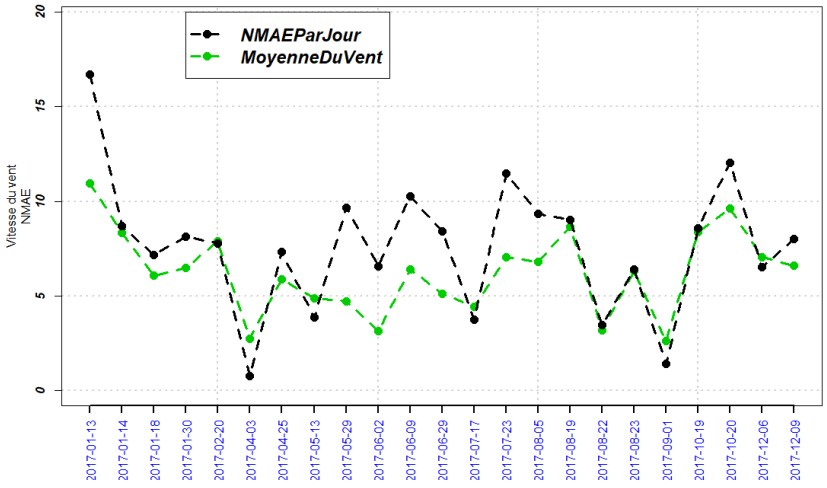


FIGURE 3.7 – NMAE ($NMAEParJour$) et la moyenne des vitesses du vent ($MoyenneDuVent$) pour les jours à forte corrélation canonique.

Cependant, il y a certains jours où la somme des valeurs propres est élevée et que les erreurs de prévision sont assez grandes. Nous estimons qu'il s'agit de périodes où les prévisions sont stables mais qu'elles ne reflètent pas

la réalité. Ces résultats empiriques montrent que l'analyse canonique peut permettre d'anticiper sur la dynamique des prévisions météorologiques. En perspective, une généralisation de l'analyse canonique entre plusieurs points de grille différents pourrait permettre de mieux comprendre la dynamique du vent si on parvenait à savoir quels points choisir ainsi que la distance entre les points de grille.

3.6 Conclusion

Ce chapitre a été l'objet d'une étude empirique de l'apport des variations spatio-temporelles sur la prévision du productible de parcs éoliens. Nous avons montré que la modélisation des variations spatio-temporelles permet de réduire les erreurs de prévision. L'étude de la turbulence du vent (accessible à travers les modèles météorologiques) quant à elle montre que la turbulence n'est pas importante dans la prévision éolienne. En effet l'intensité de turbulence est très faible à cause d'une forte corrélation des vitesses du vent d'un point de grille à un autre. Nous avons aussi étudié la sélection des points de grille (sélection de groupe de variables) permettant de regrouper les points géographiques du maillage du modèle météorologique les plus pertinents pour expliquer la production d'un parc éolien.
Dans le chapitre suivant, nous avons étudié l'intervalle de prévision pour les prévisions d'énergie éolienne avec les forêts aléatoires.

Chapitre 4

Incertitudes liées à la prévision de la production

Sommaire

4.1 Introduction

Au cours de ces dernières années, la production d'énergie éolienne et solaire s'est accrue au niveau mondial et on s'attend à ce qu'un pourcentage important de la production totale d'énergie provienne de ces sources d'énergie. Cependant, elles présentent une variabilité inhérente qui implique des fluctuations de la production d'énergie. Cette variabilité est souvent source d'erreurs et d'incertitudes pouvant engendrer des pénalités financières lors de la vente d'énergie sur le marché. Ainsi, les erreurs de prévision jouent un rôle considérable dans les impacts et les coûts de l'intégration, de la gestion et de

la commercialisation des énergies renouvelables.
La problématique est que nous avons un modèle indirect en deux étapes : un modèle de prévision du vent avec les forêts aléatoires et un modèle de courbe de puissance réelle permettant de passer le vent prévu en prévision de production (figure 4.1).

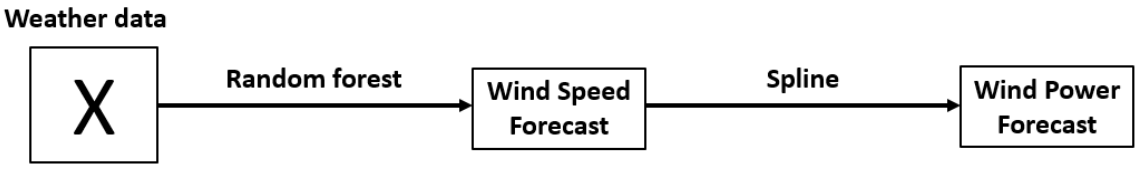


FIGURE 4.1 – Modèle de prévision indirect de la production

Nous avons deux sources d'incertitudes : l'incertitude liée à la prévision du vent qui découle de la forêt aléatoire et des données météorologiques et l'incertitude liée à la courbe de puissance estimée par une spline. Nous nous sommes appuyés sur ces deux incertitudes pour construire l'intervalle de prévision associé aux prévisions de production éolienne dans le cadre d'un modèle indirect. Ces travaux ont été présenté à la conférence "13th International Conference on CFE-CMStatistics 2019".

L'analyse des erreurs de prévision est un aspect important pour évaluer les performances d'un modèle de prévision. Une bonne connaissance des erreurs de prévision peut aider à améliorer la qualité d'un modèle de prévision afin de réduire les risques sur la vente d'énergie. Les mesures d'erreurs classiques ($NMAE$ et $NRMSE$) ne prennent pas en compte le décalage horizontal ou la distance temporelle entre les prévisions et les observations. C'est pourquoi, nous avons également étudié les erreurs temporelles de la prévision éolienne dans ce chapitre.

Les méthodes de construction d'intervalles prévision sont étudiées dans la section 4.2. La section 4.3 explore l'estimation des incertitudes sur la courbe de puissance. La procédure de construction d'intervalle de prévision est développée dans la section 4.4. Les résultats de la construction d'intervalle de prévision sont discutés dans la section 4.5. Les erreurs de phase sont étudiées dans la section 4.6 et la section 4.7 conclut ce chapitre.

4.2 Intervalle de prévision avec les forêts aléatoires

L'estimation ponctuelle d'une variable aléatoire continue fournit en général une valeur unique. Il est indispensable d'estimer la variabilité de l'erreur associée à l'estimation ponctuelle. Cette variabilité peut être utilisée pour construire un intervalle de prévision. Nous allons présenter trois méthodes de construction d'intervalles de prévision utilisant les forêts aléatoires.

4.2.1 Forêts de régression quantile

Les forêts de régression quantile ont été introduites par [Meinshausen, 2006] pour estimer la fonction de distribution cumulée $F_Y(y/\mathbf{X} = \mathbf{x}_h) = P(Y \leqslant y/\mathbf{X} = \mathbf{x}_h)$ d'une variable d'intérêt Y sachant une valeur $\mathbf{x}_h$ des variables d'entrée $\mathbf{X}$ et en déduire des quantiles qui permettent d'obtenir un intervalle de prévision. La régression quantile est donc une méthode qui peut être directement utilisée pour obtenir des intervalles de prévision. L'algorithme donne les détails de l'estimation de la fonction de distribution cumulative de la variable Y. Soit $\hat{F}$ l'estimation de cette fonction. L'algorithme proposé par [Meinshausen, 2006] identifie d'abord chaque nœud terminal de l'arbre de décision contenant une nouvelle observation $\mathbf{x}_h$. Pour chaque observation utilisée dans la construction de la forêt aléatoire φ, l' algorithme parcourt chaque nœud terminal identifié dans l'étape précédente et calcule la fréquence avec laquelle l'observation apparaît au moins une fois dans le nœud terminal, où la fréquence est évaluée par rapport au nombre total d'observations qui apparaissent au moins une fois dans le nœud terminal (Equation 4.2). Pour chaque observation, l'algorithme calcule une moyenne des fréquences sur tous les arbres de décision de la forêt aléatoire (Equation 4.3) et utilise le poids résultant pour construire une fonction de distribution cumulative conditionnelle empirique (Equation 4.4).

[Meinshausen, 2006] a montré que sous certaines hypothèses (voir l'article pour plus de détails), la fonction de distribution cumulative conditionnelle empirique $\hat{F}_Y(y/\mathbf{X} = \mathbf{x}_h)$ obtenue converge en probabilité vers la vraie fonction de distribution cumulative conditionnelle $F_Y(y/\mathbf{X} = \mathbf{x}_h)$ si la taille de l'échantillon n tend vers l'infini. Si la fonction de distribution cumulative conditionnelle empirique tend vers la vraie fonction, alors un intervalle de prévision $(1-\alpha)100\%$ peut être obtenu en sous paramétrant le domaine de la fonction de distribution cumulative conditionnelle empirique aux quantiles appropriés.

L'intervalle $[\hat{Q}_{\alpha_1}, \hat{Q}_{\alpha_2}]$ où

$$\hat{Q}_{\alpha_i} = inf\{y : F_Y(y/\mathbf{X} = \mathbf{x}_h) \geq \alpha_i\} \tag{4.1}$$

est l'intervalle de prévision $(1-\alpha)100\%$ si $\alpha_2 - \alpha_1 = \alpha$.
Dans la suite on note $\mathbf{Z} = (\mathbf{X}, Y)$ l'échantillon d'observations et $\mathbf{z}_h = (\mathbf{x}_h, y_h)$ une observation de $\mathbf{Z}$. L'algorithme des forêts de régression quantile est le suivant :

Algorithme 1 : Algorithme de la régression quantile par forêt aléatoire

procedure ESTIMATECDF (sample $\mathbf{Z}$, random forest φ built on $\mathbf{Z}$, test observation $\mathbf{z}_h$)
    let n denote the sample size of $\mathbf{Z}$
    let B denote the number of trees in φ
    let $v_b(\mathbf{x})$ be the index of the b^{th} tree's terminal node that contains x
    for i in $1, \ldots, n$ **do**
        for b in $1, \ldots, B$ **do**
            let $w_i^b(\mathbf{x}_h)$ be defined by

$$w_i^b(\mathbf{x}_h) = \frac{\mathbb{1}(v_b(\mathbf{x}_i) = v_b(\mathbf{x}_h))}{|\{j : v_b(\mathbf{x}_j) = v_b(\mathbf{x}_h)\}|} \tag{4.2}$$

        end for
        let $w_i(\mathbf{x}_h)$ be defined by

$$w_i(\mathbf{x}_h) = \frac{1}{B}\sum_{b=1}^{B} w_i^b(\mathbf{x}_h) \tag{4.3}$$

    end for
    compute the estimated conditional cumulative distribution function $\hat{F}_Y(y/\mathbf{X} = \mathbf{x_h})$ by

$$\hat{F}_Y(y/\mathbf{X} = \mathbf{x_h}) = \sum_{i=1}^{n} w_i(\mathbf{x}_h)\mathbb{1}(Y_i \leq y) \tag{4.4}$$

    return $\hat{F}_Y(y/\mathbf{X} = \mathbf{x_h})$
end procedure.

Les forêts de régression quantile permettent d'obtenir directement des intervalles de prévision. Les tests empiriques de [Meinshausen, 2006] ont montré que des intervalles de prévision $(1-\alpha)100\%$ qui contiennent la vraie valeur d'intérêt $(1-\alpha)100\%$ du temps peuvent être obtenus.

4.2.2 Estimation de la variance des forêts aléatoires

Considérons $\hat{\Theta}_B^{RF}$ un prédicteur de forêt aléatoire construit à partir de B échantillons bootstrap. On cherche à trouver un estimateur de l'écart-type (ou de la variance) de la forêt aléatoire,

$$se\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) \equiv \sqrt{Var\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)},$$

où $\hat{\Theta}_B^{RF}(\mathbf{x}_h)$ est la prévision de la forêt aléatoire sachant les valeurs des variables explicatives $\mathbf{x}_h$. $\hat{\Theta}_B^{RF}(\mathbf{x}_h)$ n'est rien d'autre que $\hat{y}_h$. Dans les paragraphes suivants nous passons en revue des méthodes d'estimation de $se\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)$.

[Sexton and Laake, 2009] ont proposé trois méthodes d'estimation de l'écart-type $se\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)$ par bootstrap. Soit $\hat{\Theta}_B^{RF}(\mathbf{x}_h)$ la prévision de la variable réponse de l'observation $\mathbf{x}_h$ par une forêt aléatoire ajustée sur un échantillon $\mathbf{Z} = (\mathbf{X}, Y)$ de taille n. Considérons $\mathbf{Z}^m$, $m = 1, \ldots, M$, M échantillons bootstrap issus de $\mathbf{Z}$. La prévision de la réponse y_h sur le m^{iem} échantillon bootstrap est :

$$\hat{\Theta}_B^{RF,m}(\mathbf{x}_h) = \frac{1}{B}\sum_{b=1}^{B} T_b^m(\mathbf{x}_h),$$

où T_b^m est un arbre de décision construit sur le b^{iem} échantillon bootstrap du m^{iem} échantillon bootstrap de $\mathbf{Z}$. Le premier estimateur de l'écart-type de la forêt aléatoire au point $\mathbf{x}_h$ proposé par [Sexton and Laake, 2009] est :

$$\hat{se}^{BF}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) = \left(\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\Theta}_B^{RF,m}(\mathbf{x}_h) - \bar{\hat{\Theta}}_B^{RF}(\mathbf{x}_h)\right)^2\right)^{1/2},$$

où

$$\bar{\hat{\Theta}}_B^{RF}(\mathbf{x}_h) = \frac{1}{M}\sum_{m=1}^{M}\hat{\Theta}_B^{RF,m}(\mathbf{x}_h)$$

est la moyenne des prévisions des forêts aléatoires de la réponse y_h sur les M échantillons bootstrap.
Cet estimateur est appelé « Brute Force estimator » en raison du nombre important d'arbres de décision construits dans la procédure ($M \times B$ arbres de décision). Le coût de calcul de cette procédure est énorme. C'est pourquoi, [Sexton and Laake, 2009] l'ont utilisé comme base de référence avec laquelle ils ont comparé les deux autres méthodes suivantes.

La deuxième méthode proposée dans [Sexton and Laake, 2009] est une méthode similaire à la première appelée l'estimateur « Biaised Bootstrap ». La

seule différence est que le « Biaised Bootstrap » ajuste une petite forêt aléatoire construite sur $R < B$ arbres à chacun des M échantillons bootstrap.
L'estimateur « Biaised Bootstrap » est défini par :

$$\hat{se}^{BB}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) = \left(\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\Theta}_R^{RF,m}(\mathbf{x}_h) - \bar{\hat{\Theta}}_R^{RF}(\mathbf{x}_h)\right)^2\right)^{1/2},$$

où

$$\bar{\hat{\Theta}}_R^{RF}(\mathbf{x}_h) = \frac{1}{M}\sum_{m=1}^{M}\hat{\Theta}_R^{RF,m}(\mathbf{x}_h).$$

C'est un estimateur biaisé de l'écart-type.

La troisième méthode proposée par [Sexton and Laake, 2009] est un estimateur corrigeant le biais de l'estimateur « Biaised Bootstrap » appelé « Noisy Bootstrap estimator » mais nécessite toujours deux niveaux de bootstrapping. L'estimateur « Noisy Bootstrap estimator » de l'écart-type de la forêt aléatoire à $\mathbf{x}_h$ est :

$$\hat{se}^{NB}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) = \left(\hat{Var}^{BB}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) - \hat{bias}\left(\hat{Var}^{BB}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)\right)\right)^{1/2},$$

où

$$\hat{Var}^{BB}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) \equiv \left(\hat{se}^{BB}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)\right)^2$$

et

$$\hat{bias}\left(\hat{Var}^{BB}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)\right) = \frac{1/R - 1/B}{MR(R-1)}\sum_{i=1}^{M}\sum_{r=1}^{R}\left(T_r^m(\mathbf{x}_h) - \hat{\Theta}_R^{RF,m}(\mathbf{x}_h)\right)^2.$$

[Wager et al., 2014] ont proposé d'estimer $se\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)$ par des procédures de jackknife. Ils ont défini deux estimateurs : l'estimateur Jackknife-après-Bootstrap et l'estimateur Infinitesimal Jackknife ([Efron, 1992] et [Efron, 2013]). Dans les deux cas, des versions corrigées des biais sont identifiées. On considère toujours $\hat{\Theta}_B^{RF}(\mathbf{x}_h)$ comme la prévision d'une observation $\mathbf{x}_h$ par une forêt aléatoire ajustée sur un échantillon $\mathbf{Z}$ de taille n. L'estimateur Jackknife-après-Bootstrap de l'écart-type de la forêt aléatoire au point $\mathbf{x}_h$ est :

$$\hat{se}^{J}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) = \left(\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\Theta}_{(-i)}^{RF}(\mathbf{x}_h) - \hat{\Theta}^{RF}(\mathbf{x}_h)\right)^2\right)^{1/2}$$

où

$$\hat{\Theta}_{(-i)}^{RF}(\mathbf{x}_h) = \frac{1}{|\{b : \langle \mathbf{z}_i\rangle_b = 0\}|}\sum_{b:\langle \mathbf{z}_i\rangle_b = 0} T_b(\mathbf{x}_h) \tag{4.5}$$

avec $\langle \mathbf{z} \rangle_b$ le nombre de fois où $\mathbf{z}$ apparaît dans le b^{iem} échantillon bootstrap de la forêt aléatoire. Ainsi, l'équation 4.5 donne la réponse moyenne prévue de $\mathbf{y}_h$ sur seulement les arbres dont la i^{iem} observation de $\mathbf{Z}$ est « out-of-bag » (en dehors de l'échantillon).
L'estimateur Jackknife-après-Bootstrap avec correction du biais est :

$$\hat{se}^{J-U}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) = \left[\left(\hat{se}^{J}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)\right)^2 - (e-1)\frac{n}{B^2}\sum_{b=1}^{B}\left(T_b(\mathbf{x}_h) - \bar{T}(\mathbf{x}_h)\right)^2\right]^{1/2}$$

où $\bar{T}$ est la moyenne des $T_b(\mathbf{x}_h)$.

L'estimateur Infinitesimal Jackknife quant à lui est défini par :

$$\hat{se}^{IJ}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) = \left[\sum_{i=1}^{n}\left(\frac{1}{B}\sum_{b=1}^{B}\left(\langle \mathbf{z}_i \rangle_b - 1\right)\left(T_b(\mathbf{x}_h) - \bar{T}(\mathbf{x}_h)\right)\right)^2\right]^{1/2}.$$

Et enfin [Wager et al., 2014] ont défini une version de l'estimateur Infinitesimal Jackknife avec un biais corrigé. Cette version avec un biais corrigé est donnée par :

$$\hat{se}^{IJ-U}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right) = \left[\left(\hat{se}^{IJ}\left(\hat{\Theta}_B^{RF}(\mathbf{x}_h)\right)\right)^2 - \frac{n}{B^2}\sum_{b=1}^{B}\left(T_b(\mathbf{x}_h) - \bar{T}(\mathbf{x}_h)\right)^2\right]^{1/2}.$$

Ils proposent qu'une fois l'écart-type de la forêt aléatoire $se\left(\hat{\Theta}_B^{RF}(\mathbf{x_h})\right)$ estimé, de prendre l'intervalle $\hat{\Theta}_B^{RF}(\mathbf{x_h}) \pm Q_\alpha * se\left(\hat{\Theta}_B^{RF}(\mathbf{x_h})\right)$, où Q_α est le quantile $1-\alpha$ d'une distribution normale.

Les méthodes proposées par [Sexton and Laake, 2009] et [Wager et al., 2014] cherchent en réalité à estimer comment les prévisions de la réponse des observations avec un certain ensemble de prédicteurs varient d'une forêt aléatoire à une autre. Elles sont coûteuses en calcul et exigent deux niveaux de bootstrapping. De plus, l'écart type des prévisions de la forêt aléatoire n'est pas la mesure de variabilité nécessaire pour construire un intervalle de prévision. Par exemple dans le cas où les prévisions des forêts aléatoires sont biaisées (c'est-à-dire que la moyenne des prévisions est différentes de la moyenne des observations) ou lorsque les variations des prévisions sont différentes des variations des observations (exemple les prévisions éoliennes ont tendance à être plus lisses que les observations), ces estimateurs ne peuvent pas servir à construire un bon intervalle de prévision. Une alternative consiste à estimer la distribution des erreurs de prévision.

4.2.3 Estimation de la distribution des erreurs de prévision

Les méthodes récentes de construction d'intervalle de prévision avec les forêts aléatoires s'appuient sur la distribution des erreurs de prévision. On cherche à estimer l'erreur quadratique moyenne ou les quantiles de la distribution des erreurs pour construire un intervalle de prévision. Ainsi, [Lu and Hardin, 2017] ont proposé deux estimateurs de l'erreur quadratique moyenne de prévision (Mean Squared Prediction Error) : $MSPE1$ et $MSPE2$. La notation $\hat{\theta}_B^{RF}(\mathbf{x}_h)$ comme la prévision aléatoire de $\mathbf{x}_h$ par la forêt est abandonnée dans les algorithmes qui suivent. En effet, on passe d'une prévision de la forêt aléatoire comme une statistique avec une variabilité d'échantillonnage d'intérêt à une prévision de la forêt aléatoire comme une prévision de la réponse d'une observation particulière.

Le $MSPE1$ calcule directement l'erreur quadratique moyenne de prévision de la forêt aléatoire. Pour chaque observation, il identifie les arbres de décision pour lesquels l'observation est en dehors de l'échantillon bootstrap (out-of-bag en anglais) puis calcule la prévision moyenne de la réponse de l'observation en utilisant uniquement ces arbres de décision (équation 4.6). L'estimation de l'erreur quadratique moyenne de la prévision est alors la somme de la différence quadratique entre la vraie réponse de chaque observation et sa prévision moyenne en utilisant seulement les arbres de décision pour lesquels l'observation est out-of-bag (équation 4.7). Intuitivement si on veut estimer directement l'erreur quadratique moyenne de la prévision de forêt aléatoire, on peut utiliser de nombreuses observations n'ayant pas servi à la construction de la forêt. Par contre en pratique, on utilise habituellement toutes les observations pour construire la forêt, donc une alternative est de traiter de façon itérative chacune des observations comme dans une procédure semblable à une validation croisée. Plus concrètement, pour chaque observation $(\mathbf{X_i}, Y_i)$, on pourrait l'omettre de l'échantillon, construire une forêt sur l'échantillon sans l'observation et obtenir la prévision de la réponse de l'observation omise. Ceci suppose la construction d'une forêt aléatoire pour chaque observation. Cependant l'algorithme des forêts aléatoires permet d'effectuer la procédure sans construction de forêt supplémentaire. En effet les arbres de décision dans la forêt aléatoire sont construits sur des échantillons bootstrap qui ne contiennent pas certaines observations. S'il y a un grand nombre d'arbres, alors chaque observation sera out-of-bag pour beaucoup d'entre eux. C'est d'ailleurs [Zhang et al., 2019] qui ont montré qu'à partir de l'étape de ré-échantillonnage de l'algorithme de forêt aléatoire,

approximativement $(\frac{n-1}{n})^n \approx exp(-1)$ des B arbres de la forêt aléatoire sont construits sans l'observation $(\mathbf{X_i}, Y_i)$. Donc pour tout $i = 1, \ldots, n$, on peut construire une sous-forêt $RF_{(i)}$ de la forêt originale avec approximativement $B.exp(-1)$ arbres sans l'observation $(\mathbf{X_i}, Y_i)$. Une telle forêt aléatoire est facilement disponible pour chaque observation comme un sous ensemble d'arbre de la forêt originale. Cette procédure modifiée permet d'estimer $MSPE1$.

Algorithme 2 : Mean Squared Prediction Error 1

1 : **procedure** GetMSPE1 (sample **Z**, random forest φ built on **Z**)
2 : let n denote the sample size of **Z**
3 : let B denote the number of trees in φ
4 : let $\langle \mathbf{z} \rangle_b$ be the number of times **z** appears in the b^{th} bootstrap sample

5 : # for each observation
6 : **for** i in $1, \ldots, n$ **do**

7 : # identify the trees for which it is out of bag
8 : let $S = \{b \in \{1, \ldots, B\} : \langle \mathbf{z}_i \rangle_b = 0\}$

9 : # get its mean predicted response over those trees
10 : let $\hat{y}_{(-i)}$ be defined by

$$\hat{y}_{(-i)} = \frac{1}{|S|} \sum_{b \in S} T_b(\mathbf{x}_i) \tag{4.6}$$

11 : **end for**
12 : **return**

$$MSPE1 = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{(-i)} - y_i)^2 \tag{4.7}$$

13 : **end procedure**

Le $MSPE1$ calcule l'erreur quadratique de prévision une fois pour chaque observation d'entraînement. Le $MSPE2$ calcule l'erreur quadratique de prévision uniquement pour les cohabitantes out-of-bag de la nouvelle observation dans chaque arbre. Il permet les répétitions, de sorte que si une observation est out-of-bag cohabitante d'une nouvelle observation dans m arbres, son erreur de prévision au carré sera comptée m fois (voir [Lu and Hardin, 2017]).

Algorithme 3 : Mean Squared Prediction Error 2

procedure GetMSPE2 (sample Z, random forest φ built on Z)
    let n denote the sample size of Z
    let B denote the number of trees in φ
    let $\langle \mathbf{z} \rangle_b$ be the number of times $\mathbf{z}$ appears in the b^{th} bootstrap sample
    let $v_b(\mathbf{x})$ be the index of the b^{th} tree's terminal node that contains $\mathbf{x}$

    # for each decision tree
    for i in $1, \ldots, B$ **do**

        # get the sample observations that are out-of-bag cohabitants of the new observation
        let $S_b = \{i \in \{1, \ldots, n\} : \langle \mathbf{z_i} \rangle_b = 0 \wedge v_b(\mathbf{x}_i) = v_b(\mathbf{x}_h)\}$

        # for each out-of-bag cohabitant
        for i in S_b **do**

            # identify the trees for which it is out of bag
            let $Q_i = \{b \in \{1, \ldots, B\} : \langle \mathbf{z}_i \rangle_b = 0\}$

            # get its mean predicted response over those trees
            let $\hat{y}_{(-i)}$ be defined by

$$\hat{y}_{(-i)} = \frac{1}{|Q_i|} \sum_{b \in Q_i} T_b(\mathbf{x}_i) \tag{4.8}$$

        end for
    end for

    # count how many times each observation is an out-of-bag cohabitant of the test observation
    let c_i be the total number of times i appears in $S_1, \ldots, S_B$
    return

$$MSPE2 = \frac{1}{\sum_{i=1}^{n} c_i} \sum_{i:c_i>0} (\hat{y}_{(-i)} - y_i)^2 \tag{4.9}$$

end procedure

Par simulations empiriques quelques propriétés des deux estimateurs sont données dans [Lu and Hardin, 2017]. La restriction du $MSPE2$ aux cohabi-

tantes de l'observation out-of-bag garantit qu'il mesure l'erreur quadratique de prévision des seules observations qui sont proches de la nouvelle observation. En effet, les nœuds des arbres de décision sont construits en tenant compte des observations les unes des autres ainsi que des différences entre leurs réponses. Cette formulation repose sur l'idée que les observations proches d'une nouvelle observation dans l'espace prédicteur sont tirées de la population d'intérêt ; l'erreur de prévision des observations tirées de cette population devrait ressembler étroitement à l'erreur de prévision de la nouvelle observation. Ceci rend probablement préférable le $MSPE2$ lorsqu'il y a une variation importante mais systématique de la variance des valeurs de réponse des observations par rapport à l'espace prédicteur ([Lu and Hardin, 2017]). Cependant le $MSPE2$ peut ignorer certaines observations de l'échantillon par ce qu'elles ne sont pas des cohabitantes de la nouvelle observation dans l'arbre de décision de la forêt aléatoire alors que le $MSPE1$ inclut toutes les observations dans son calcul, de sorte qu'on s'attend à ce qu'elle soit plus stable que le $MSPE2$. Ceci rend probablement préférable le $MSPE1$ lorsque la distribution des réponses présente une variance constante dans l'espace prédicteur. Dans un tel contexte, toutes les observations sont également utiles pour estimer l'erreur quadratique moyenne de prévision d'une nouvelle observation. Le $MSPE1$ en tire pleinement parti en incluant toutes les observations dans son calcul, alors que le $MSPE2$ n'inclut que les cohabitantes out-of-bag de la nouvelle observation, ignorant ainsi les données utiles. Ainsi, chaque estimateur a ses avantages et ses inconvénients.

Si ϕ est une forêt aléatoire et $\mathbf{x}_h$ une observation dont la réponse y_h est inconnue. Un intervalle de prévision à $(1-\alpha)100\%$ de confiance pour y_h est donné par

$$\phi(\mathbf{x}_h) \pm Q_\alpha * \sqrt{MSPE1}$$

et par

$$\phi(\mathbf{x}_h) \pm Q_\alpha * \sqrt{MSPE2},$$

où Q_α est le quantile $1-\alpha$ d'une distribution normale.

Ces algorithmes permettent d'obtenir l'intervalle de prévision des vitesses du vent. Il faut passer maintenant à l'intervalle de prévision de la courbe de puissance réelle estimée par une spline.

4.3 Incertitudes sur la courbe de puissance réelle

Nous avons un modèle de forêt aléatoire pour la prévision du vent, ensuite la courbe de puissance réelle (estimée par une spline en utilisant les vitesses de vent réel et les productions réelles) permet de convertir les prévisions de vitesse du vent en prévisions de production. Une première manière d'obtenir l'intervalle de prévision de la production pourrait être d'utiliser la courbe de puissance réelle estimée pour convertir les bornes de l'intervalle de prévision des vitesses du vent en intervalle de prévision de la production (voir figure 4.2). Cependant on n'intégrera pas l'incertitude sur la courbe de puissance. En effet, dans la réalité, à une vitesse de vent donnée correspondent plusieurs productions. D'où la nécessité d'estimer l'incertitude liée à la courbe de puissance.

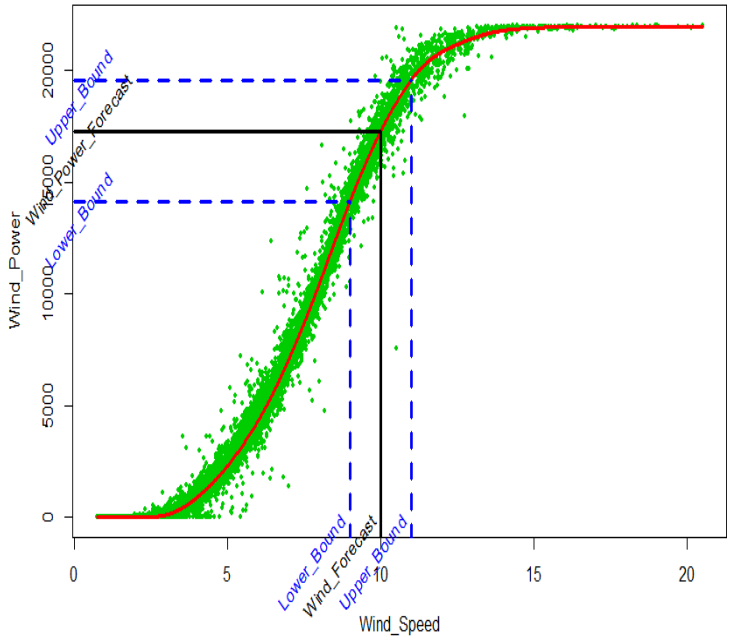


FIGURE 4.2 – Passage direct de l'intervalle de prévision du vent à l'intervalle de prévision de la production

Pour estimer l'incertitude sur la courbe de puissance, nous avons d'abord fait une 10-validation croisée de la spline sur l'échantillon d'apprentissage de deux ans pour obtenir la production estimée par la spline sur la période d'apprentissage. Notons $\widehat{P}_{spline}$ cette production estimée. On cherche ensuite à estimer l'écart type des erreurs de prévision $\widehat{\sigma}_{spline}$ liées à la spline. On sait

que cet écart type varie en fonction de x. Par exemple pour les vitesses de vent inférieures à $4\,m/s$ ou supérieures $14\,m/s$, il y a moins d'incertitudes sur la spline et donc $\widehat{\sigma}_{spline}$ est plus faible. Pour chaque x, on considère l'intervalle $I_x = [x - \frac{1}{2}; x + \frac{1}{2}]$. L'écart type des erreurs de la spline est estimée par :

$$\widehat{\sigma}_{spline}(x) = \sqrt{\frac{\sum_{i=1}^{n_x}(e_i - \bar{e_x})^2}{n_x}},$$

où, $e_i = P(x_i) - \widehat{P}_{spline}(x_i)$ avec $x_i \in I_x$ est l'erreur d'estimation de la spline, P la production réelle, $\widehat{P}_{spline}$ la production estimée par la spline, $\bar{e_x}$ est la moyenne des erreurs sur l'intervalle I_x et n_x le nombre d'observation dans l'intervalle I_x. L'intervalle de prévision de la spline est défini par :

$$\hat{P}_{spline} \pm Q_\alpha * \widehat{\sigma}^k_{spline}.$$

Ce qui nous donne les deux bornes délimitant la dispersion au tour de la courbe de puissance estimée avec $\alpha = 0.2$, figure 4.3.

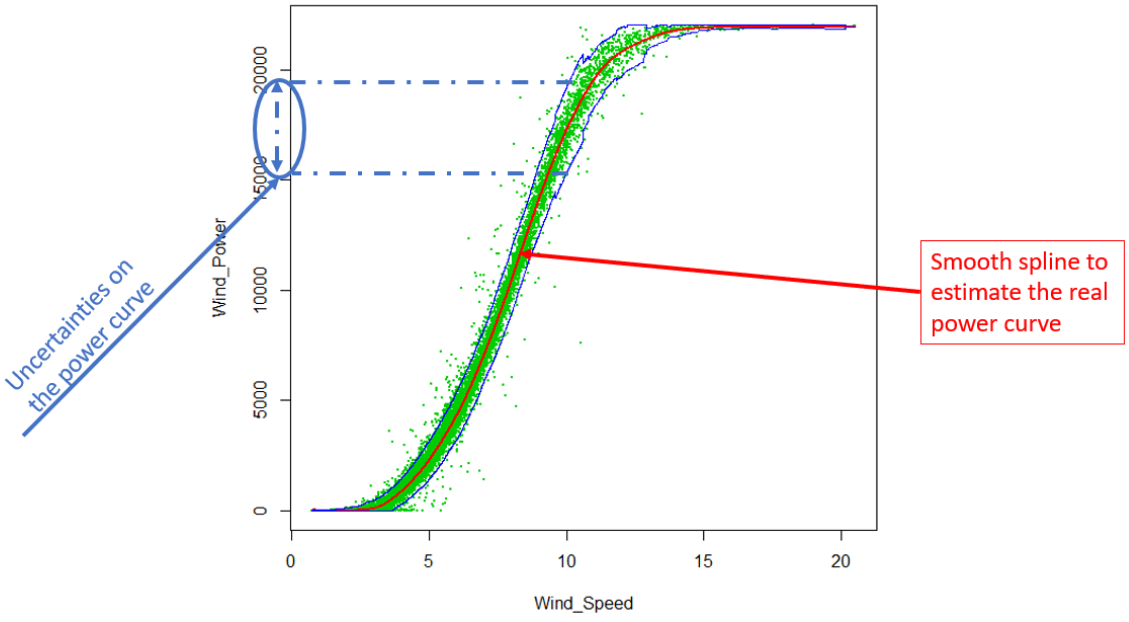


FIGURE 4.3 – Incertitudes liées à la courbe de puissance réelle

4.4 Combinaison incertitudes des forêts aléatoires et incertitudes sur la courbe de puissance

Nous disposons maintenant de l'intervalle de prévision de la forêt aléatoire des vitesses du vent. Si nous avons une vitesse de vent prévue par la forêt aléatoire avec l'intervalle de prévision associé, deux cas de figure se présentent :

- L'intervalle de prévision des vitesses du vent est petit ; dans ce cas, selon la valeur de le vitesse du vent prévue, l'incertitude liée à la courbe de puissance peut être plus large que l'incertitude liée uniquement au modèle de prévision du vent (voir figure 4.4) ;
- L'intervalle de prévision des vitesses du vent est grand ; dans ce cas, selon la valeur de le vitesse du vent prévue, l'incertitude liée uniquement au modèle de prévision du vent peut couvrir l'incertitude liée à la courbe de puissance (voir figure 4.5).

D'où la nécessité de combiner les deux incertitudes pour obtenir l'intervalle de prévision de la production avec le modèle indirect. L'intervalle de prévision de la production est l'union entre l'incertitude sur la courbe de puissance et l'incertitude de la forêt aléatoire.

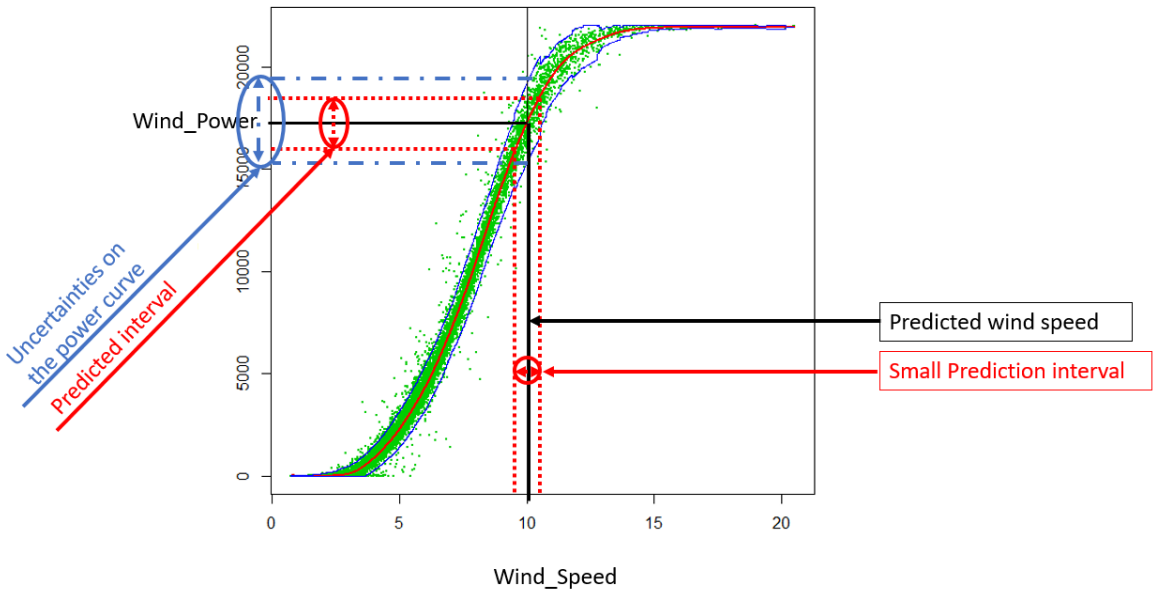


FIGURE 4.4 – Incertitudes liées à la courbe de puissance réelle et incertitudes liées à la prévision du vent par forêt aléatoire

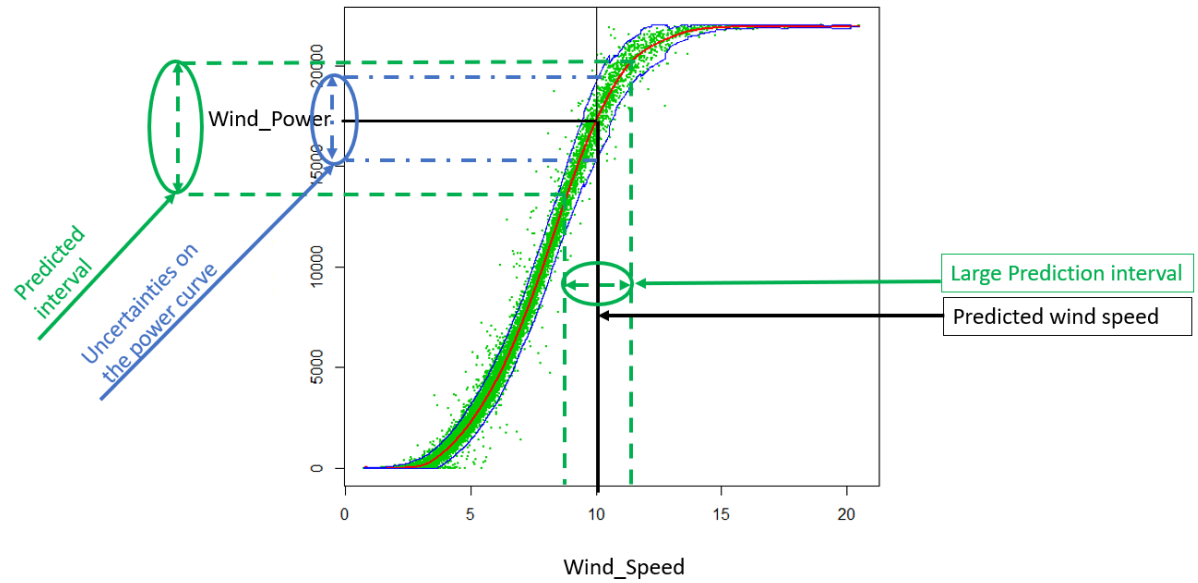


FIGURE 4.5 – Incertitudes liées à la courbe de puissance réelle et incertitudes liées à la prévision du vent par forêt aléatoire

4.5 Intervalle de prévision des prévisions de production éolienne

4.5.1 Applications

Dans cette partie nous avons appliqué les trois méthodes de construction d'intervalle de prévision avec les forêts aléatoires aux prévisions d'énergie éolienne sur l'année 2017 (parc éolien 1).

- La méthode de construction d'intervalle de prévision par quantile forest (algorithme 1) est du même principe que les forêts aléatoires. La différence principale réside de l'estimation de la fonction de distribution dans les nœuds terminaux pour les forêts de régression quantile pour en déduire les bornes de l'intervalle de prévision alors qu'avec les forêts aléatoires on estime une moyenne.

- La deuxième méthode ($MSPE1$, Algorithme 2) calcule l'erreur quadratique moyenne de prévision de la forêt aléatoire permettant de définir les bornes de l'intervalle de prévision.

- La troisième méthode ($MSPE2$, Algorithme 3) est similaire à la deuxième

méthode : le $MSPE1$ calcule l'erreur quadratique moyenne de prévision une fois pour chaque observation d'entrainement alors que le $MSPE2$ la calcule pour un sous-ensemble des observations d'entrainement et peut le faire plusieurs fois pour une observation.

Comme déjà expliqué dans la section 4.3, après avoir obtenu l'intervalle de prévision du vent par forêt aléatoire, la courbe de puissance réelle est utilisée pour obtenir les bornes de l'intervalle de prévision. Mais ces premières bornes ne tiennent pas compte de l'incertitude liée à la courbe de puissance. Il faut donc l'union de cet intervalle avec l'intervalle donné par l'incertitude sur la courbe de puissance pour obtenir l'intervalle de prévision finale de la production. Ces trois méthodes combinées à l'estimation de l'erreur de prévision de la spline nous permettent d'obtenir les trois intervalles de prévision de la production éolienne que nous avons notés $QF + Spline$ (quantile forests et spline), $MSPE1 + Spline$ ($MSPE1$ et spline) et $MSPE2 + Spline$ ($MSPE2$ et spline).

Les figures 4.6 ,4.7 et 4.8 sont des exemples des résultats d'intervalles de prévision des trois méthodes (chacun associé aux incertitudes sur la courbe de puissance). Il est assez difficile de comparer des intervalles de prévision sur simple observation. Ainsi nous utilisons la probabilité de couverture, la longueur moyenne de l'intervalle de prévision et l'indice introduit dans la sous-section 4.5.2.

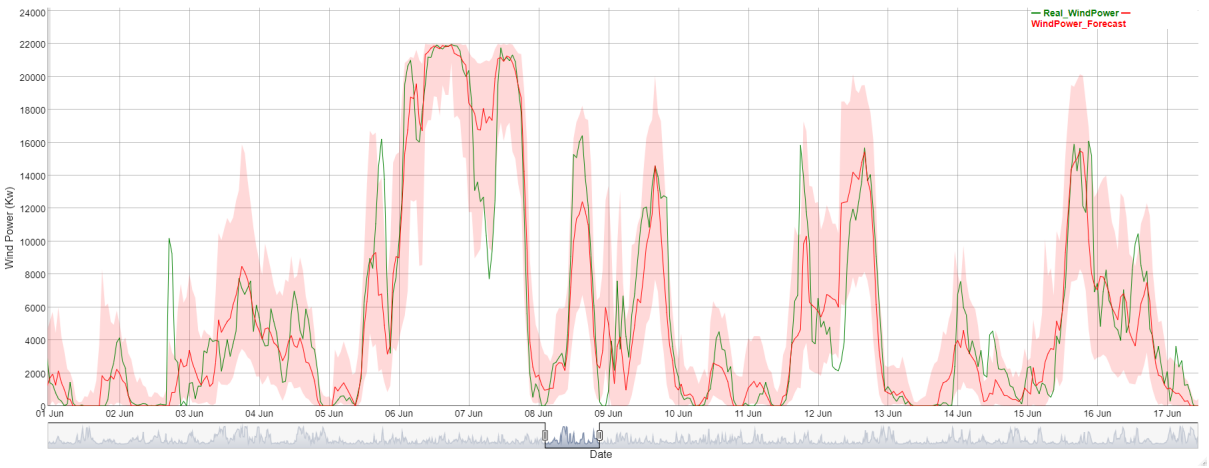


FIGURE 4.6 – Intervalle de prévision $QF + Spline$.

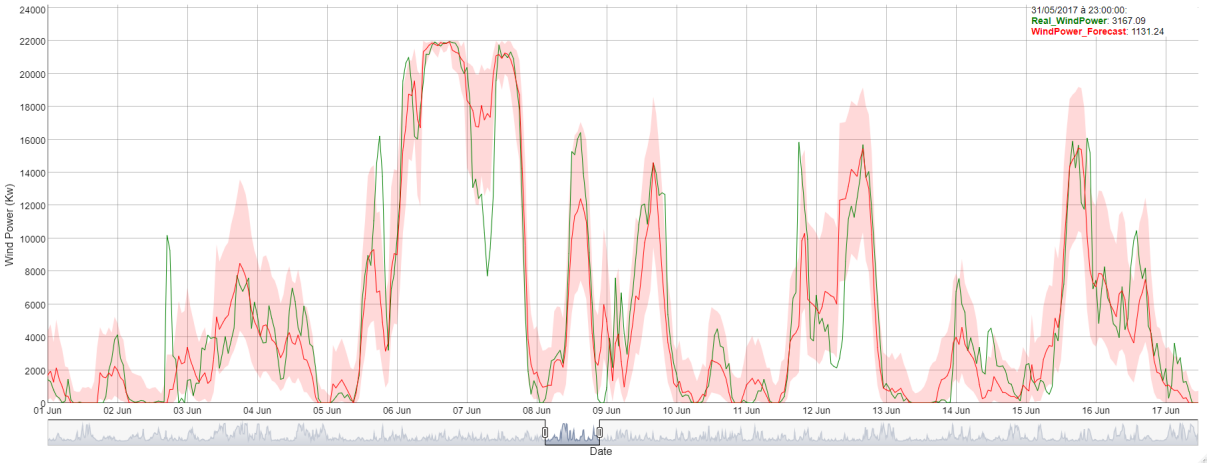


FIGURE 4.7 – Intervalle de prévision $MSPE1 + Spline$.

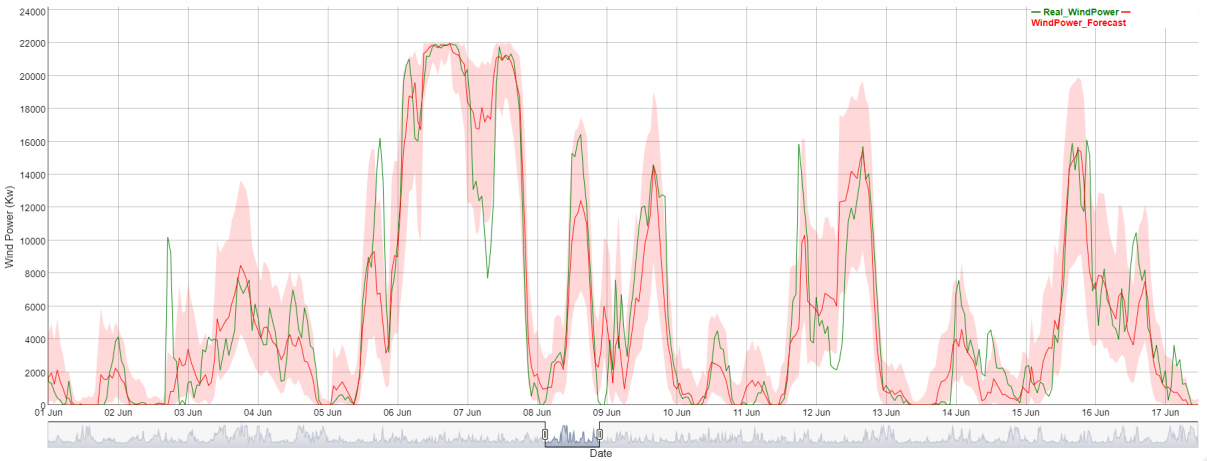


FIGURE 4.8 – Intervalle de prévision $MSPE2 + Spline$.

4.5.2 Comparaison des intervalles de prévision

Dans un problème statistique de prévision, plusieurs estimateurs d'intervalles de prévision peuvent être disponibles. Par conséquent, le choix d'un estimateur approprié est important. Le critère d'un bon estimateur est d'avoir

une probabilité de couverture élevée proche du niveau nominal et une longueur d'intervalle plus courte. Cependant, ces deux concepts s'opposent l'un à l'autre : les couvertures élevées ou faibles sont associées respectivement à des intervalles plus longs ou plus courts.
Certaines méthodes, comme l'échantillonnage bootstrap, modifient le niveau nominal pour améliorer la couverture et permettre ainsi la sélection d'intervalles basée uniquement sur la longueur des intervalles. Néanmoins, ces méthodes sont coûteuses sur le plan informatique. Ainsi, [Minkah and Wet, 2018] ont proposé un indice pour comparer des estimateurs d'intervalle de confiance (ou de prévision) basé sur un compromis entre la probabilité de couverture et la longueur de l'intervalle de confiance. On a utilisé cet indice pour comparer les intervalles de prévision. Soient R intervalles de prévision, $\eta = \{\eta_1, \ldots, \eta_R\}$ et $\mathbf{L} = \{L_1, \ldots, L_R\}$ les vecteurs des probabilités de couverture réalisés et les longueurs moyennes des intervalles respectivement. L'indice de l'intervalle de prévision est défini par :

$$I(L_j, \eta_j; \alpha) = k_\alpha \left(1 - \frac{1}{2} \left(\frac{1 + H(\eta_j; \alpha)}{1 + \left(\frac{\eta_j}{1 + L_j} \right)} \right) \right), \; L_j \geq 0, \; 0 \leq \eta_j \leq 1, \; j = 1, \ldots, R,$$

où k_α est une constante dépendant du niveau de significativité α. H pénalise l'écart entre la probabilité empirique de couverture et $1 - \alpha$:

$$H(\eta_j; \alpha) = |1 - \alpha - \eta_j|, \; 0 \leq \eta_j \leq 1, \; j = 1, \ldots, R.$$

Le paramètre d'échelle k_α est fixé à :

$$k_\alpha = \frac{4 - 2\alpha}{3 - 2\alpha},$$

pour obtenir des valeurs de $I(L_j, \eta_j; \alpha)$ dans le voisinage de la probabilité de couverture. Pour trouver les valeurs possibles de $I(L_j, \eta_j; \alpha)$, on examine les cas extrêmes :

I. $L_j \to 0, \; \eta_j \to 0 \Longrightarrow I(L_j, \eta_j; \alpha) \to \dfrac{k_\alpha \alpha}{2}$

II. $L_j \to \infty, \; \eta_j \to 0 \Longrightarrow I(L_j, \eta_j; \alpha) \to \dfrac{k_\alpha \alpha}{2}$

III. $L_j \to \infty, \; \eta_j \to 1 - \alpha \Longrightarrow I(L_j, \eta_j; \alpha) \to \dfrac{k_\alpha}{2}$

IV. $L_j \to 0, \; \eta_j \to 1 - \alpha \Longrightarrow I(L_j, \eta_j; \alpha) \to 1.$

Ainsi, $I(L_j, \eta_j; \alpha)$ est dans l'intervalle $[k_\alpha \alpha/2,\ 1]$. Un mauvais intervalle de prévision (c'est-à-dire un intervalle avec une faible probabilité de couverture et une grande longueur) correspond au cas I et II, avec $I(L_j, \eta_j; \alpha) \to \frac{k_\alpha \alpha}{2}$. D'un autre côté, un bon intervalle de prévision (c'est-à-dire le cas IV) a un indice $I(L_j, \eta_j; \alpha) \to 1$. Il faudra noter que d'autres fonctions de pénalité peuvent être choisies, par exemple une perte quadratique et dans ces cas des valeurs appropriées de k_α peuvent être déterminées analytiquement (voir [Minkah and Wet, 2018] pour plus de détails). Nous avons utilisé dans la suite cet indice avec la probabilité de couverture et la longueur moyenne des intervalles pour comparer les différents intervalles de prévision. Dans la suite nous avons fais un choix en interne de $\alpha = 0.2$ pour plus de précision autour de la valeur prédite.

Dans le tableau 4.1 on a comparé les trois méthodes par la probabilité de couverture, la longueur moyenne de l'intervalle de prévision et l'indice définie dans sous-section 4.5.2 qui est un compromis entre ces deux premiers indicateurs. L'intervalle de prévision généré à partir des quantile forests a une probabilité de couverture supérieure à la valeur nominale qui est égale à 80%, tandis que l'intervalle de prévision généré par la méthode $MSPE1 + Spline$ a une probabilité de couverture légèrement supérieure à celle nominale.

TABLE 4.1 – Comparaison des intervalles de prévision

Intervalle de prévision	Probabilité de couverture	Longueur moyenne	Indice
$QF + Spline$	86.84	6.24 Mw	0.64
$MSPE1 + Spline$	81.95	5.16 Mw	0.68
$MSPE2 + Spline$	82.5	5.32 Mw	0.68

Globalement les intervalles de prévision sont un peu larges (au moins 5 Mw pour une production maximale de 22 Mw). Nous estimons qu'il s'agit du caractère très variable du vent et des difficultés des modèles météorologiques à suivre correctement ces variations. On a également observé que les méthodes $MSPE1 + Spline$, $MSPE2 + Spline$ ont tendance à avoir des intervalles de prévision moins larges que celles produites par les quantile forests.

En observant l'indice de comparaison (qui est un bon compromis entre la probabilité de couverture et la longueur de l'intervalle de prévision), on note que les deux méthodes de construction d'intervalle de prévision $MSPE1 + Spline$ et $MSPE2 + Spline$ sont légèrement meilleures que la méthode $QF + Spline$.

4.6 Incertitude de phase

L'analyse des erreurs de prévision est un aspect important pour évaluer les performances d'un modèle de prévision. Une bonne connaissance des erreurs de prévision peut aider à améliorer la qualité d'un modèle de prévision afin de réduire les risques sur la vente d'énergie. Nous avons vu dans le chapitre 2 que le $NMAE$ ou le $NRMSE$ sont généralement utilisés. Cependant ces méthodes d'évaluation comparent les paires des séries temporelles, prévisions et observations et mesurent la distance verticale entre ces deux séries temporelles. Un autre aspect très important mérite d'être pris en compte dans l'évaluation des erreurs de prévision : le décalage horizontal ou la distance temporelle entre les prévisions et les observations. En effet un modèle peut prévoir correctement les pics de production mais avec un certain décalage temporel. La connaissance de ce type d'erreur est utile à la vente d'énergie et à la maintenance industrielle sur les sites éoliens.
L'indice de distorsion temporelle (ou TDI, Temporal Distortion Index en anglais) mesure la dissimilarité entre des séries temporelles ([Frías-Paredes et al., 2016] et [Frías-Paredes et al., 2017]). [Gastón et al., 2017] ont utilisé le TDI pour analyser des prévisions d'irradiation solaire. La section suivante résume le principe du TDI et sa combinaison avec le $NMAE$ pour évaluer les décalages verticaux et temporels entre les prévisions de production éolienne et les observations réelles.

4.6.1 Indice de distorsion temporelle et application

La caractérisation de l'erreur temporelle des prévisions est basée sur le principe de distorsion temporelle dynamique (ou Dynamic Time Warping, DTW) qui permet d'obtenir l'alignement optimal de deux séries temporelles en appliquant une optimisation dynamique à un problème du chemin le plus court. Le DTW procède à un ensemble de modifications de la série temporelle des prévisions pour obtenir un meilleur alignement par rapport aux observations réelles (ou réalisations) comme on peut le voir sur la figure 4.9.

On considère une série de prévisions notée $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_N)$ et une série

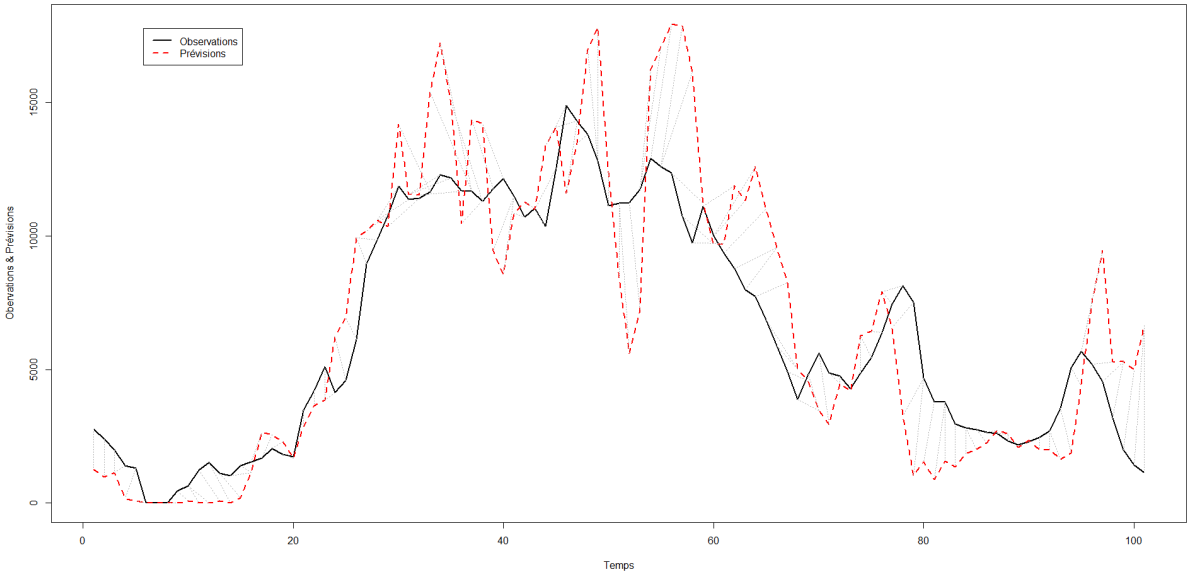


FIGURE 4.9 – Exemple représentatif du DTW.

de réalisations $Y = (Y_1, Y_2, \ldots, Y_M)$. On choisit $N = M$.
L'étape première consiste à calculer la distance locale entre les paires d'éléments $\hat{Y}_i$ et Y_i définie comme une fonction $f : \Phi \times \Phi \to \mathbb{R}^+$ avec

$$f(\hat{Y}_i, Y_j) = d(\hat{Y}_i, Y_j) = d_{ij} = \|\hat{Y}_i - Y_j\| \geq 0.$$

L'algorithme DTW commence par calculer une matrice des distances locales ($d \in \mathbb{R}^{N \times N}$), appelée matrice des coûts locaux (Local Cost Matrix) qui contient toutes les paires de distances correspondantes entre les deux séries. Après avoir défini la matrice des coûts locaux, le concept de chemin entre les séries est introduit sous la forme d'une séquence de points $w = (w_1 = (i_1, j_1), w_2 = (i_2, j_2), \ldots, w_k = (i_k, j_k))\ k \in \mathbb{N}$ où $w_l = (i_l, j_l) \in [1 : N] \times [1 : N]$ pour tout $l \in [1 : k]$ sous les contraintes suivantes.

- Conditions aux bornes : $w_1 = (1, 1)$ et $w_k = (N, N)$. Cette condition exige que le chemin commence et se termine respectivement au premier et au dernier point de la séquence.

- Condition de monotonicité : si $w_l = (i_l, j_l)$ alors $w_{l-1} = (i_{l-1}, j_{l-1})$ et $i_l - i_{l-1} \geq 0$ et $j_l - j_{l-1} \geq 0$. Cette condition garantit que les points sont triés en fonction du temps à travers le chemin.

- Condition de continuité : si $w_l = (i_l, j_l)$ alors $w_{l-1} = (i_{l-1}, j_{l-1})$ et $i_l - i_{l-1} \leq$

1 et $j_l - j_{l-1} \leq 1$. Cette condition assure que le chemin ne présente pas de grands sauts et est limité aux points voisins.

Le chemin peut être tracé sur une grille $N \times N$ où l'axe des abscisses représente les indices temporels dans la série des prévisions et l'axe des ordonnées les indices temporels dans la série des observations (figure 4.10). Le

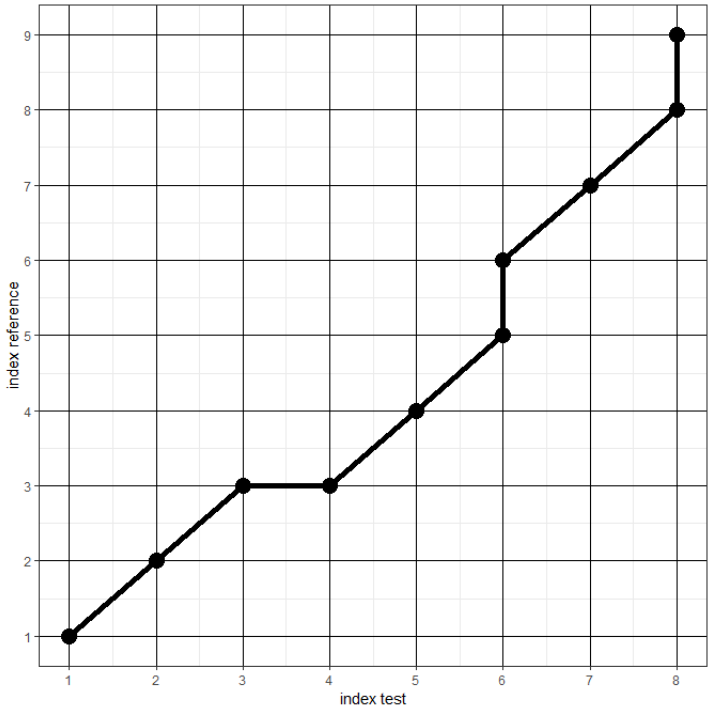


FIGURE 4.10 – Exemple de chemin de la relation entre les indices temporelles des séries d'observations et de prévisions.

coût total associé au chemin w entre les séries $\hat{Y}$ et Y par rapport à sa distance locale $c_w(\hat{Y}, Y)$ est définie par l'expression :

$$c_w(\hat{Y}, Y) := \sum_{l=1}^{k} d(\hat{Y}_{i_l}, Y_{j_l}) \text{ où } (i_l, j_l) = w_l.$$

Par conséquent, un chemin optimal entre les séries $\hat{Y}$ et Y est un chemin w^* qui présente un coût total minimal prenant en compte tous les chemins possibles :

$$c_{w^*}(\hat{Y}, Y) = min\{c_w(\hat{Y}, Y) | w \textit{ est un chemin}\}.$$

Le chemin optimal w^* est obtenu en utilisant la programmation dynamique (Dynamic Programming DP). La programmation dynamique est une procédure permettant de résoudre des problèmes d'optimisation en les décomposant en

problèmes plus simples. La base de la programmation dynamique est le principe d'optimalité de Bellman ([Bellman, 1957]).
On identifie les séquences $\hat{Y}(1:i) := (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_i)$ où $i \in [1:N]$ et $Y(1:j) := (Y_1, Y_2, \ldots, Y_j)$ où $j \in [1:N]$ et on définit :

$$D(i,j) = c_{w^*}(\hat{Y}(1:i), Y(1:j)),$$

alors, $D(i,j)$ est le coût associé au meilleur chemin correspondant aux séries $(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_i)$ et $(Y_1, Y_2, \ldots, Y_j)$. Ainsi les valeurs $D(i,j)$ définissent une matrice $D \in \mathbb{R}^{N \times N}$ connue sous le nom de la matrice des coûts cumulés. On a :

$$D(N,N) = c_{w^*}(\hat{Y}, Y).$$

En tenant compte des contraintes citées dans la définition du chemin, le calcul des éléments de la matrice des coûts peut être automatisé à l'aide de la formule récursive suivante :

$$D(i,j) = min \begin{cases} D(i-1,j) & + & d(i,j) \\ D(i,j-1) & + & d(i,j) \\ \underbrace{D(i-1,j-1)}_{\text{coût cumulé}} & + & \underbrace{d(i,j)}_{\text{coût actuel}} \end{cases}$$

Notez qu'il y a trois transitions possibles d'une paire dans le chemin à la suivante, toutes vérifiant la condition de continuité. La figure 4.11 contient le schéma de notation de la formule récursive, également appelée Step Pattern.

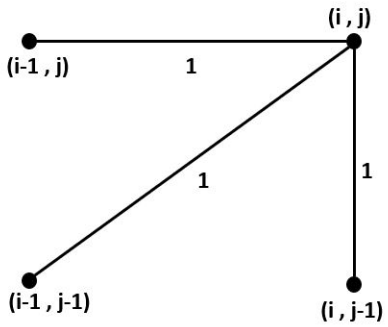


FIGURE 4.11 – Schéma de la formule récursive. Ce Step Pattern est connu sous le nom de Symetric1.

La condition de continuité incluse dans la définition du chemin w peut être assouplie en introduisant une transition supplémentaire possible entre les éléments consécutifs de la trajectoire. De plus, en mesurant les distances

locales, il est possible de pénaliser certains mouvements. Par conséquent, certaines modifications à l'ensemble des contraintes peuvent être incluses afin d'obtenir un meilleur contrôle des chemins possibles en fonction des besoins du cas d'étude. Il existe donc d'autres types de schéma récursif (voir [Frías-Paredes et al., 2017] pour plus de détails).

Une fois le chemin optimal obtenu, la série temporelle des prévisions est transformée en une série appelée "série alignée" : $S = (S_1, S_2, \ldots, S_N)$. Avec une différence verticale plus faible par rapport à la série des observations que la série des prévisions.
Soit $w^* = (w_1 = (i_1, j_1), w_2 = (i_2, j_2), \ldots, w_k = (i_k, j_k))$ le chemin optimal, $S_j = \hat{Y}_{f(j)}$ $(j = 1, \ldots, N)$ où f est une fonction d'interpolation vérifiant $f(j_l) = i_l$ $(l = 1, \ldots, k)$ et $S_j = g(f(j))$ avec $f(j) \in \mathbb{R}$ et g représentant la fonction d'interpolation qui vérifie $g([f(j)]) = \hat{Y}_{[f(j)]}$ et $g([f(j)] + 1) = \hat{Y}_{[f(j)]+1}$. Il faut noter que le chemin optimal recueille la relation entre l'indice temporel de la série des observations Y et les séries des prévisions $\hat{Y}$ pour obtenir la série alignée S.
Si le chemin optimal w^* passe par le point $w_l = (i_l, j_l)$, on peut en déduire qu'il y a un mouvement temporel de $|i_l - j_l|$ unités de temps entre la série des prévisions et la série alignée. Par conséquent, cette différence doit être prise en compte pour obtenir l'élément $D(i, j)$ de la matrice des coûts, surtout lorsque le minimum de la formule récursive est atteint dans deux branches différentes ou plus. Par exemple si deux prédécesseurs possibles du point (i, j), (i^*, j^*) et (i^{**}, j^{**}), donnent la même valeur de $D(i, j)$, la paire produisant la modification mineure dans la série des prévisions doit être sélectionnée, c'est-à-dire $min\{|i^* - j^*|, |i^{**} - j^{**}|\}$. Ainsi, la série alignée résultante est celle, parmi toutes les séries associées à un chemin optimal, qui fournit un désalignement temporel mineur de la série des prévisions. De même, la série alignée est celle dont le chemin optimal associé est le plus proche du chemin identité (le chemin formé par les points $\{(1, 1), (2, 2), \ldots, (N - 1, N - 1), (N, N)\}$).
Une mesure globale de la distorsion temporelle réalisée dans la série des prévisions, afin d'obtenir la série alignée, est donnée par l'aire entre le chemin optimal résultant et le chemin identité. Cette mesure est l'indice de distorsion temporelle (TDI, temporal distortion index). Ce paramètre est utilisé pour décrire l'erreur temporelle. On définit

$$P_l = \int_{i_l}^{i_{l+1}} x - \left(\frac{(x - i_l)(j_{l+1} - j_l)}{(i_{l+1} - i_l)} + j_l \right) dx$$

pour chaque étape de la trajectoire optimale qui ne traverse pas la diagonale principale (le chemin identité). Dans le cas contraire, le segment du chemin

optimal (i_l, j_l) , (i_{l+1}, j_{l+1}) traversera la diagonale principale au point

$$\left(\frac{i_l j_l + j_l i_{l+1} - i_l j_{l+1} - j_l i_l}{i_{l+1} - i_l - (j_{l+1} - j_l)}, \frac{i_l j_l + j_l i_{l+1} - i_l j_{l+1} - j_l i_l}{i_{l+1} - i_l - (j_{l+1} - j_l)}\right)$$

et on définit

$$\begin{aligned} P_l &= \int_{i_l}^{\frac{i_l j_l + j_l i_{l+1} - i_l j_{l+1} - j_l i_l}{i_{l+1} - i_l - (j_{l+1} - j_l)}} x - \left(\frac{(x - i_l)(j_{l+1} - j_l)}{(i_{l+1} - i_l)} + j_l\right) dx \\ &+ \int_{\frac{i_l j_l + j_l i_{l+1} - i_l j_{l+1} - j_l i_l}{i_{l+1} - i_l - (j_{l+1} - j_l)}}^{i_{l+1}} x - \left(\frac{(x - i_l)(j_{l+1} - j_l)}{(i_{l+1} - i_l)} + j_l\right) dx, \end{aligned}$$

l'indice de distorsion temporelle est donné par :

$$TDI = \frac{2\sum_{l=1}^{k-1} |P_l|}{N^2}.$$

L'indice de distorsion temporelle est un nombre sans dimension variant dans l'intervalle $[0, 1]$, où 0 correspond à une distorsion temporelle nulle et 1 à une distorsion temporelle maximale. La figure 4.12 est un exemple de l'expression du TDI. Le TDI dans cet exemple est le quotient entre la zone rouge et la zone bleue.

L'indice de distorsion temporelle est combiné avec le MAE ou le $NMAE$ pour définir une erreur bidimensionnelle (BE, Bidimensional Error) avec une composante statique et une composante temporelle :

$$BE_{FR}(\hat{Y}, Y) = (TDI, NMAE(S))_{FR},$$

où FR est la formule récursive utilisée et $NMAE(S) = (100/PI) * \sum_{i=1}^{N} \frac{|S_i - Y_i|}{N}$. Il est important de noter que les différents alignements (obtenus à partir de différentes formules récursives) produisent différents vecteurs d'erreur bidimensionnelle. L'erreur bidimensionnelle entre la série des observations et la série des prévisions dans le cas de non-alignement ($TDI = 0$) est notée et définie par :

$$BE_0(\hat{Y}, Y) = (0, NMAE(\hat{Y})).$$

4.6.2 Application aux prévisions éoliennes

La méthodologie du calcul de l'erreur bidimensionnelle est appliquée aux prévisions d'énergie éolienne. La figure 4.13 (a) représente une période de quatre jours avec des variations brusques du vent provoquant des pics de

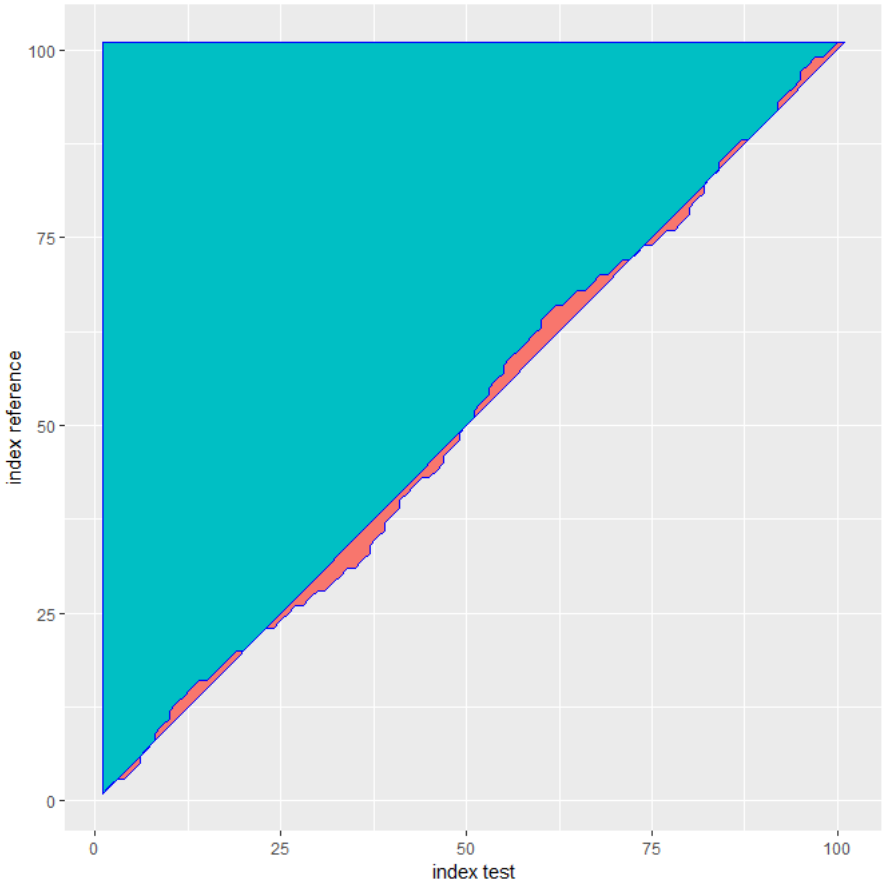


FIGURE 4.12 – Exemple de chemin identité, du chemin optimal et des surfaces utilisées dans l'expression du TDI.

production. On peut noter certains décalages des prévisions éoliennes par rapport à aux observations. Le modèle de prévision peine à prévoir les pics de production aux dates où ils sont observés. Ces problèmes de désalignement sont dus en général aux modèles météorologiques dont les données sont utilisées en entrée par les modèles de prévision de la production éolienne. Les modèles météorologiques ont également des difficultés à reproduire les variations brusques du vent sur des intervalles de temps courts. Les prévisions des modèles météorologiques ont tendance à être plus lisses que les observations réelles, ceci se reproduit sur les modèles de prévision de la production éolienne. La figure 4.13 (b) montre la série alignée des prévisions éoliennes et les observations réelles. On note une correction du décalage temporel avec la série alignée qui est plus proche des observations réelles.

L'erreur bidimensionnelle est représentée sur la figure 4.14 (c). Lorsque la composante temporelle de l'erreur bidimensionnelle est nulle ($TDI = 0$), le $NMAE$ est de l'ordre de 9% (indicateur classique entre les prévisions et les

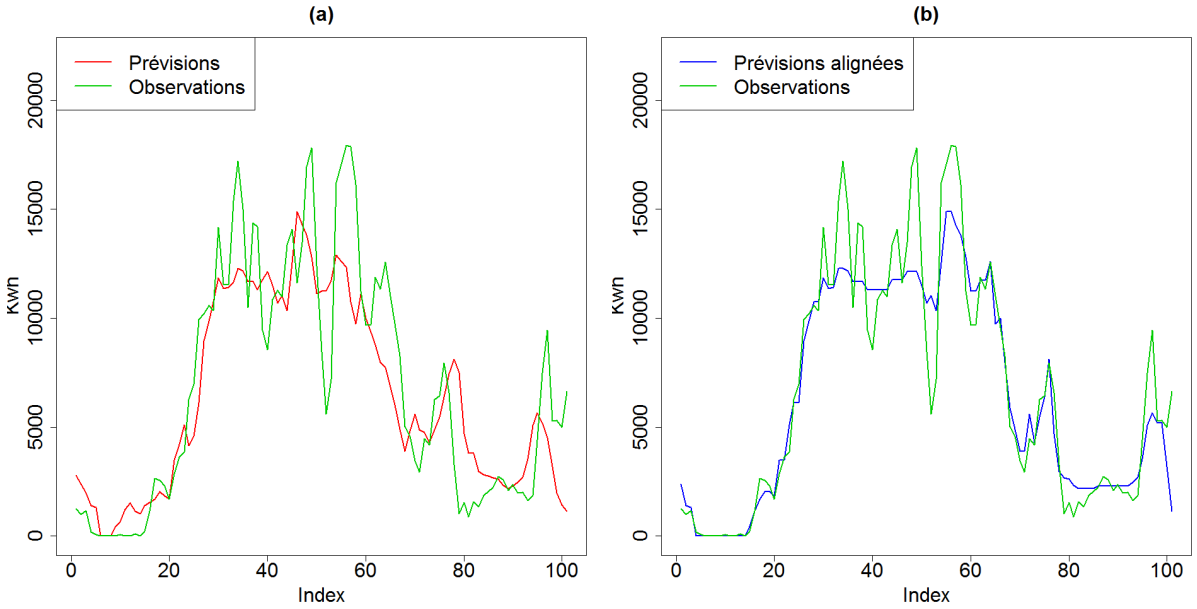


FIGURE 4.13 – Application TDI pour des prévisions éoliennes. La figure (a) montre les prévisions (courbe rouge) et les observations (courbe verte). La figure (b) représente la série alignée (courbe bleue) obtenue après application du DTW et les observations (courbe verte).

observations). Après évaluation, l'indice de distorsion est de l'ordre de 4% avec une diminution du $NMAE$ jusqu'à 5% (point bleu sur la figure 4.14 (c)). On peut dire dans ce cas que le désalignement temporel est responsable de 50% de l'erreur de prévision mesurée par le $NMAE$.

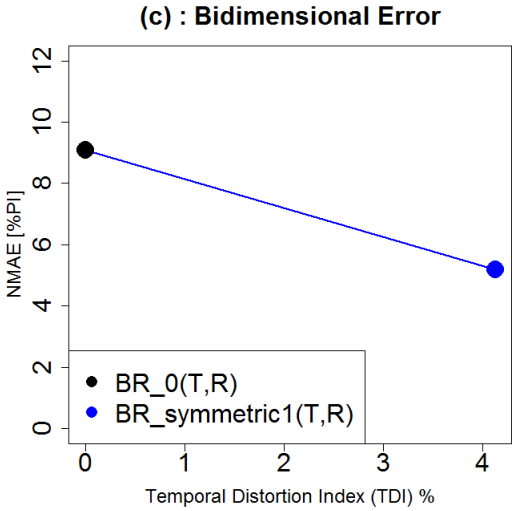


FIGURE 4.14 – L'erreur bidimensionnelle avant (point noir) et après alignement (point bleu).

La figure 4.15 est une application avec trois formules récursives différentes (symmetric1, symmetric2 et la formule récursive de Rabiner Juang figure 4.16). On observe que la formule récursive "symmetric2" minimise l'erreur bidimensionnelle de prévision.

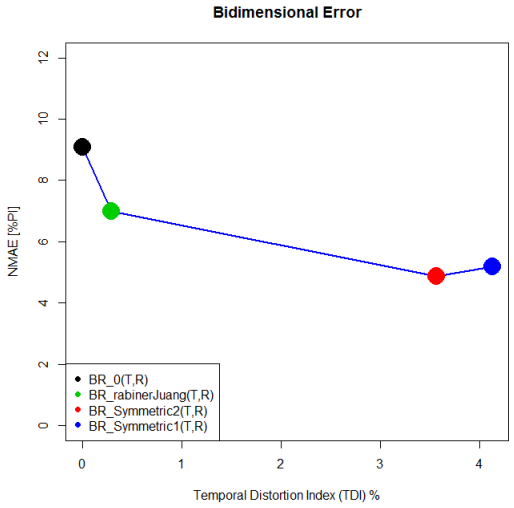


FIGURE 4.15 – L'erreur bidimensionnelle en fonction de différentes formules récursives.

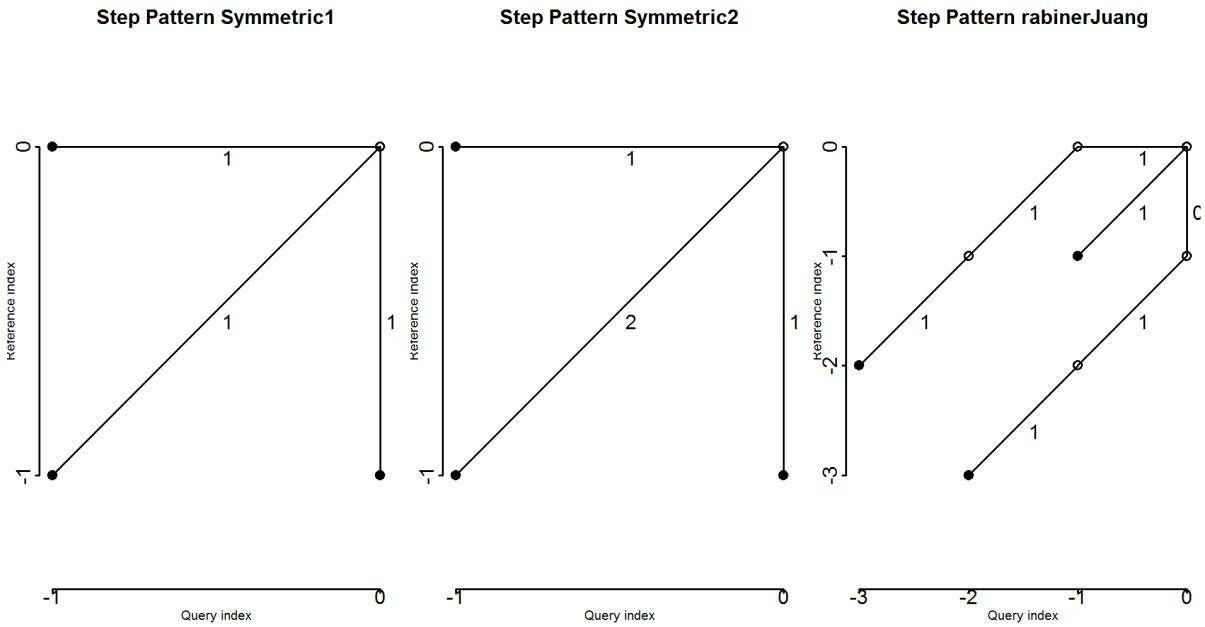


FIGURE 4.16 – Trois formules récursives : Symmetric1, symmetric2 et rabinerJuang-stepPattern.

4.7 Conclusion

Dans ce chapitre nous avons étudié les intervalles de prévision par forêt aléatoire. Nous avons utilisé un indice qui est un compromis entre la probabilité de couverture et la longueur de l'intervalle de prévision pour évaluer les méthodes de construction d'intervalle de prévision.
Nous avons aussi étudié l'incertitude de phase (la distorsion temporelle entre les prévisions et les réalisations) à l'aide de l'indice de distorsion temporelle. Cet indice permet d'évaluer les écarts verticaux et le décalage temporel entre les prévisions éoliennes et les réalisations. Le désalignement temporel semble être en grande partie responsable de l'erreur de prévision observée sur la prévision de la production éolienne.

Chapitre 5

Industrialisation du modèle de prévision court terme

Sommaire

5.1 Introduction

La prévision d'énergie éolienne jouera un rôle déterminant dans la transition énergétique notamment sur la maîtrise des coûts d'intégration de l'énergie éolienne dans le réseau électrique et le marché électrique. L'énergie éolienne commençant à être directement vendue sur le marché par les producteurs à cause de la rupture des contrats d'obligation d'achat, les gestionnaires de réseaux de distribution et les gestionnaires de réseaux de transport demandent ou même obligent les producteurs à fournir au moins des prévisions de production un jour à l'avance pour rééquilibrer le réseau. Ainsi, c'est suite à ces besoins et pour avoir un modèle de prévision interne que les travaux de cette

thèse ont été lancés. Le modèle de prévisions découlant de ces travaux a été développé pour permettre à ENGIE d'avoir chaque jour ses propres prévisions pour l'ensemble de ses parcs éoliens.
Jusqu'à présent pour ses besoins de prévisions, ENGIE est lié à un fournisseur par un contrat de fourniture de prévisions de production pour une partie de son périmètre (France, Afrique du sud, Belgique, Roumanie, Pologne et Mexique). Dans ce chapitre, nous avons abordé la mise en production de modèle découlant de nos travaux et la comparaison de la qualité de nos prévisions avec celle du fournisseur. Les résultats des analyses ont montré que nous fournissons les meilleures prévisions et que les prévisions internes calculées dans Darwin font baisser la facture de façon drastique .

La section 5.2 revient sur le contexte du développement industriel du projet DARWIN qui intègre notre modèle et le service Darwin Forecaster qui fournit nos prévisions (qui sont nommées Darwin-Production). La section 5.3 étudie la stabilité du modèle dans le temps. Dans la section 5.4, nous avons comparé notre modèle de prévision au modèle du fournisseur actuel de prévisions. La section 5.5 analyse les coûts du produit Darwin et l'apport potentiel de nos prévisions pour ENGIE. La conclusion de ce chapitre est donnée dans la section 5.6.

5.2 Projet Darwin et le service Darwin Forecaster

L'injection d'énergie renouvelable (très variable pour certaines énergies renouvelables comme l'éolien) introduit des déséquilibres dans le système énergétique. Par conséquent, des prévisions de production précises permettent de réduire les pénalités pouvant être liées à une sur-estimation ou à une sous-estimation de la production réelle. En parallèle, elles jouent aussi un rôle important dans la réduction du coût d'exploitation en servant à optimiser la planification des travaux de maintenance. Par exemple, planifier les maintenances lorsque les prévisions de production sont faibles.
La gestion du système électrique et l'optimisation de la vente d'énergie éolienne reposent sur des prévisions de qualité. Les prévisions éoliennes (et solaires) seront omniprésentes dans les activités quotidiennes des producteurs. Et le fait de disposer de modèles internes permettra de réduire les coûts et de créer de nouveaux services.

5.2.1 Le projet Darwin

Le projet Darwin est conçu pour améliorer les performances opérationnelles des parcs et développer la maintenance prédictive. Le projet s'inscrit dans le cadre de la digitalisation des opérations d'ENGIE. Darwin permet :

- une meilleure gestion des périodes de la maintenance : les données collectées sont transformées en plan d'action pour faciliter maintenance des installations ;
- une optimisation des coûts : Darwin aide les clients d'ENGIE à cibler les périodes durant lesquelles les activités de maintenance sont les plus rentables ;
- une amélioration de la performance : la cartographie des forces et des faiblesses des parcs grâce aux algorithmes de data analytics et aux modèles prédictifs permet d'augmenter la rentabilité des parcs de production ;
- une augmentation de l'acceptabilité des parcs via une meilleure communication sur leur finalité.

Darwin d'ENGIE est la plateforme digitale unique pour optimiser les actifs renouvelables d'ENGIE à travers le monde. Le fonctionnement se résume sur la figure 5.1.

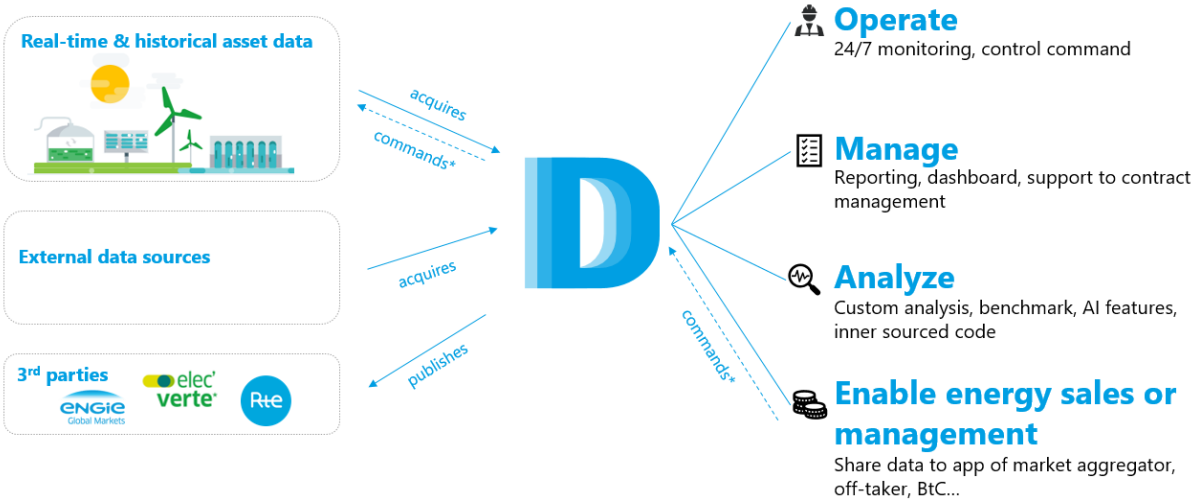


FIGURE 5.1 – Résumé des services de Darwin. Source ENGIE.

Darwin est donc une plateforme offrant plusieurs services mais nous choisissons de développer dans la section suivante le service dédié à la prévision : Darwin Forecaster. En effet le modèle de prévision éolienne découle de nos travaux. Darwin Forecaster intègre aussi les prévisions éoliennes d'autres modèles (en particulier le fournisseur actuel de prévisions). C'est la raison pour laquelle nous avons aussi évalué les qualités de nos prévisions par rapport à celles du fournisseur afin d'expliquer l'apport industriel de nos travaux.

5.2.2 Le service Darwin Forecaster

Le premier modèle qui est implémenté dans Darwin est un modèle indirect avec les variations spatio-temporelles du vent sur trois points de grille (modèle avec le point le plus proche et les deux autres points qui entrent dans le calcul des variations spatiales du vent). Le modèle est entrainé sur l'historique disponible de données météorologiques, de données réelles de production et de vitesse du vent pour chaque parc éolien. La collecte des données météorologiques par la plateforme Darwin étant ressente, il y a au mois de mars 2020 environ 5 mois d'historique. Pour les nouveaux parcs qui n'ont pas d'historique de données, la prévision est faite par la courbe de puissance théorique (le vent est juste converti en prévision de production en utilisant la courbe de puissance théorique). L'entrainement du modèle est réactualisé chaque dimanche à minuit. Chaque jour les données météorologiques sont récupérées pour ensuite prédire la production des parcs éoliens. Même si nos travaux ont portés sur un horizon de prévision de 24 à $47h$, il a été décidé dans Darwin de prédire la production jusqu'à un horizon de dix jours (avec le même modèle) et d'afficher aussi la prévision cumulée (figure 5.2). Aujourd'hui, 313 parcs éoliens et de 138 centrales photovoltaïques sont connectés à Darwin pour une capacité de $6.1GWc$.
Bien qu'elles soient nommées dans Darwin (voir figure 5.2) "Puissance ACTIVE" et "Production CUMULÉE", il s'agit respectivement de la "prévision de la puissance active" et de la "prévision de la production cumulée".
Darwin Forecaster est le produit dédié à la prévision des actifs éoliens et solaires du groupe ENGIE. Le modèle de prévision fournissant les prévisions de production éolienne dans ce produit est le fruit des travaux de cette thèse et d'une collaboration entre ENGIE North America, Engie Green France et Engie digital. Le produit Darwin Forecaster intègre aussi les prévisions du fournisseur actuel de prévisions dont nous avons comparé les prévisions à celles du modèle interne dans la section suivante. L'une des étapes de l'évolution dans le développement de Darwin est l'intégration des intervalles de prévision que

nous avons étudiés dans le chapitre 4 et des prévisions des vitesses du vent dans le modèle indirect. C'est la raison pour laquelle il n'y a pas d'intervalle de prévision et de prévisions du vent sur la figure 5.2.

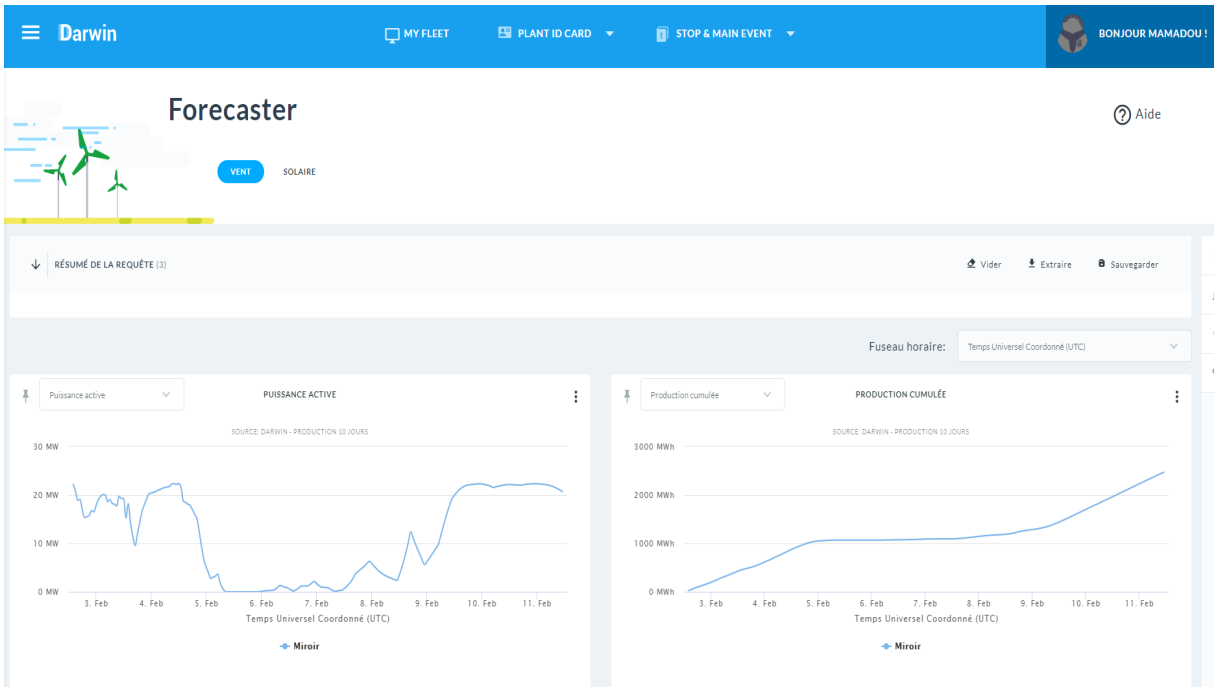


FIGURE 5.2 – Exemple des prévisions de production (à gauche) et des prévisions cumulées (à droite) du 02/2020 au 11/02/2020 de notre modèle développé dans Darwin.

Darwin fournit en parallèle les prévisions d'un fournisseur externe. La section suivante explique le modèle de prévision de ce fournisseur.

5.2.3 Le fournisseur de prévisions

Aujourd'hui ENGIE Green a un contrat avec un fournisseur de prévisions. Le modèle de prévision du fournisseur s'appuie sur les données météorologiques du modèle ECMWF. Le fournisseur externe ne fait pas de prévisions de production proprement dit contrairement à nous. En effet, son modèle ne fournit que des prévisions de vitesses du vent et la courbe de puissance théorique de l'éolienne est directement appliquée aux prévisions de vent fournies pour prédire la production des parcs éoliens d'une partie du périmètre d'ENGIE avec un horizon de prévision allant jusqu'à neuf jours (voir figure 5.3). Les prévisions du fournisseur s'accompagnent aussi de deux intervalles de prévision : un intervalle à 50% de confiance et un intervalle à 80% de confiance. Les intervalles de prévision et les prévisions des vitesses de vent du fournisseur sont déjà développés dans Darwin.

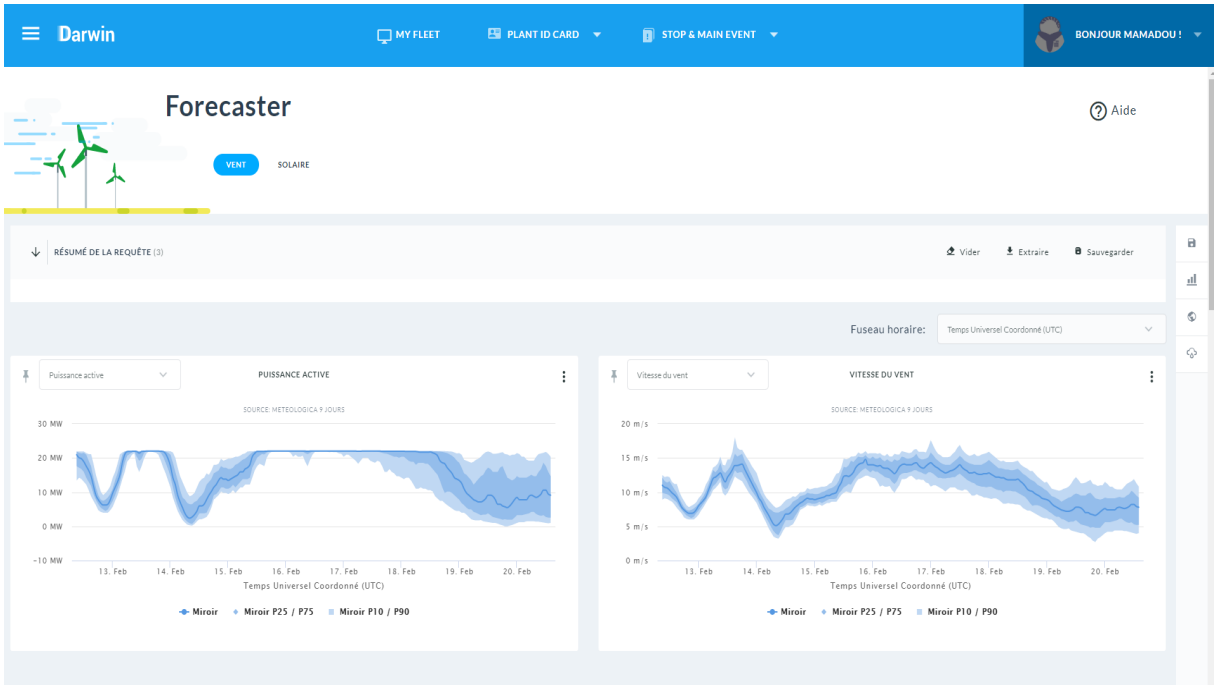


FIGURE 5.3 – Prévisions de production et de vitesse du vent du modèle fournisseur sur la période du 12/02/2020 au 20/02/2020.

Nous verrons dans la section suivante que le modèle développé en interne donne de meilleures prévisions que celle du fournisseur actuel de prévisions (en prévision $J+1$ plus utile pour la vente d'énergie).

5.3 Stabilité du modèle dans le temps et en fonction des périodes d'apprentissage

Dans cette section nous avons étudié la stabilité de la performance du modèle de prévision dans le temps et en fonction la période d'apprentissage. En effet dans Darwin on utilise l'historique disponible comme période d'apprentissage. Du coup, le nombre d'années dans la période d'apprentissage varie. On s'est donc posé la question de savoir s'il y a des variations importantes des erreurs de prévision en calibrant le modèle avec deux ans, trois ans ou quatre ans d'historique de données ? La deuxième question que nous nous sommes aussi posée est comment évoluent les erreurs de prévision d'une année prévision à une autre ? Pour apporter une réponse à ces questions, nous avons travaillé avec le modèle à un point de grille qui est implémenté dans Darwin. Pour ce faire, nous disposons d'un historique de données de production et de données météorologiques de cinq ans (2015 à 2019) sur le parc

éolien 1 (en local et non dans la plateforme Darwin). A noter que pour l'année 2019, à cause des dysfonctionnements observés sur le parc éolien entre le mois de septembre et le mois de décembre, nous avons uniquement retenu la période du premier janvier au quinze septembre dans l'analyse.
Nous savons déjà que sur ce parc éolien, avec un apprentissage sur les années 2015 et 2016 et une prévision sur l'année 2017, les erreurs de prévision étaient égales à 7.45% (Chapitre 3, section 3.2.2). Pour analyser la stabilité du modèle dans le temps et en fonction de la période d'apprentissage nous avons étudié les scénarios suivants :

- apprendre le modèle sur les années 2015, 2016 et 2017 puis prédire la production sur l'année 2018 ;
- apprendre le modèle sur les années 2016 et 2017 puis prédire la production sur l'année 2018 ;
- apprendre le modèle sur les années 2017 et 2018 puis prédire la production sur l'année 2019 ;
- apprendre le modèle sur les années 2016, 2017 et 2018 puis prédire la production sur l'année 2019 ;
- apprendre le modèle sur les années 2015, 2016, 2017 et 2018 puis prédire la production sur l'année 2019.

Les résultats observés dans le tableau 5.1 peuvent être analysés sous deux angles. D'une part si on s'intéresse à la variation du $NMAE$ dans le temps (c'est-à-dire les erreurs de prévision en 2017, 2018 et 2019) indépendamment du nombre d'années dans la période d'apprentissage. On observe que les erreurs de prévision sont relativement stables sur les années 2017, 2018 et 2019 avec une diminution en 2019. Nous estimons que cette diminution est due au fait que la période du 16/09/2019 au 31/12/2019 n'est pas prise en compte comme expliqué précédemment. D'autre part si on analyse les erreurs de prévision en fonction du nombre d'années dans la période d'apprentissage (2 ans, 3 ans ou 4 ans), on note également qu'il n'y a pas une grande différence de l'erreur de prévision en 2017, 2018 et 2019. Les erreurs de prévision sur l'année 2019 avec un modèle appris sur quatre ans sont plus faibles mais comme nous l'avons si bien remarqué dans la première partie de l'analyse, la

période du 16/09/2019 au 31/12/2019 n'est pas prise en compte.
En définitive, avec deux ans d'apprentissage, les erreurs de prévision restent relativement stables dans le temps.

TABLE 5.1 – Erreur de prévision dans le temps et en fonction de la période d'apprentissage

Période d'apprentissage	Période test	NMAE
2015-2016	2017	7.45%
2016-2017	2018	7.23%
2015-2016-2017	2018	7.21%
2017-2018	2019	7.05%
2016-2017-2018	2019	7.04%
2015-2016-2017-2018	2019	6.98%

5.4 Comparaison modèle fournisseur/modèle Darwin-Production

La valeur ajoutée est la contribution au projet DARWIN avec la prévision de la production pour tous les actifs éoliens de la flotte d'ENGIE connectée à DARWIN. Il s'agit d'une innovation tant au niveau du produit lui-même que des moyens utilisés pour le mettre en œuvre. Tout d'abord, nous industrialisons un modèle de prévision de la production éolienne à la pointe de la technologie qui produit des résultats fiables et robustes. Il a été mis en place dans le cadre de cette thèse et validé dans un "datascience challenge". Le modèle de prévision s'appuie sur les prévisions météorologiques ECMWF disponible au niveau mondial.

Deuxièmement, l'industrialisation d'un modèle de prévision à l'échelle mondiale est une réussite en soi et a été rendue possible par une première collaboration numérique de ce type au sein du groupe ENGIE avec des acteurs d'ENGIE North America, d'Engie Green France et d'Engie Digital. Jusqu'à présent pour ses besoins de prévisions de production, ENGIE Green a un contrat avec un fournisseur pour un coût d'environ 10 000 euros par an.

5.4.1 Comparaison des prévisions

Nous proposons dans cette partie une étude comparative entre les prévisions de production du modèle du fournisseur et celles issues d'une simulation en local du modèle Darwin-Production sur la période de janvier à juillet 2019. Nous avons choisi un parc éolien de $22Mw$ (Wind farm 1 du chapitre 2). Le $NMAE$ sur la période, par mois et par horizon de prévision est l'indicateur statistique utilisé. La production totale du parc en un instant t découle d'une part, d'une moyenne des 60 relevées de la production par minute précédant t pour chaque éolienne (pour passer de la fréquence minute en fréquence horaire) et d'autre part, de la somme des productions horaires des éoliennes (voir chapitre 2). Nous rappelons la définition de la production totale du parc que nous avions qualifiée de type de production 2.
Type de production 2 : on ne tient compte que des périodes où au moins une éolienne n'a eu d'arrêt ou de dysfonctionnement durant les 60 minutes précédant l'instant t. On calcule la moyenne horaire de production des éoliennes qui ont fonctionné durant toutes les 60 minutes comme suit :

$$ProdH_1^E(t) = \frac{1}{60} \sum_{i=t-59}^{t} ProdMin^E(i).$$

On note Ne ce nombre d'éoliennes. La production horaire du parc est :

$$ProdH_2^{parc}(t) = \left(\frac{1}{Ne} \sum_{E=1}^{Ne} ProdH_1^E(t) \right) \times NT.$$

Si toutes les éoliennes ont fonctionné durant les 60 minutes, le type de production 2 correspond à la production du parc sans aucun dysfonctionnement. Si au moins une éolienne n'a pas fonctionné toutes les 60 minutes, $ProdH_2^{parc}(t)$ est une production corrigée du parc en utilisant uniquement les éoliennes ayant fonctionné durant les 60 minutes.
Le tableau 5.2 résume les erreurs de prévision moyennes des deux modèles sur la période étudiée. Le résultat de ce tableau montre que le modèle Darwin-Production a fourni les meilleures prévisions. Ce résultat nous donne une indication sur les performances moyennes des deux modèles de prévision. Nous

avons aussi analysé en détail les prévisions des deux modèles en s'intéressant aux performances par horizon de prévision et par mois.

TABLE 5.2 – $NMAE$ parc éolien 1 : modèle Darwin-Production vs modèle du fournisseur.

	Darwin-Production	Fournisseur
NMAE (%)	7.04	10.64

La figure 5.4 montre la distribution des erreurs de prévision des deux modèles pour chaque horizon de prévision. Le modèle Darwin-Production fait moins d'erreurs de prévision à tous les horizons exceptés les horizons $37h$ ($14h$ dans la journée) et $39h$ ($16h$ dans la journée). il est possible que cette contre-performance du modèle Darwin-Production à $14h$ et $16h$ vient du modèle météorologique ECMWF qui a tendance à être moins précis en milieu de journée.

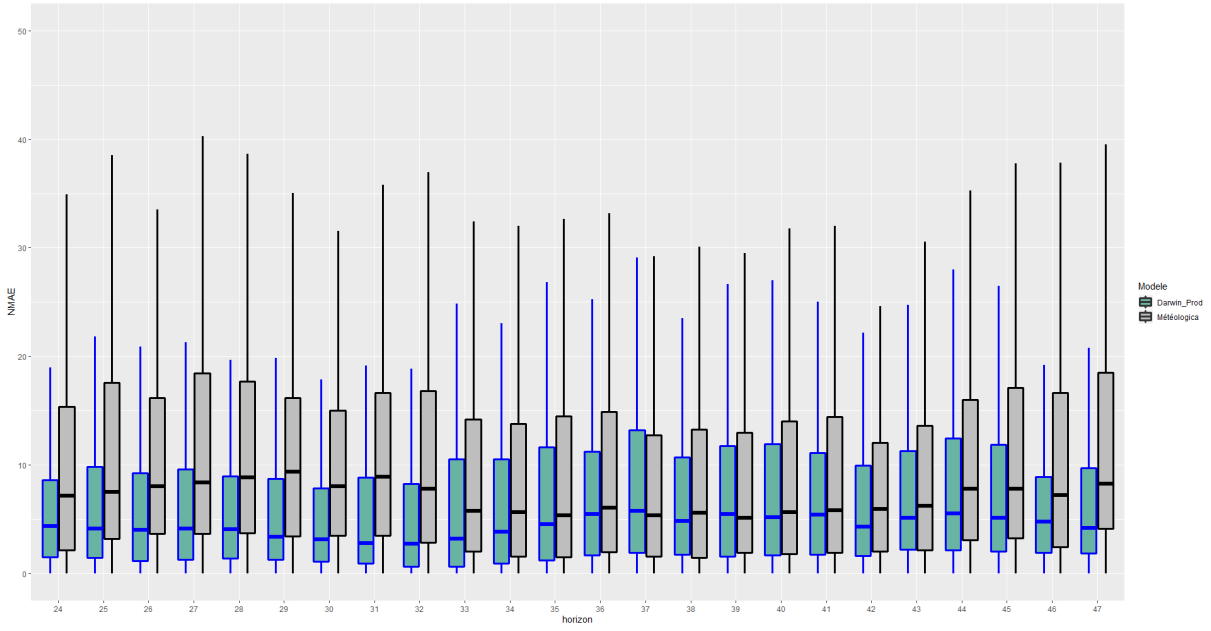


FIGURE 5.4 – NMAE par horizon de prévision.

Par contre la variabilité des prévisions (que nous allons analyser à la section suivante) du modèle Darwin est bien plus faible que la variabilité des prévisions du fournisseur.

Nous nous sommes aussi intéressés aux performances des deux modèles par mois, figure 5.5. Nous avons la même conclusion : le modèle Darwin-Production est meilleur que le modèle du fournisseur.

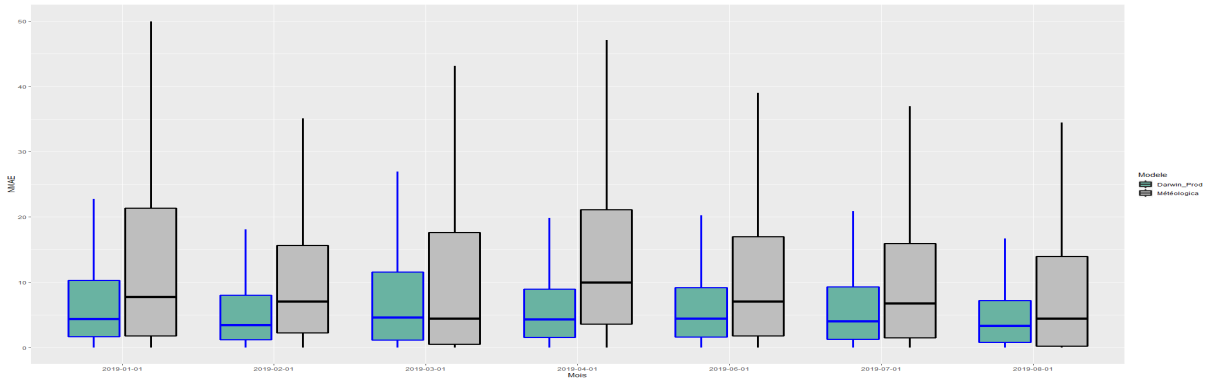


FIGURE 5.5 – NMAE par mois sur l'année 2019 en intégrant les périodes d'arrêt et de dysfonctionnement des éoliennes.

5.4.2 Comparaison des intervalles de prévision

Dans cette partie nous avons comparé l'intervalle de prévision à 80% du fournisseur de prévisions à l'intervalle de prévision à 80% étudié dans le chapitre 4, $MSPE2 + Spline$ et qui sera prochainement développé dans Darwin. Nous avons pris comme indicateurs statistiques les mêmes indicateurs de comparaison d'intervalle de prévision utilisés dans le chapitre 4. Il s'agit de la probabilité de couverture, de la longueur moyenne de l'intervalle de prévision et de l'indice de compromis entre la probabilité de couverture et la longueur moyenne de l'intervalle de prévision (voir chapitre 4, section 4.5.2).

TABLE 5.3 – Comparaison des intervalles de prévision.

Intervalle de prévision	Probabilité de couverture	Longueur moyenne	Indice
Darwin-Production	82%	5.20 Mw	0.68
Fournisseur	55%	4.5 Mw	0.52

Comme nous l'avions expliqué dans le chapitre 4, le critère d'un bon intervalle de prévision est d'avoir une probabilité de couverture élevée proche du niveau nominal et une longueur d'intervalle plus courte. Cependant, ces deux concepts s'opposent l'un à l'autre : les couvertures élevées ou faibles sont associées respectivement à des intervalles plus longs ou plus courts. C'est exactement ce qu'on observe sur le tableau 5.3. Darwin-Production à une

probabilité de couverture plus élevée et une plus grande longueur moyenne d'intervalle de prévision. Cependant Darwin-Production a une probabilité de couverture égale à 82%, légèrement supérieur à la valeur nominale (80%) alors que la probabilité de couverture du fournisseur (55%) est très inférieure à sa valeur nominale (80%). La longueur moyenne de l'intervalle de prévision du fournisseur est légèrement plus petite que celle du modèle Darwin-Production. Pour pallier cette opposition entre la probabilité de couverture et la longueur moyenne de l'intervalle de prévision, [Minkah and Wet, 2018] ont proposé un indice pour comparer des estimateurs d'intervalle de prévision. Ainsi cet indice montre que l'intervalle de prévision de Darwin-Production est meilleur que l'intervalle de prévision du modèle du fournisseur.
En définitive le modèle Darwin-production fournit de meilleures prévisions que le modèle du fournisseur.

5.4.3 Modèle Darwin-Production vs Modèle fournisseur dans un cadre opérationnel

Dans cette partie nous avons évalué les prévisions du modèle Darwin-Production et celle du fournisseur dans le cadre opérationnel. Nous avons extrait de la plateforme Darwin les prévisions du modèle Darwin-Production, du fournisseur et des productions réelles de sept parcs (déjà connectés à Darwin et pour lesquels ces données sont disponibles) sur la période du 01/01/2020 au 31/03/2020. Il s'agit de deux parcs situé dans le sud, trois parcs dans le nord et deux autres dans l'ouest (figure 5.6). Darwin-Production étant jusqu'en ce moment (04/2020) en développement, pour beaucoup de parcs, les données de prévisions du fournisseur, de Darwin-Production et de production réelle ne sont pas disponibles. C'est pourquoi nous avons effectué l'analyse sur sept parcs mais il en existe plus d'une centaine en France.

FIGURE 5.6 – Situation géographique des sept parcs.

Les résultats observés sont résumés dans le tableau 5.4. Sur cinq des sept parcs étudiés, Darwin-production a un NMAE plus faible. La production des parcs situés dans le nord semblent être plus difficiles à prévoir mais nous n'avons pas assez de parcs dont les données sont disponibles pour confirmer cette tendance. Une étude plus élargie avec une centaine de parcs bien répartis sur le territoire pourrait permettre d'évaluer les prévisions selon la position géographique (l'est, l'ouest, le nord, le sud ou le centre). Il faut noter que les erreurs de prévision des deux modèles sont trop élevées.

TABLE 5.4 – NMAE des sept parcs

Parcs éoliens	FRCRU	FRLDC	FRLPT	FRMIR	FRPTE	FRROQ	FRSVT
NMAE Darwin-Production	12.98%	14.45%	18.04%	15.43%	17.69%	13.38%	10.64%
NMAE Fournisseur	11.82%	12.83%	19.03%	17.07%	17.74%	15.79%	14.54%

La non-disponibilité de l'indice de fonctionnement pour filtrer les données

de production explique en partie cette hausse des erreurs de prévision. En effet, en observant la figure 5.7 qui donne les courbes de puissance des sept pars sur la période test, on note beaucoup de variabilités sur les courbes de puissance.

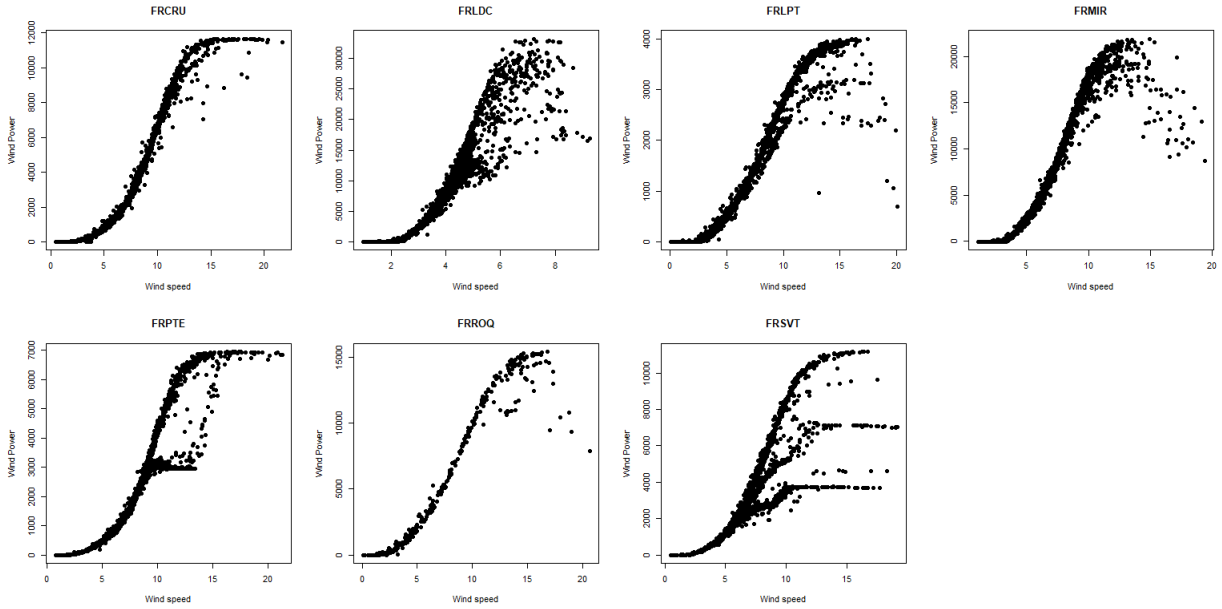


FIGURE 5.7 – Courbe de puissance des sept parcs sur les trois mois.

Ces variabilités sont dues à des dysfonctionnements, des problèmes de mesures et potentiellement des bridages (réduction de la capacité de production d'une éolienne pour certaines vitesses du vent ou durant certaines heures de la journée). Des contraintes internes font que l'indice de fonctionnement des éoliennes n'est plus disponible dans Darwin. Aucun filtre n'est fait sur les données de production réelle sur les trois mois de test. Les erreurs de prévision vont donc intégrer les arrêts machines (qui ne sont pas prévus par les modèles de prévision), les périodes de dysfonctionnement etc. C'est pourquoi les NMAEs obtenus sont largement au-dessus des NMAEs qui ont été obtenus avant l'industrialisation. Les équipes de Darwin travaillent sur une méthode de nettoyage générale. Nous avons proposé d'itérer des régressions isotoniques afin de détecter les résidus négatifs importants qui indiqueraient un palier donc un bridage.
Cependant, la courbe de puissance du parc "FRCRU" (figure 5.7) a moins d'incertitude que celle du parc "FRSVT" alors que le NMAE du modèle Darwin-Production est plus faible sur "FRSVT" que sur "FRCRU" et le NMAE du modèle fournisseur plus faible sur "FRCRU" que sur "FRSVT". Donc les incertitudes sur les courbes de puissance n'expliquent pas seules les erreurs de prévision.

Une autre explication des grosses erreurs de prévision est la taille de l'historique d'apprentissage. En effet, la collecte des données météorologiques par la plateforme Darwin étant ressente, il y a au mois de mars 2020 environ 5 mois d'historique. Donc puisque le modèle Darwin-Production est actualisé chaque dimanche, les prévisions du 01/01/2020 au 31/03/2020 sont issues de modèles entraînés avec un historique de 3 à 5 mois environ. Or un historique d'au moins un an est nécessaire pour tenir compte des effets saisonniers. Pour le modèle du fournisseur, nous ne disposons pas d'information sur l'historique d'apprentissage

5.5 Résultats attendus/observés sur l'industrialisation du modèle de prévision

Dans cette partie nous avons analysé les coûts du produit Darwin et l'apport potentiel de nos prévisions. Comme indiqué dans le contexte, demain les producteurs d'énergie renouvelable seront obligés de fournir des prévisions de production. Au moins pour continuer à assurer la stabilité du réseau électrique.
Comme indiqué dans la section 5.2.2, 313 parcs éoliens et de 138 centrales photovoltaïques sont connectés à Darwin pour une capacité de $6.1GWc$. Si nous devions acheter des prévisions externes pour tous ces actifs, cela nous coûterait 563 750 euros par an, en considérant un prix moyen de 1 250 euros/actifs/an pour les prévisions de production. Les prévisions internes calculées dans Darwin font baisser la facture de façon drastique : les coûts totaux pour le stockage, le traitement des données, l'apprentissage et la maintenance des machines ainsi que le coût lissé pour le développement devraient s'élever à 167 019 euros en 2020, soit trois fois moins que pour un service externe (tableau 5.5). Le gain net est donc de 396 731 euros (563 750 - 167 019) la première année. D'ici 2021, Engie a déjà obtenu +9 GW de projets éoliens et solaires. Au-delà, les projections de la Renewable Global Business Line (RGBL) prévoient une augmentation de capacité annuelle de +3 GW. Ainsi, avec un volume de 27,1 GWc, le gain net pourrait passer à 2 423 733 euros d'ici 2025. En outre, si nous décidions d'offrir ce service à des acteurs extérieurs, nous pourrions générer des revenus supplémentaires.

TABLE 5.5 – Projection sur les coûts et les gains du produit Darwin Forecaster

Year	2020	2021	2022	2023	2024	2025
Capacity WIND & SOLAR [GW]	6.1	15.1	18.1	21.1	24.1	27.1
Number of Assets	451	1 160	1 417	1 673	1 931	2 188
Cost for forecasts from provider (number assets * 1250euros/asset/an) [euros]	563 750	1 450 000	1 771 250	2 091 250	2 413 750	2 735 000
Cost for internal forecasts in Darwin [euros]	167 019	216 373	265 934	265 935	265 936	311 267
Cost/asset/year for internal forecasts in Darwin [euros]	370.3	186.5	187.7	159.0	137.7	142.3
Net gain [euros]	396 731	1 233 627	1 505 316	1 825 315	2 147 814	2 423 733

5.6 Conclusion

Ce chapitre est la concrétisation d'un point de vue industriel des travaux de cette thèse. Chaque jour, le modèle prévoit pour des parcs éoliens la production attendue pour un horizon allant jusqu'à dix jours. Le développement industriel du modèle de prévision éolienne dans le projet Darwin permettra de réduire les coûts d'intégration de l'éolienne dans le marché électrique. Les résultats des analyses ont montré que nous produisons de meilleures prévisions que le fournisseur actuel et par conséquent les coûts de l'intégration de l'énergie éolienne dans le système électrique peuvent être réduits en utilisant des prévisions en interne.

Conclusion générale

La prévision court terme de la production éolienne a été l'objet des travaux de cette thèse. Puisque les modèles météorologiques sont souvent utilisés dans le cadre de la prévision court terme, nous avons d'abord évalué les performances de deux grands modèles météorologiques (GFS et ECMWF) très souvent utilisés dans la littérature. Ainsi les résultats ont montré que les prévisions du modèle ECMWF ont moins d'erreurs. Nous avons décidé de travailler par la suite avec les données du modèle ECMWF même si une combinaison des deux modèles est une perspective. Mais nous ne disposions pas d'un historique suffisant de GFS.

Une première modélisation de la prévision de parcs éoliens a ainsi été faite en mettant l'accent sur l'approche directe qui consiste à prévoir directement la production à partir des données météorologiques et l'approche indirecte qui consiste à prévoir d'abord le vent à la hauteur des éoliennes à partir des données météorologiques et de convertir le vent prévu en prévision de production en se servant de la courbe de puissance réelle. En appliquant différents algorithmes d'apprentissage automatique, les résultats empiriques ont montré que si la courbe de puissance réelle est « lisse », la prévision du vent à la hauteur des éoliennes puis la transformation de ce vent en prévision de production est en général plus performante que la prévision directe de la production éolienne.

Nous avons ensuite cherché à améliorer les prévisions en intégrant les variations spatio-temporelles des prévisions météorologiques. En effet jusque-là, la production en un instant t était expliquée par les données météorologiques au même instant t. Nous avons estimé que la production en un instant t est potentiellement liée aussi aux variations des prévisions météorologiques dans le temps et dans l'espace. Ainsi, en tenant compte des variations spatio-temporelles, nous avons amélioré les prévisions de production d'environ 3%. La modélisation et la prise en compte de la turbulence du vent météorologique n'ont pas eu d'impact sur la prévision de la production.
Nous avons choisi au début des travaux d'utiliser les prévisions météorolo-

giques aux points de grilles (1, 4 et 16) les plus proches des parcs éoliens. On s'est donc posé la question d'un choix statistique des points de grille, basé sur la mesure d'importance des forêts aléatoires dans le cadre d'un groupe de variables. Ainsi nous avons montré qu'un choix des points de grille peut se faire de manière statistique mais qu'en général les points de grille les plus proches ont un poids plus important sur la prévision de la production.

Pour mieux prendre en compte les incertitudes liées à la prédiction de la production éolienne, les intervalles de prévision ont été développés dans le cadre d'un modèle à deux étapes. En effet, nous avons un modèle de prévision du vent à la hauteur des éoliennes d'une part et un modèle de conversion du vent prévu en prévision de production d'autre part. Le premier cité est une forêt aléatoire et le second une courbe de puissance réelle estimée par une spline. Nous avons donc deux sources d'incertitudes : une liée au modèle de forêt aléatoire et une autre liée à la courbe de puissance. Nous avons donc estimé et combiné ces deux incertitudes pour construire un intervalle de prévision associé aux prévisions de production éolienne.

L'objectif de cette thèse ayant été de mettre en place un modèle de prévision court terme de la production éolienne pour le groupe ENGIE Green, le modèle est développé dans la plateforme Darwin conçue pour améliorer les performances opérationnelles des parcs et développer la maintenance prédictive. Le service Darwin Forecaster de cette plateforme fournit les prévisions des parcs éoliens d'ENGIE Green. La comparaison des prévisions opérationnelles du modèle qui a découlé des travaux de cette thèse avec celles du fournisseur externe de prévision a montré que le modèle interne fournit généralement les meilleures prévisions.

Les travaux de cette thèse peuvent être poursuivis dans différentes directions. Nous avons choisi dans un premier temps les 16 points de grille les plus proches. Nous avons aussi développé une méthode de choix des points de grille par un algorithme de sélection. Cette méthode pourrait être élargie sur une zone plus large de point de grille pour détecter des clusters de points de grille qui expliqueraient mieux la production d'un parc.
Un autre aspect à exploiter et qui pourrait contribuer à une meilleure précision des prévisions éoliennes est l'utilisation des trajectoires du vent. En effet, les données des trajectoires du vent permettent de remonter dans le temps et de savoir à chaque instant le parcours du vent qui arrive sur un site éolien. Ainsi, des profils de trajectoires peuvent être identifiés et des modèles de prévision peuvent être calibrés selon ces profils.
Nous avons utilisé en entrée les données ECMWF pour la prévision éolienne.

Nous avons agrégé par poids exponentiels des modèles de GAM, SVM, forêts aléatoire mais étant donné qu'ils utilisaient les mêmes entrées, le modèle résultant de l'agrégation n'a pas donné de meilleurs résultats. En utilisant d'autres sources météorologiques et en ajustant un modèle sur chaque source, on pourra ensuite faire une agrégation de modèles qui pourrait améliorer les erreurs de prévision.

Une combinaison de l'approche statistique et l'approche probabiliste qui utilise les prévisions ensemblistes des modèles météorologiques est une direction dans l'amélioration des prévisions éoliennes. L'approche probabiliste n'est pas nouvelle dans la littérature mais à notre connaissance, sa combinaison avec l'approche statistique n'a pas été étudiée.

Annexe A

Annexes

A.1 Variables provenant des éoliennes

TABLE A.1 – Description des données provenant des éoliennes

Variables	Description
Date	Date de récupération de la donnée, heure locale
Vent	Vitesse du vent nacelle en m/s
P	Angle de pitch en degré (orientation de la nacelle)
Production	Production de l'éolienne en KW
State	codage des états de l'éolienne : 0,1=Emergency, 2=Pause, 3=Run
Etat	Etat de l'éolienne spécifique à chaque constructeur (Run/ Emergency/ Pause/ Menu service/ Ambiant)
TurbBrut	disponibilité de l'éolienne (l'éolienne fonctionne-t-elle ?)
TurbOK	correction de turbBrut
GridBrut	disponibilité du réseau électrique (l'éolienne peut-elle produire ?)
GribOK	correction gribBrut
Figee	S'il y a pas de variation du vent, du pitch, de la production, de la vitesse de rotation de la génératrice et de la température de la génératrice pendant 5min la donnée est considérée comme figée
Manquante	Lorsque, pour une éolienne, les données ne se suivent pas à une min près, on comble ce trou en dupliquant la donnée à la min précédente : données considérée comme manquante (et figée)
Vitesse génératrice	pour détection des données figées
Température génératrice	pour détection des données figées
Fonctionnement	Codé en 0 si arrêt et 1 si fonctionnement
Catégorie	A quatre niveaux par exemple niveau 1 fonctionnement ou arrêt, niveau 2 incohérent ou correct et niveau 3 compteur non ok

A.2 Variables provenant des pylônes de mesure

TABLE A.2 – Description des données provenant des Pylônes de mesure

Variables	Description
Date	Date et heure de la mesure, heure GMT
MIN_BT_0, MAX_BT_0, AVG_BT_0 et MET_BT_0	Minimum, maximum, moyenne et écart-type du voltage de la batterie
MIN_TI_0, MAX_TI_0, AVG_TI_0 et MET_TI_0	Minimum, maximum, moyenne et écart-type de la température au sol
MAX_3S_TI_0 et VALIDE_TI_0	Température interne de la centrale de mesure
MIN_CH_7, MAX_CH_7, AVG_CH_7 et MET_CH_7	Minimum, maximum, moyenne et écart-type de l'humidité à 7m au-dessus du sol
MIN_ST_7, MAX_ST_7, AVG_ST_7 et MET_ST_7	Minimum, maximum, moyenne et écart-type de la température à 7m au-dessus du sol
MIN_SP_i, MAX_SP_i, AVG_SP_i et MET_SP_i	Minimum, maximum, moyenne et écart-type de la Pression à i mètres au-dessus du sol
MIN_AA_i, MAX_AA_i, AVG_AA_i et MET_AA_i	Minimum, maximum, moyenne et écart-type de la vitesse du vent mesurée par l'anémomètre à i mètres au-dessus du sol
MIN_GG_i, MAX_GG_i, AVG_GG_i et MET_GG_i	Minimum, maximum, moyenne et écart-type de la direction du vent mesurée par la girouette à i mètres au-dessus du sol
..._AH_...	..._Vitesse horizontale_...
..._AV_...	..._Vitesse Verticale_...

A.3 Le test FEPA permettant de catégoriser les arrêts et fonctionnements des machines

TABLE A.3 – Le test FEPA

Formule	GMAO	Pitch OK	Etat	Alarme	Affectation
F1	0	1	Run	0	FO
F2	0	1	Run	1	FO
F3	0	1	Stop Emerg	1	ANI
F4	0	1	Stop Emerg	0	ANI
F5	0	1	Pause	1	ANI
F6	0	1	Pause	0	ANI
F7	0	1	Menu serv	1	ANI
F8	0	1	Menu serv	0	ANI
F9	0	1	Ambiant	0	ANI
F10	0	1	Ambiant	1	ANI
F11	0	0	Run	0	ANI
F12	0	0	Run	1	ANI
F13	0	0	Stop Emerg	1	ANI
F14	0	0	Stop Emerg	0	ANI
F15	0	0	Pause	1	ANI
F16	0	0	Pause	0	ANI
F17	0	0	Menu serv	1	ANI
F18	0	0	Menu serv	0	ANI
F19	0	0	Ambiant	0	FO
F20	0	0	Ambiant	1	FO
F21	1	1	Run	0	FO
F22	1	1	Run	1	FO
F23	1	1	Stop Emerg	1	AI
F24	1	1	Stop Emerg	0	AI
F25	1	1	Pause	1	AI
F26	1	1	Pause	0	AI
F27	1	1	Menu serv	1	AI
F28	1	1	Menu serv	0	AI
F29	1	1	Ambiant	0	AI
F30	1	1	Ambiant	1	AI
F31	1	0	Run	0	AI
F32	1	0	Run	1	AI
F33	1	0	Stop Emerg	1	AI
F34	1	0	Stop Emerg	0	AI
F35	1	0	Pause	1	AI
F36	1	0	Pause	0	AI
F37	1	0	Menu serv	1	AI
F38	1	0	Menu serv	0	AI
F39	1	0	Ambiant	0	FO
F40	1	0	Ambiant	1	FO

A.4 Catégorisation des états des éoliennes

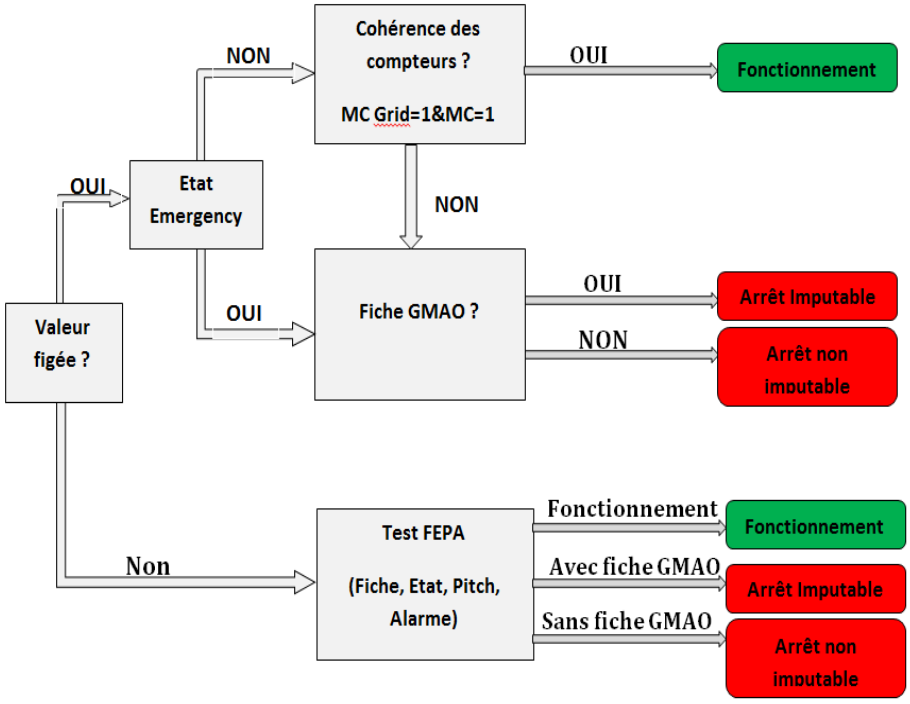


FIGURE A.1 – Catégorisation automatique du fonctionnement des éoliennes

Bibliographie

[Sim, 1996] (1996). *Smoothing methods in statistics*. 2nd ed. Springer.

[ENR, 2015] (2015). Panorama de l'électricité renouvelable en 2015. Technical report, Syndicat des Énergies Renouvelables.

[PN2, 2017] (2017). Panorama de l'électricité renouvelable au 31 mars 2017. Technical report, Syndicat des Énergies Renouvelables.

[FEE, 2019] (2019). *LA RÉGLEMENTATION EN FRANCE*. France énergie éolienne.

[Baïle, 2010] Baïle, R. (2010). Analyse et modélisation multifractales de vitesses devent. application à la prévision de la ressource éolienne. *Océan, Atmosphère. Université Pascal Paoli*.

[Bellanger et al., 2006] Bellanger, L., Baize, D., and Tomassone, R. (2006). L'analyse des corrélations canoniques appliquée à des données environnementales. *Revue de statistique appliquée, tome 54, no 4 (2006)*, pages 7–40.

[Bellman, 1957] Bellman, R. (1957). Dynamic programming. *Princeton : Princeton University Press*.

[Breiman, 2001] Breiman, L. (2001). Random forest. *Machine Learning*.

[Chakraborty and Pal, 2008] Chakraborty, D. and Pal, N. (2008). Selecting useful groups of features in a connectionist framework. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 19 :381–96.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*.

[Costa et al., 2008] Costa, A., Crespo, A., Navarron, J., Lizcano, G., Madsen, H., and Feitosa, E. (2008). A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*.

[Dione, 2019] Dione, M. (2019). Prévision de production éolienne par forêts aléatoires, agrégation et alerte de rampes. *In Proceedings of the 51es "Journées des Statistiques"*.

[Dione and Matzner-Løber, 2019] Dione, M. and Matzner-Løber, E. (2019). Short-term forecast of wind turbine production with machine learning methods : Direct and indirect approach. In Valenzuela, O., Rojas, F., Pomares, H., and Rojas, I., editors, *Theory and Applications of Time Series Analysis*, pages 301–315, Cham. Springer International Publishing.

[Efron, 1992] Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 83–127.

[Efron, 2013] Efron, B. (2013). Estimation and accuracy after model selection. *Journal of the Royal Statistical Society, (just-accepted)*.

[Fanglin, 2015] Fanglin, Y. (July 03, 2015). Comparison of forecast skills between ncep gfs four cycles and on the value of 06z and 18z cycles.

[Feng and Zhang, 2018] Feng, C. and Zhang, J. (2018). *Wind Power and Ramp Forecasting for Grid Integration*, pages 299–315.

[Ferreira et al., 2010] Ferreira, C., Gama, J., Matias, L., Botterud, A., and Wang, J. (2010). A survey on wind power ramp forecasting. *Argonne National Laboratory, Tech. Rep.*

[Frías-Paredes et al., 2016] Frías-Paredes, F., Mallor, F., León, T., and Gastón-Romeo, M. (2016). Introducing the temporal distortion index to perform a bidimensional analysis of renewable energy forecast. *Energy*, 94 :180–194.

[Frías-Paredes et al., 2017] Frías-Paredes, F., Mallor, F., León, T., and Gastón-Romeo, M. (2017). Assessing energy forecasting inaccuracy by simultaneously considering temporal and absolute errors. *Energy Conversion and Management*, 142 :433–546.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *In Proceedings of the 13th International Conference on Machine Learning*, pages 148–156.

[Fugon et al., 2008] Fugon, L., Juban, J., and Kariniotakis, G. (2008). Data mining for wind power forecasting. *European Wind Energy Conference & Exhibition EWEC 2008*, page 6.

[Fujimoto et al., 2019] Fujimoto, Y., Takahashi, Y., and Hayashi, Y. (2019). Alerting to rare large-scale ramp events in wind power generation. *IEEE Transactions on Sustainable Energy*, 10(1) :55–65.

[G. Giebel and Brownsword, 2003] G. Giebel, G. K. and Brownsword, R. (2003). The state-of-the-art in short-term prediction of wind power-a literature review. *[Online] : Available :*

http ://ecolo.org/documents/documents_in_english/wind-predict-ANEMOS.pdf.

[Gallego-Castillo et al., 2015] Gallego-Castillo, C., Cuerva-Tejero, A., and Lopez-Garcia, O. (2015). A review on the recent history of wind power ramp forecasting. *Renewable and Sustainable Energy Reviews*, 52 :1148 – 1157.

[Garcia and De-La-Torre-Vega, 2009] Garcia, A. R. and De-La-Torre-Vega, E. (2009). A statistical wind power forecasting system – a mexican wind-farm case study. *European Wind Energy Conference & Exhibition – EWEC Parc Chanot, Marseille, France*.

[Gastón et al., 2017] Gastón, M., Frías, L., Fernández-Peruchena, C., and Mallor, F. (2017). Assessing energy forecasting inaccuracy by simultaneously considering temporal and absolute errors. *AIP Conference Proceedings 1850*, pages 433–546.

[Gensler et al., 2016] Gensler, A., Sick, B., and Vogt, S. (2016). A review of deterministic error scores and normalization techniques for power forecasting algorithms. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–9.

[Genuer, 2010] Genuer, R. (2010). *Forêts aléatoires : aspects théoriques, sélection de variables et applications*. PhD thesis, Université Paris Sud 11.

[Genuer et al., 2010] Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31 :2225–2236.

[Ghaderi et al., 2017] Ghaderi, A., Sanandaji, M. B., and Faezeh, F. F. (2017). *Deep Forecast : Deep Learning-based Spatio-Temporal Forecasting*.

[Giebel et al., 2011] Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., and Draxl, C. (2011). The state of the art in short-term prediction of wind power a literature overview, 2nd edition. Technical report.

[Giebel and Kariniotakis, 2009] Giebel, G. and Kariniotakis, G. (2009). Best practice in short-term forecasting. a users guide. Technical report.

[Göçmen and Giebel, 2016] Göçmen, T. and Giebel, G. (2016). Estimation of turbulence intensity using rotor effective wind speed in lillgrund and horns rev-i offshore wind farms. *Renewable Energy.*, 99 :524–532.

[Gregorutti et al., 2014] Gregorutti, B., Michel, B., and Saint Pierre, P. (2014). Correlation and variable importance in random forests.

[Gregorutti et al., 2015] Gregorutti, B., Michel, B., and Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis.

[Gupta et al., 2016] Gupta, S., Shrivastava, N. A., Khosravi, A., and Panigrahi, B. K. (2016). Wind ramp event prediction with parallelized gradient boosted regression trees.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3) :389–422.

[Hasti and Tibshirani, 1986] Hasti, T. and Tibshirani, R. (1986). Generalized additive models prediction. *Statistical Science*.

[He et al., 2014] He, M., Yang, L., Zhang, J., and Vital, V. (2014). A spatio-temporal analysis approach for short-term forecast of wind farm generation. *IEEE Transactions on power systems*, 29(4) :1611–1622.

[He and Yu, 2010] He, Z. and Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational biology and chemistry*, 34 :215–225.

[Hotelling, 1936] Hotelling, H. (1936). Relations between two sets variables. *Biometrika*, 28 :321–377.

[Juban et al., 2008] Juban, J., Fugon, L., and Kariniotakis, G. (2008). Uncertainty estimation of wind power forecasts : Comparison of probabilistic modelling approaches. *European Wind Energy Conference & Exhibition EWEC*.

[Kariniotakis and Giebel, 2017] Kariniotakis, G. and Giebel, G. (2017). Wind power forecasting-a review of the state of the art. *Woodhead Publishing*.

[Kerns and Chen, 2014] Kerns, B. W. and Chen, S. S. (Received 5 SEP 2013, Accepted 13 MAR 2014). Ecmwf and gfs model forecast verification during dynamo : Multiscale variability in mjo initiation over the equatorial indian ocean. *Journal of Geophysical Research : Atmospheres, 10.1002/2013JD020833*.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22 :79–86.

[Kumar et al., 2017] Kumar, K., Kishtawal, C. M., and Pal, P. K. (2017). Theoretical and applied climatology. *Springer*.

[Kusiak et al., 2008] Kusiak, A., Zheng, H., and Song, Z. (2008). Wind farm power prediction : a data-mining approach. *Wind Energy*, 12(3) :275–293.

[Lange and Focken, 2008] Lange, M. and Focken, U. (2008). New developments in wind energy forecasting. *IEEE Power and Energy Society General Meeting 2008 - Conversion and Delivery of Electrical Energy in the 21st Century*, pages 1–8.

[Lenzi et al., 2017] Lenzi, A., Steinsland, I., and Pinson, P. (2017). *Benefits of spatio-temporal modelling for short term wind power forecasting at both individual and aggregated levels.*

[Lu and Hardin, 2017] Lu, B. and Hardin, J. (2017). Constructing prediction intervals for random forests.

[Madsen et al., 2005] Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H. A., and Nielsen, T. S. (2005). Standardizing the performance evaluation of shortterm wind power prediction models. *Wind Engineering*, 29(6) :475–489.

[Meinshausen, 2006] Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, page 983–999.

[Minkah and Wet, 2018] Minkah, R. and Wet, T. D. (2018). Comparison of confidence interval estimators : An index approach.

[Moreau and Glorieux-Freminet, 2019] Moreau, S. and Glorieux-Freminet, A. (2019). *Chifres clés des énergies renouvelables*. Commissariat général au développement durable.

[Najac, 2012] Najac, J. (2012). La prévision de production éolienne et photovoltaïque à edf. *Séminaire In'Tech, INRIA*.

[Nzobounsana and Gaymard, 2010] Nzobounsana, V. and Gaymard, S. (2010). Les analyses canoniques simple et généralisée linéaires : applications à des données psychosociales. *Math. and Sci. hum. / Mathematics and Social Sciences*, 1 :69–101.

[Persson, 2015] Persson, A. (2015). User guide to ecmwf forecast products. *ECMWF*.

[REN21, 2017] REN21 (2017). Renewables 2017 global status report. *Paris : REN21 Secretariat*.

[Sexton and Laake, 2009] Sexton, J. and Laake, P. (2009). Standard errors for bagged and random forest estimators. *Computational Statistics and Data Analysis*, pages 53 :801–811.

[Stone et al., 1984] Stone, C. J., Breiman, L., Friedman, J., and Olshen, R. (1984). Classification and regression trees. *Chapman & Hall/CRC*, pages 5–32.

[Tastu et al., 2011] Tastu, J., Pinson, P., Kotwa, E. K., Madsen, H., and Nielsen, H. A. (2011). Spatiotemporal analysis and modeling of shortterm wind power forecast errors. *Wind Energy*, 14(1) :43–60.

[Taylor et al., 2009] Taylor, J. W., McSharry, P. E., and Buiza, R. (2009). Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion.*, 24 :775–782.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58 :267–288.

[Wager et al., 2014] Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests : The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, pages 15 :1625–1651.

[Wang et al., 2011] Wang, X., Guo, P., and Huang, X. (2011). A review of wind power forecasting models. *Energy Procedia*, 12 :770–778.

[Wedam et al., 2009] Wedam, B., Garrett, Mcmurdie, L., and F. Mass, C. (2009). Comparison of model forecast skill of sea level pressure along the east and west coasts of the united states. *Weather and Forecasting*, 24.

[Yang, 2014] Yang, F. (2014). Review of gfs forecast skills in 2014. *Environmental Modeling Center National Centers for Environmental Prediction*.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68 :49–67.

[Zhang et al., 2019] Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*, 0(0) :1–15.

[Zhang et al., 2008] Zhang, H. H., Liu, Y., Wu, Y., and Zhu, J. (2008). Variable selection for the multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics*, 2 :149–167.

[Zhang et al., 2013] Zhang, J., Chowdhury, S., Messac, A., and Castillo, L. (2013). A multivariate and multimodal wind distribution model. *Renewable Energy*, 51 :436–447.

[Zhang et al., 2017] Zhang, J., Cui, M., Hodge, B.-M., Florita, A., and Freedman, J. (2017). Ramp forecasting performance from improved short-term wind power forecasting over multiple spatial and temporal scales. *Energy*, 122 :528 – 541.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Prévision court terme de la production éolienne par Machine learning

Mots clés : Apprentissage statistique, Énergie éolienne, Prévision court terme, Intervalle de prévision

Résumé : La loi de transition énergétique votée par l'État français a des implications précises sur les énergies renouvelables, en particulier sur leur mécanisme de rémunération. Jusqu'en 2015, un contrat d'obligation d'achat permettait de vendre l'électricité d'origine éolienne à un tarif fixe. À partir de 2015 certains parcs éoliens ont commencé à sortir de l'obligation d'achat. En effet, l'énergie éolienne commence à être directement vendue sur le marché par les producteurs à cause de la rupture des contrats d'obligation d'achat. Les gestionnaires de réseaux de distribution et les gestionnaires de réseaux de transport demandent ou même obligent les producteurs à fournir au moins des prévisions de production un jour à l'avance pour rééquilibrer le marché. Une surestimation ou une sous-estimation pourrait être exposée à des pénalités. Il existe donc un besoin énorme de prévisions précises.
C'est dans ce contexte que cette thèse a été lancée avec pour objectif de proposer un modèle de prévision de la production des parcs éoliens par apprentissage statistique. Nous disposons de données de production et de mesures réelles du vent ainsi que des données de modèles météorologiques. Nous avons d'abord comparé les performances des modèles GFS et ECMWF et étudié les relations entre ces deux modèles par l'analyse de corrélation canonique. Nous avons ensuite appliqué des modèles de machine learning pour valider un premier modèle de prévision par forêts aléatoires. Nous avons ensuite modélisé la dynamique spatio-temporelle du vent et l'avons intégrée dans le modèle de prévision ce qui a amélioré l'erreur de prévision de 3%. Nous avons aussi étudié la sélection de points de grille par une mesure d'importance de groupe de variables à l'aide des forêts aléatoires. Les intervalles de prévision par forêt aléatoire associés aux prévisions ponctuelles de la production des parcs éoliens sont aussi étudiés. Le modèle de prévisions découlant de ces travaux a été développé pour permettre au Groupe ENGIE d'avoir chaque jour ses propres prévisions pour l'ensemble de ses parcs éoliens.

Title : Machine learning for short term wind power forecasting

Keywords : Machine learning, wind power, Short term forecasting, prediction interval

Abstract : The energy transition law passed by the French government has specific implications for renewable energies, in particular for their remuneration mechanism. Until 2015, a purchase obligation contract made it possible to sell electricity from wind power at a fixed rate. From 2015 onwards, some wind farms began to be exempted from the purchase obligation. This is because wind energy is starting to be sold directly on the market by the producers because of the breach of the purchase obligation contracts. Distribution system operators and transmission system operators require or even oblige producers to provide at least a production forecast one day in advance in order to rebalance the market. Over- or underestimation could be subject to penalties. There is, therefore, a huge need for accurate forecasts.
It is in this context that this thesis was launched with the aim of proposing a model for predicting wind farms production by machine learning. We have production data and real wind measurements as well as data from meteorological models. We first compared the performances of the GFS and ECMWF models and studied the relationships between these two models through canonical correlation analysis. We then applied machine learning models to validate a first random forest prediction model. We then modeled the spatio-temporal wind dynamics and integrated it into the prediction model, which improved the prediction error by 3%. We also studied the selection of grid points by a variable group importance measure using random forests. Random forest prediction intervals associated with point forecasts of wind farm production are also studied. The forecasting model resulting from this work was developed to enable the ENGIE Group to have its own daily forecasts for all its wind farms.

Institut Polytechnique de Paris
91120 Palaiseau, France