Caractérisation de l'incertitude de production éolienne

Pierre Haessig

13 juillet 2011

Résumé

En prélude à de futurs travaux sur le dimensionnement et la gestion d'un système de stockage d'énergie couplé à un système éolien de production d'électricité, cette étude se focalise sur un problème sous-jacent : la *variabilité* de la production éolienne. Il s'agit de comprendre et de modéliser ce *caractère incertain* à partir de séries de données de terrains. Les deux résultats principaux sont *i*) la modélisation de l'incertitude de la relation entre la puissance produite par une éolienne et la vitesse du vent, en mode de fonctionnement à vitesse variable et *ii*) la modélisation d'une erreur de prévision au moyen d'un modèle conditionnellement hétéroscédastique.

Ces modélisations se sont faites *en statique*, sans prendre en compte l'aspect temporel. L'importance des phénomènes temporels suggère de les étudier prochainement.

1 Introduction

Avant de présenter le contexte et les objectifs de ce stage, je vais résumer le contexte du problème de l'intermittence des énergies renouvelables, en particulier éoliennes, car il justifie la présente étude centrée sur la "Caractérisation de l'incertitude de production éolienne".

1.1 Contexte général

Actuellement, plus de 80 % de la production mondiale d'électricité se fait dans des centrales thermiques (conventionnel ou nucléaire, cf figure 1). Cela pose plusieurs types de problèmes :

- consommation de ressources non renouvelables (charbon, hydrocarbures, uranium, ...)
- émission de substances toxiques (oxydes de soufre et d'azote, ...)
- émission de gaz à effet de serre (dioxyde de carbone, ...)

Pour ces raisons, les moyens basés sur des énergies dites "renouvelables" sont l'objet d'une grande attention à la fois du monde de la recherche, de l'industrie et de la société en générale. Ainsi *l'Umweltrat*[1] propose, dans un rapport d'expertise publié en janvier 2011 [10], plusieurs scénarios économiquement étayés visant à atteindre une production d'énergie électrique basée *en totalité* sur des ressources renouvelables à faible impact sur l'environnement à l'horizon 2050. À plus faible échelle, mais à plus proche échéance (2012) l'île d'El Hierro (Canaries) prévoit de produire 80 % de son électricité à partir du vent [7].

1. Comité fédéral consultatif pour la politique environnementale Allemande

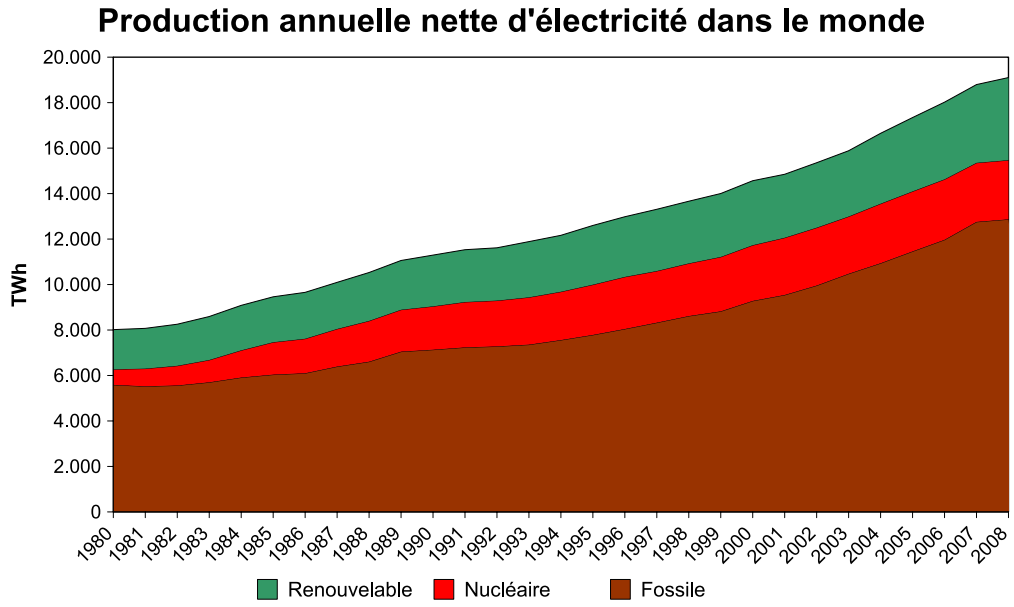

FIGURE 1 – graphique Wikipedia, d'après les données de `www.eia.gov`

1.1.1 Problème de l'intermittence

Les productions à base d'énergies renouvelables – du type éolien et solaire en particulier – sont qualifiées *d'intermittentes* car la puissance qu'elles produisent est fluctuante et difficilement prévisible. Du point de vue du gestionnaire de réseau (RTE pour la France) qui doit assurer à tout instant le difficile équilibre entre la production et la consommation électrique, ces sources d'énergies sont donc qualifiées de "non dispatchables [2]" et de "fatales".

Ce caractère fatal a conduit à plafonner le taux de pénétration des ces énergies intermittentes dans la production totale pour garantir le bon fonctionnement du réseau. Par exemple en France, un arrêté du 23 avril 2008 fixe la limite à 30 % de la puissance circulant sur le réseau à un instant donné. Au delà de cette limite, un producteur peut être déconnecté ce qui entraîne pour lui une perte de revenus.

1.1.2 Gestion de l'intermittence

Le caractère intermittent de la production éolienne ou photovoltaïque rend impossible son *utilisation seule* sur un réseau électrique. Une première façon de gérer l'intermittence est d'apporter un *complément* par des sources dispatchables mais non renouvelables telles que les génératrices diesel [3]. Une seconde solution est d'ajouter au réseau des moyens de stockage d'énergie pouvant absorber les pointes de production et compenser les creux. L'hybridation de ces deux techniques est bien sûr possible, voire nécessaire [7].

Les technologies de stockage Les différents moyens de stockage d'énergie sur le réseau se distinguent essentiellement par leur coût, leur capacité, leur puissance et leur temps de

2. qui ne peuvent pas être allumées et éteintes à volonté
3. ce qui alimente les critiques basés sur "les éoliennes augmentent les rejets de CO_2"

réponse. Sans entrer dans les détails, les principales technologies actuellement utilisés ou utilisables sont :

- *Stockage Gravitationnel*, dit aussi Stockage par Transfert d'Énergie par Pompage (STEP) où l'on transvase de l'eau par pompage-turbinage entre deux réservoirs d'altitudes différentes. Cela permet de réaliser les plus grandes capacités de stockage au coût le plus bas. Par contre le temps de réponse est mauvais.
- *Stockage Électrochimique* ("batteries"), utilisé généralement dans les appareils électriques portatifs, mais également utilisable sur les réseaux. Les caractéristiques dépendent de la technologie utilisée (Plomb-acide, Lithium-ion, ...)
- *Stockage Inertiel*, inspiré des technologies de centrifugation, où l'on fait varier la vitesse de rotation d'une masse cylindrique
- *Stockage Électrostatique*, c'est à dire basé sur des super-condensateurs, où l'on fait varier la tension pour extraire ou stocker des charges électriques.

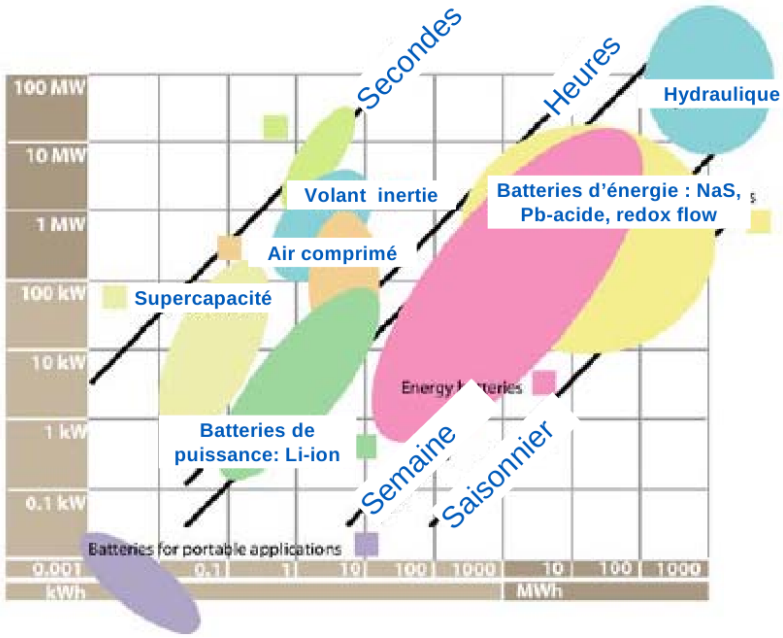

FIGURE 2 – Comparaison (Puissance, Capacité) des technologies de stockage. Source [15, p. 27]

La description précise de ces moyens n'est pas l'objet de ce rapport et un séminaire au Collège de France a récemment été consacré au sujet [15]. La figure 2 donne une comparaison des puissance et des capacités de chacun. On trouve d'autres éléments de comparaisons sur le site de l'Electricity Storage Association (`http://www.electricitystorage.org`). Vu la diversité des technologies il est nécessaire de bien comprendre les besoins de stockage, en particulier en terme de capacité, de puissance à échanger et de temps de réponse, Or les besoins de stockage sont directement liés à la variabilité de la ressource. Celle-ci se doit donc d'être étudiée et c'est précisément le sujet de ce stage.

1.2 Contexte institutionnel

En plus d'une dynamique globale favorable au stockage de l'électricité, on assiste en France à une reconnaissance de ces besoins par les acteurs institutionnels de l'électricité.

Appel d'offre éolien La Commission de Régulation de l'Énergie (CRE) a publié le 9 novembre 2010 un appel d'offre pour des installations éoliennes terrestres en Corse et Outre-mer dans l'objectif de dépasser la limite des 30 % d'énergie intermittente (cf. partie 1.1.1). L'appel d'offre spécifie un certain nombre de "services aux réseaux" que ces parcs devront assurer pour permettre de passer hors du cadre défini par l'arrêté. En particulier, le parc éolien devra pouvoir stocker de l'électricité.

D'après l'une des clauses principales de l'appel d'offre, la production de ces parcs devra être annoncée 24 heures à l'avance, par tranche de 30 minutes. L'annonce constituera un engagement contractuel qui devra être respecté avec une marge de $\pm$ 15 % et si la puissance produite sort de ces marges, le producteur sera pénalisé financièrement (électricité rachetée à la moitié de son prix).

Pour illustrer *l'impossibilité* de respecter ce cahier des charges sans stockage, on a représenté sur la figure 3 la production d'une éolienne sur 40 heures. La puissance, mesurée à chaque seconde, a été moyennée par pas de 30 minutes (tracé bleue). L'intervalle vert représente la marge de $\pm$ 15 % autorisée par l'appel d'offre [4] alors que la zone bleu clair indique l'étendue de la distribution de la puissance dans le pas correspondant. On peut constater que la puissance produite sort souvent de la marge autorisée, principalement autour de l'heure 35. Un lissage de la production est donc nécessaire.

Expérimentations industrielles Les technologies rapidement présentées à la fin de la partie 1.1 sont actuellement l'objet de plusieurs mises en oeuvre industrielles. En 2010, EDF a par exemple installé à La Réunion une batterie Sodium Soufre de 1 MW / 7,2 MWh [15] dans le cadre d'un projet pilote. Sur la base d'un financement du Département américain à l'Énergie (DoE), un système de stockage par volants d'inertie vient d'être inauguré dans l'état de New York (juin 2011), pour une puissance de 20 MW.

Dans des scénarios à long terme impliquant des quantités massives de stockage [8], il s'agit d'utiliser les pays disposant d'un grand potentiel hydroélectrique, tels que la Norvège et la Suisse. Ce stockage *centralisé* nécessiterait cependant la construction de grandes interconnexions par opposition avec le stockage par batterie ou volant d'inertie qui est beaucoup plus *distribuable*.

1.3 Contexte académique

Ce paragraphe est un rapide tour d'horizon des activités scientifiques liées au problème de l'intermittence des énergies renouvelables et en particulier des éoliennes.

1.3.1 Prévision de la ressource

Un des points clés de la maîtrise de la variabilité est la prévision de la ressource éolienne à *brève échéance* (à un ou deux jours). Ces prévisions se basent sur des modèles

4. en se plaçant dans un cas idéal où l'on aurait pu prévoir 24h à l'avance *sans erreur* la production moyenne !

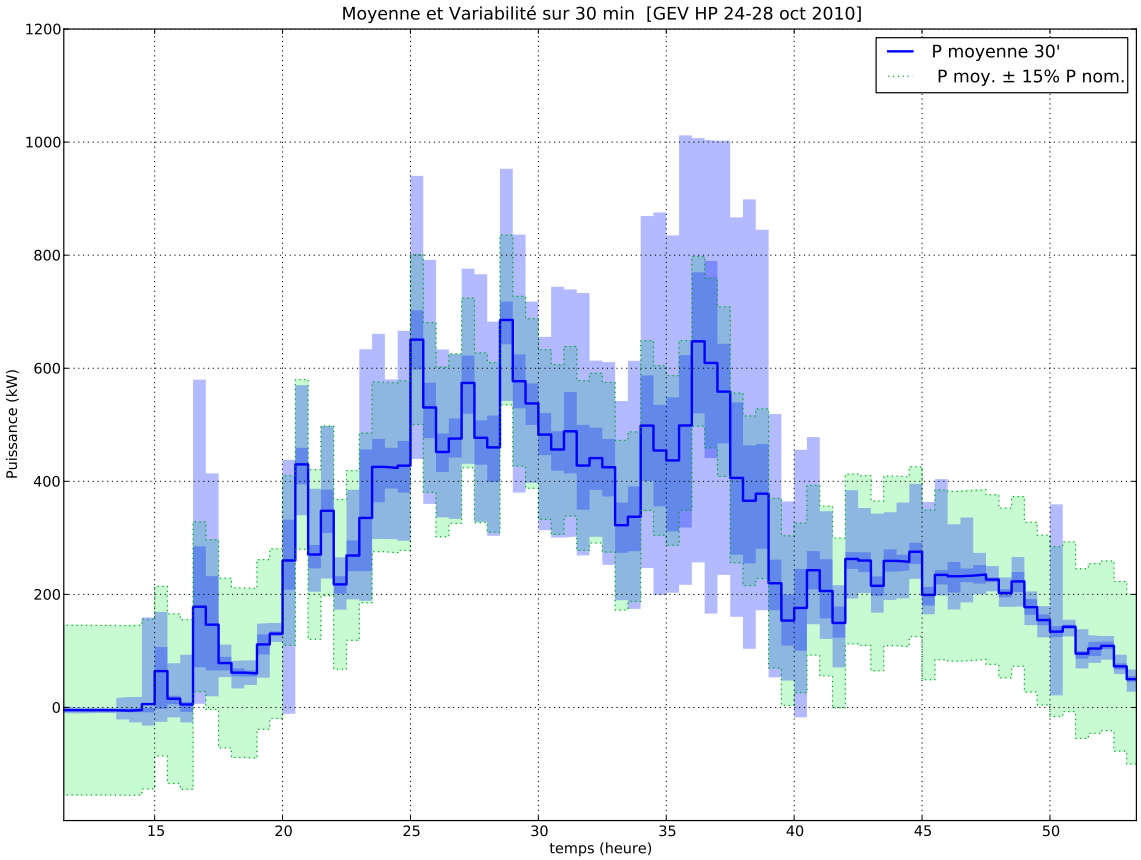

FIGURE 3 – Illustration de la variabilité de production éolienne

météorologiques [5] qui prédisent la vitesse et la direction du vent dont on peut déduire la puissance produite. Au moins deux problèmes compliquent le travail de prévision de la puissance :

- le maillage spatial de ces modèles est trop lâche ce qui masque les effets topologiques locaux.
- la relation entre la vitesse du vent et la puissance produite est complexe (cf. figure 7), et souvent mal connue.

Il faut donc nécessairement une méthode de *correction et de raffinement* de la prédiction, qui se base généralement sur une analyse des historiques d'observations (vent et/ou puissance produite) issues du site étudié. Une grande variété de méthodes sont ou ont été employées pour ce raffinage et de très nombreuses publications sont parues à ce sujet. Les méthodes les plus populaires sont non linéaires, non paramétriques, du type "boîte grise" avec par exemple l'utilisation de réseaux de neurones artificiels (ANN).

Un état de l'art poussé de la prédiction du vent à court terme se trouve dans les délivrables du projet ANEMOS.plus [11]. L'introduction du rapport (pages 4 – 25) donne une bonne vision de ce qu'est la prévision à court terme et de ses enjeux. Une description plus historique (sur quatre décennies) des études sur la prédiction a été publiée en 2008 [6].

5. à part pour des horizons courts (inférieurs à 3 – 6 h) où des modèles purement statistiques peuvent suffire

Projets ANEMOS La recherche européenne sur le sujet s'est réunie dans le cadre de deux projets européens de recherche : ANEMOS (2002 – 2006) et ANEMOS.plus (2008 – 2011). Cest deux projets sont pilotés par l'équipe Énergies Renouvelables du Centre Énergétique et Procédés (CEP, MINES ParisTech, situé à Sophia Antipolis) et incluent des partenaires essentiellement originaires des grands pays éoliens européens : Danemark, Allemagne et Espagne.

Qualité de la prévision a posteriori Un des grands objectifs du projet ANEMOS a été de définir des normes pour *évaluer la qualité* de la prévision. Il y a en effet de nombreuses façons de comptabiliser les erreurs de prédiction (erreur quadratique moyenne, en valeur absolue, ...). De plus la qualité des prévisions dépend beaucoup de la topologie du lieu, selon que le terrain est plus ou moins accidenté. Par exemple, en terme d'erreur NMAE [6], on peut passer de 10 % pour un terrain dit facile à 30 % pour un terrain dit compliqué [18]. Les partenaires du projet ont donc publié des *protocoles d'évaluation* définissant les critères d'erreur à employer ainsi que les "fermes type" sur lesquelles il faut mesurer la performance[17]. Ils rappellent également l'importance de *séparer* les données utilisées pour l'apprentissage du modèle des données utilisées pour tester sa qualité.

Qualité de la prévision en ligne Alors qu'au début des années 2000, les outils de prévisions fournissaient uniquement des *prédictions ponctuelles*, il est apparu le besoin d'accompagner ces prévisions d'un *intervalle de confiance*. Ce n'est pas une tâche évidente car les caractéristiques statistiques de l'erreur varient selon la situation météorologique et selon le niveau de puissance prévu. On ne peut donc pas utiliser directement les outils d'inférence statistique classique [5, chap. 3] liés à l'hypothèse gaussienne.

La *régression de quantile* [14] introduite par Koenker en 1978 pour l'économétrie a été utilisée par Nielsen [19] pour prédire l'incertitude à partir d'une régression sur des archives de prévisions et de mesures. Son modèle d'incertitude dépend principalement du niveau de puissance prévu. La méthode dite de *rééchantillonage* a également été utilisée [22]. Dans les deux cas il s'agit de modélisations non paramétriques, c'est à dire sans a priori sur la forme de la distribution de l'erreur. Le mémoire de thèse de Pierre Pinson [20] au CEP porte spécifiquement sur l'incertitude de prédiction, et son article de 2007 pose le cadre de la prévision probabiliste de l'énergie éolienne [24].

L'information d'incertitude est cruciale dans plusieurs contextes :

1. trading d'électricité éolienne [21], pour permettre au trader d'évaluer son degré d'exposition au risque,
2. le dimensionnement du stockage, si le stockage doit absorber les erreurs de prévisions (cf. l'appel d'offre de la CRE, partie 1.2) et
3. la gestion en ligne du stockage, pour choisir une stratégie minimisant les pénalités.

Ces deux derniers points nous intéressent plus particulièrement. De façon très intéressante Pierre Pinson a jeté un pont entre prévision de la ressource et stockage en appliquant sa méthode [23] au dimensionnement d'un stockage [25].

Du côté des méthodes paramétriques, des modèles du type GARCH, inspirés de la finance pourraient peut-être aussi servir à modéliser l'incertitude grâce à leur hétéroscédasticité conditionnelle. La modélisation de la moyenne et de la volatilité du vent a été introduite par Ewing en 2006 [9]. Malheureusement, il reste un pas pour passer du vent

6. erreur quadratique moyenne normalisée

à la puissance produite. Ce pas qui exige une transformation non-linéaire n'a pas, à ma connaissance, encore été franchi.

Metnext En Guadeloupe et à la Réunion, l'exploitant de parcs éoliens Aérowatt fournit au gestionnaire de réseau (EDF SEI) des prévisions de production issues d'un système de prévision fourni par la société Metnext, filiale de Météo France. La méthode employée pour la prévision est une Régression Adaptative de Splines Multivariée (méthode MARS). La prévision est générée par une famille de fonctions linéaires par morceaux, avec un réglage des paramètres du modèle basé sur un apprentissage des données historiques. Il est malheureusement impossible de connaître les détails du système car la prévision est un produit commercial.

1.3.2 Gestion et dimensionnement du stockage

Le problème de dimensionnement et de gestion d'un stockage n'est pas nouveau. Il déjà été abordé au laboratoire SATIE, par exemple par Judicaël Aubry [2] pour le lissage de la production issue d'un houlo-générateur [7].

Dans le contexte éolien, les travaux de Sercan Teleke, à North Carolina State University ont introduit certaines notions, en particulier le problème du critère de performance [31], qui doit mesurer la qualité de la production en sortie. Cependant son choix du critère est un peu arbitraire et Aubry montre que d'autres choix sont possibles, en particulier pour mesurer la norme de l'erreur (erreur quadratique, ou bien en valeur absolue, ...).

Teleke a mis en œuvre un mécanisme de gestion basé sur des règles simples [30] pour garantir que la batterie reste dans ses limites de fonctionnement (limites de charges, limite de courant). Dans un autre article [29], il a appliqué des techniques issues de la *commande optimale* en Automatique. Dans les deux cas, son principe de dimensionnement du stockage est empirique, par opposition à Pinson [25].

Un problème clé pour le stockage, en particulier le stockage par batteries est le *vieillissement* or cette question a malheureusement été laissée en suspens par Teleke, alors même que ses résultats montrent une très forte sollicitation du stockage. À l'inverse, Aubry met en évidence un *optimal de capacité* vis-à-vis du coût du stockage, lorsque le *cycle de vie* est pris en compte. Cependant, son travail est basé sur une technologie différente (super-condensateurs). Une prise en compte du cyclage et de l'incertitude a été proposée récemment par Li [16] mais la stratégie de commande proposée semble contre-intuitive car elle mène à un cyclage forcé.

Par ailleurs, un traitement à la fois plus abstrait du point de vue du contrôle et plus appliqué au niveau de la description des convertisseurs de puissance électroniques a été effectué à Supélec (au LSS et au LGEP) par Antonio Sánchez [28] en se basant sur le concept de Dynamic Energy Router. Ce concept formalise les échanges de puissance entre plusieurs systèmes capables d'échanger de l'énergie par l'intermédiaire du bus continu. Un des points d'intérêt du travail de Sánchez est *l'hybridation* des sources d'énergie de part le mélange d'une pile à combustible (temps de réponse long) avec des condensateurs (temps de réponse court).

Des travaux ont aussi été effectués à McGill University, Montréal par Chad Abbey [1], où il étudie l'association de productions Diesel et éolienne avec un stockage pour augmenter la part moyenne d'énergie éolienne. La littérature contient beaucoup d'autres

7. système de récupération de l'énergie des vagues pour la convertir en électricité

travaux sur la gestion et le dimensionnement du stockage. Il reste donc encore un grand travail de lecture à effectuer.

1.3.3 Problème d'échelle temporelle

Le choix de l'échelle temporelle de travail, c'est à dire de la période d'échantillonnage des données, est un autre point important du problème de gestion de l'intermittence. Il faut préciser que les études précédemment citées ont été faites sur la base de mesures de terrain qui sont généralement des moyennes par pas de 10, 15, 30 ou souvent 60 minutes.

Ces périodes sont beaucoup trop grandes pour faire face à des problèmes de stabilité du réseau qui peuvent apparaître en quelques dizaines de secondes. En conséquence, le cahier des charges de la CRE (cf. partie 1.2) précise que des mesures de puissance peuvent être effectuées sur des fenêtres temporelles pouvant descendre à 10 secondes. C'est donc une fréquence de rafraîchissement 100 fois plus importante que celle des mesures usuelles.

Comme les publications apparues dans la littérature se basent sur des moyennes de plusieurs minutes, faute de mieux, les stratégies résultantes sous-estiment les besoins de stockage, en particulier son vieillissement car les fluctuations haute fréquence sont invisibles. Il est donc impératif d'obtenir et d'utiliser des données de *haute résolution temporelle*. La collaboration avec l'entreprise Vergnet (cf. 1.4.1) permet un grand pas dans ce sens.

Enfin, le fait que les fluctuations de production se situent sur d'aussi larges échelles temporelles appelle à *l'hybridation des techniques de stockage*. Par exemple, aux îles Canaries [7] le stockage principal (hydraulique) sera complété par des batteries Lithium et des volants d'inertie pour compenser la lenteur des actionneurs du système hydraulique.

1.4 Contexte et objectifs du stage

Cette courte étude intitulée "Caractérisation de l'incertitude de production éolienne" s'est effectuée dans le cadre d'un stage de Master 2 Automatique & Traitement du Signal au Laboratoire des Signaux & Systèmes (LSS) sous l'encadrement de Pascal Bondon.

Ce stage se situe en prélude à des travaux de thèse plus orientés Génie Électrique qui démarreront en septembre 2011 au laboratoire SATIE de l'ENS Cachan à Ker Lann sous l'encadrement de Bernard Multon et Hamid Ben Ahmed. Ce stage marque également le démarrage d'une collaboration avec des industriels intéressés par les problématiques de l'intermittence éolienne.

1.4.1 Partenaires Industriels

Le laboratoire SATIE a été contacté par EDF SEI [8] alors que l'appel d'offre de la CRE pour l'outre-mer était sur le point de paraître. Pour pouvoir travailler sur des données de terrains, des partenariats (accords de secret) ont été établis débuts 2011 avec deux grands acteurs de l'éolien en outre-mer :

- *Vergnet*, qui conçoit et fabrique des éoliennes spécifiquement adaptées au marché de l'outre-mer car elles peuvent résister aux cyclones.
- *Aérowatt*, qui conçoit et exploite des parcs éoliens. En particulier en outre-mer, ce sont des éoliennes Vergnet qui sont utilisées.

8. Systèmes d'Énergie Insulaire, qui gère l'électricité en Corse et en outre-mer

Ces deux entreprises qui sont concernées au premier chef par l'appel d'offre de la CRE réfléchissent aux mécanismes de stockage à mettre en œuvre pour pouvoir satisfaire le cahier des charges. Dans le cadre de ces réflections, elles ont acceptées de partager des données qu'elles possèdent. Nous les en remercions vivement, car ces données sont à la base de toute cette étude. On va voir maintenant qu'elles sont de nature très complémentaire.

Données Vergnet Le fabriquant Vergnet dispose de sites éoliens pour tester et valider ses machines. En particulier, un prototype du modèle de plus haute puissance (GEV HP)) est installé à Greneville-en-Beauce (45). Cette machine dispose d'une instrumentation plus poussée qu'une éolienne standard. Cela a permis de recueillir des mesures de puissance et de vitesse de vent à *haute résolution temporelle*. Le pendant malheureux de cette haute résolution est la courte durée de ces acquisitions (quelques jours maximum).

Comme expliqué dans la partie 1.3.3, une observation des fluctuations rapides de la puissance est cruciale pour bien maîtriser le dimensionnement et la gestion d'un stockage. Plus de détails sur les données Vergnet sont donnés au paragraphe 2.1.

Données Aérowatt L'exploitant Aérowatt a accepté de partager des mesures issues de ses fermes éoliennes d'outre-mer (Guadeloupe et La Réunion). Ces données sont des archives de production des dernières années. La longue durée de l'acquisition les rend plus représentatives que les données Vergnet, par contre la période d'échantillonnage est au standard industriel de 10 minutes.

Le deuxième point d'intérêt de ces données est la présence concomitante de prévisions Metnext (cf. partie 1.3.1) que l'on peut comparer a posteriori aux mesures de production. Cette comparaison est l'objet de la partie 3.

1.4.2 Objectifs à moyen terme

Au-delà des trois mois de ce stage, nous avons cherché à garder à l'esprit les objectifs à plus longue échéances des futurs travaux de thèse en Génie Électrique. On rappelle rapidement ici que l'on souhaite :

- Développer une méthodologie du dimensionnement des besoins en stockage, en terme de capacité et de puissance.
- Développer une méthodologie pour la commande optimale du stockage (stratégie de gestion du stockage).
- Comprendre l'interaction entre les problèmes de dimensionnement et de gestion.

Ces objectifs s'inscrivent dans le cadre de la gestion de l'intermittence des énergies renouvelables, en particulier éoliennes (cf. parties 1.1.1 et 1.1.2) qui intéressent particulièrement l'équipe du laboratoire SATIE à Ker Lann.

1.4.3 Objectifs du stage

Le contenu de cette étude est le fruit de l'interaction entre *i*) des objectifs fixés a priori par anticipation des besoins des futurs travaux de thèse, *ii*) la disponibilité des séries de données sur lesquelles reposent ce travail *iii*) l'analyse de l'état de l'art dans les différents domaines liés à nos problèmes.

L'objectif général de l'étude est la *modélisation de la variabilité* de la puissance produite par une éolienne. En particulier on souhaite construire des Intervalles de Confiance (IC)

autour d'une régression sur les données. Les variables explicatives considérées sont, suivant la série de données, soit la vitesse du vent, soit une prévision de la puissance produite.

Il aurait été intéressant de *décomposer la variabilité* en une composante due à l'erreur de prévision météo et une composante due aux dynamiques non modélisées des éoliennes. Cette décomposition nécessite d'avoir des prévisions de vent accompagnées des mesures de vent correspondantes or nous n'avons pas ces données sous forme exploitable [9].

Au final, l'étude esquisse cette décomposition par sa structure :

- *Partie 2 :* modélisation de l'incertitude de la relation $P(v)$ avec des acquisitions à haute résolution temporelle pour une éolienne.
- *Partie 3 :* modélisation de l'incertitude entourant la prédiction de la puissance produite pour une ferme d'éoliennes.

La première partie s'intéresse donc à l'incertitude liée à la dynamique de l'éolienne alors que l'incertitude dans la deuxième partie est assez certainement due à des erreurs sur la prévision météorologique.

2 Données de puissance et de vent à haute résolution

Cette partie de l'étude se base sur les données de puissance et de vent fournies par Vergnet (cf partie 1.4.1). Ces données ont été acquises avec une grande résolution temporelle, bien plus que ce qui est usuellement disponible. L'objectif est de modéliser l'incertitude de la relation $P(v)$.

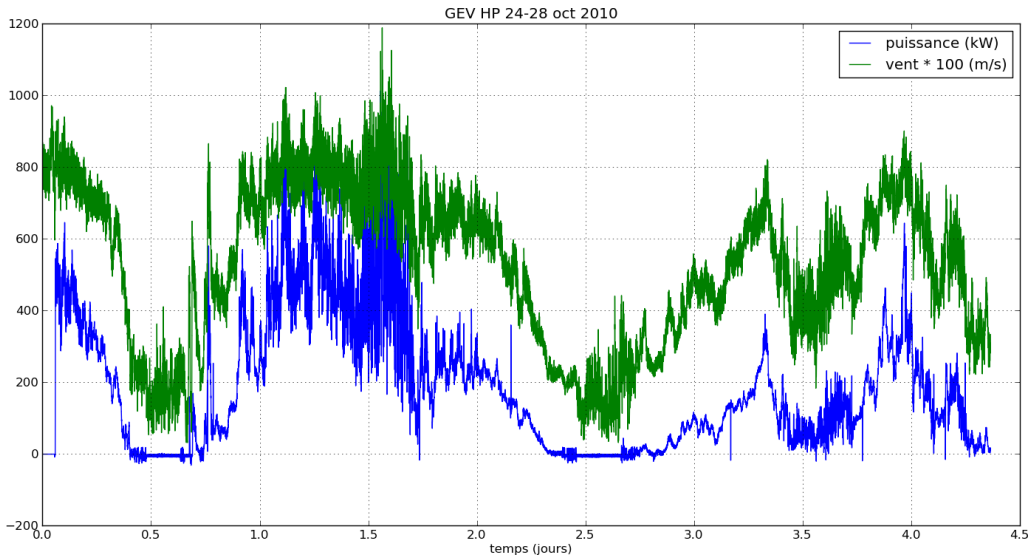

FIGURE 4 – Tracé temporel de la puissance et du vent

2.1 Présentation des données

On dispose de $N = 377169$ points de mesures de deux variables : la vitesse du vent v (m/s) et la puissance P (kW) produite par l'éolienne "GEV HP" du fabriquant Vergnet, installée sur son site d'essai à Greneville-en-Beauce (45). Sa puissance nominale est de 1

9. données sous forme d'archives multi-fichiers qui doivent encore subir plusieurs traitements

MW. Les données recueillies sont la moyenne sur une seconde à chaque seconde des grandeurs P et v. L'acquisition des N points correspond à environ 4 jours. Le tracé temporel est fait sur la figure 4.

Ces grandeurs sont issues, pour v, de l'anémomètre de la nacelle et pour P des mesures de tensions et de courant électriques. Elles sont acquises initialement par l'automate de commande de l'éolienne qui les expose sur ses sorties analogiques. Ensuite un "datalogger" Multitrend® SX de Honeywell se charge de la renumérisation et de l'enregistrement. Enfin, les données sont rapatriées à distance en un fichier au format CSV sur un ordinateur situé dans les locaux de Vergnet, près d'Orléans. Ce rapatriement se fait par intervalles de quelques jours (ce qui correspond à quelques mégaoctets de données). L'enchaînement successif des traitements a d'ailleurs été la source de problèmes.

2.1.1 Problèmes des données

Premièrement, au niveau de la *précision des mesures*, les grandeurs ont subi deux numérisations puis une conversion à une écriture décimale tronquée. Cependant, l'effet semble limité. Deuxièmement, au niveau de la *précision temporelle*, l'enregistreur Honeywell sauvegarde par défaut les données avec une méthode brevetée dite "Fuzzy Logging" qui économise de l'espace disque en n'enregistrant que les échantillons du signal lorsqu'il y a une variation significative. Cela se traduit par une perte d'échantillons, qui sont remplacés silencieusement par l'échantillon précédent lors de l'export CSV. Une observation à l'échelle de la seconde met alors en évidence un effet de marche d'escalier dû à ces répétitions d'échantillons et toute analyse temporelle fine (autocorrélation) est rendue impossible.

2.1.2 Observation des données

Le tracé de la figure 4 donne un aperçu global des données dans le domaine temporel au fil des quatre jours de mesure. Les deux grandeurs que sont la vitesse du vent v et la puissance électrique produite P sont bien sûr liées par la physique qui régit la machine (cf partie 2.2.2). C'est cette relation $P(v)$ qui va nous intéresser, et en particulier les variations autour d'un modèle moyen. L'importance de ces variations dans le domaine temporel est illustrée sur la figure 3 où les données sont agrégées par périodes de 30 minutes.

L'énergie totale produite par l'éolienne sur la période de mesure est

$$E_{prod} = \sum_i P_i.T \approx 19{,}0\ \text{MWh} \tag{1}$$

Pour information, avec un tarif de rachat [10] d'environ 0,08 €/kWh, cette énergie produite sur 4 jours représenterait environ 1500 € de chiffre d'affaire dans un contexte de production.

La caractérisation des données par des statistiques globales (moyenne, quartiles, ...) ainsi que par des histogrammes (figure 5) montre une grande différence entre d'un côté le vent dont la distribution est plutôt *symétrique* et de l'autre la puissance produite qui est *nettement asymétrique*. De plus, la puissance présente un pic de valeurs autour de zéro, lorsque l'éolienne est arrêtée faute de vent suffisant fort.

Par ailleurs, on relève que la valeur minimale de la puissance produite est négative ($P_{min} = -32\,\text{kW}$). Cela se produit lorsque l'éolienne est connectée au réseau alors que le

10. tarif éolien fixé par l'arrêté du 13 décembre 2008

	Min.	1er Q.	Méd.	Moy.	3ème Q.	Max.
Vitesse (m/s)	0,313	3,606	5,530	5,284	7,076	11,884
Puissance (kW)	-32,1	27,2	134,7	181,6	279,2	1011,9

vent est faible. Il s'agit de phases transitoires dont l'impact est plutôt faible ($\sum_{P_i<0} P_i.T \approx -83$ kWh). Le rapport entre cette énergie auto-consommée et celle produite au total est donc inférieur à 0,5 %.

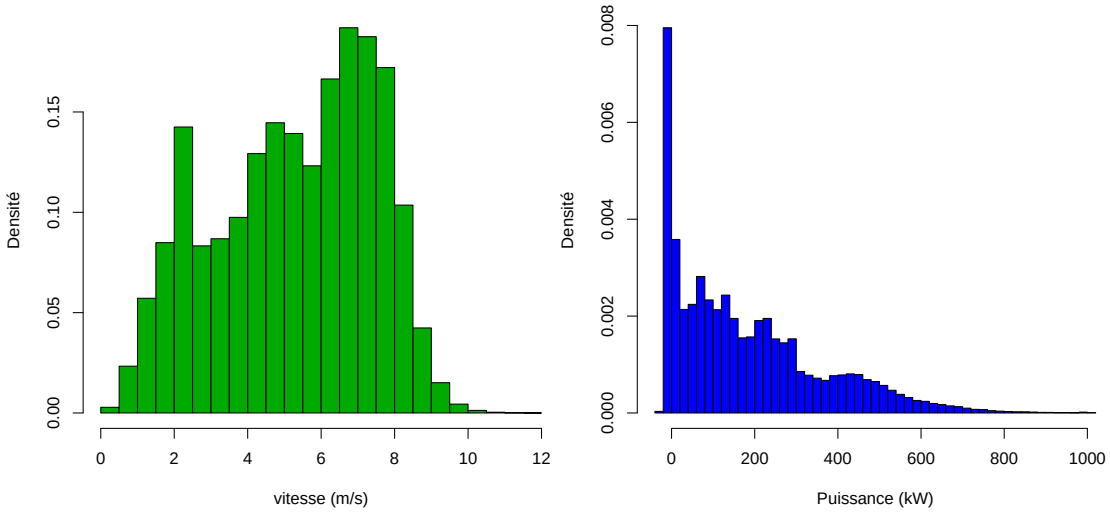

FIGURE 5 – Histogramme des mesures de vent et de puissance

2.2 Modélisation des données

On va à présent s'intéresser à la relation entre P et v d'un point de vue *statique*, sans se préoccuper de l'aspect temporel. Plus précisément, on va vouloir expliquer chaque mesure P_i à l'aide de la mesure v_i correspondante.

$$P_i \approx f(v_i) \tag{2}$$

C'est une approximation raisonnable par le principe même d'une éolienne : elle doit collecter la puissance cinétique du vent. La puissance produite dépend donc de la vitesse du vent. Cependant, ce modèle fait l'impasse sur d'autres variables explicatives telles que la température de l'air. De plus, ce modèle considère "LA" vitesse du vent alors que celle-ci n'est pas uniforme, en particulier elle augmente en moyenne avec l'altitude.

2.2.1 Modélisation statistique

Pour modéliser la relation (2) on considère les mesures $(P_i,v_i)_{i\in[1,N]}$ sous forme de couples de variables où P_i est une variable aléatoire à expliquer (ou variable *endogène*) et v est la variable explicative, ou *exogène*. La figure 6 montre la carte de densité [11] de ces couples de réalisations. La représentation par carte de densité permet d'éviter l'effet de

11. un histogramme bidimensionnel représenté comme une image

"saturation" obtenu avec une représentation classique en nuage de points le nombre de points $N = 377169$ étant grand.

Le choix de la fonction f va être inspiré d'une *analyse physique du système*, ou plus précisément une analyse issue de la mécanique des fluides.

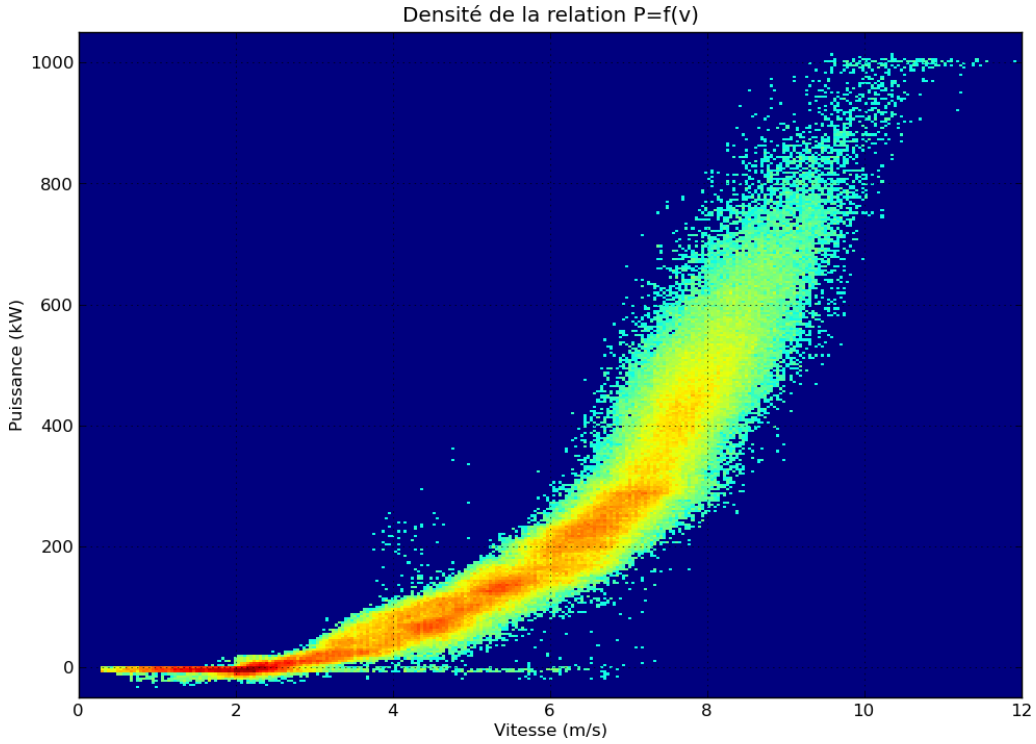

FIGURE 6 – Carte de densité des couples (P_i,v_i), calculé avec 300 * 300 cases

2.2.2 Modélisation physique

L'éolienne GEV HP est une éolienne qui fonctionne *à vitesse variable*, c'est à dire que la vitesse de rotation des pâles Ω peut être adaptée à la vitesse du vent pour maximiser le transfert d'énergie mécanique. Pour cela le système de commande maintient le rapport $\lambda = \Omega.R/v$ constant à une valeur prédéterminée par des simulations et des essais. Le physicien allemand Albert Betz a montré en 1920 [3] que la puissance maximale extractible par une éolienne est de la forme

$$P = \alpha.v^3 \tag{3}$$

où α est un réel dépendant de paramètres physiques et de la géométrie de l'éolienne. La partie 2.5 revient plus en détails sur l'interprétation physique de ce coefficient α.

Il faut préciser que la relation (3) n'est valable qu'en mode de fonctionnement à vitesse variable, lorsque le rapport λ peut être maintenu constant. Les différents modes de fonctionnements sont illustrés sur la figure 7.

On distingue 4 modes principaux de fonctionnements :

1. *production à l'arrêt* : lorsque le vent est trop faible ($v < v_D$) la commande de l'éolienne ne connecte pas la chaîne de conversion d'énergie au réseau. Pour ne pas

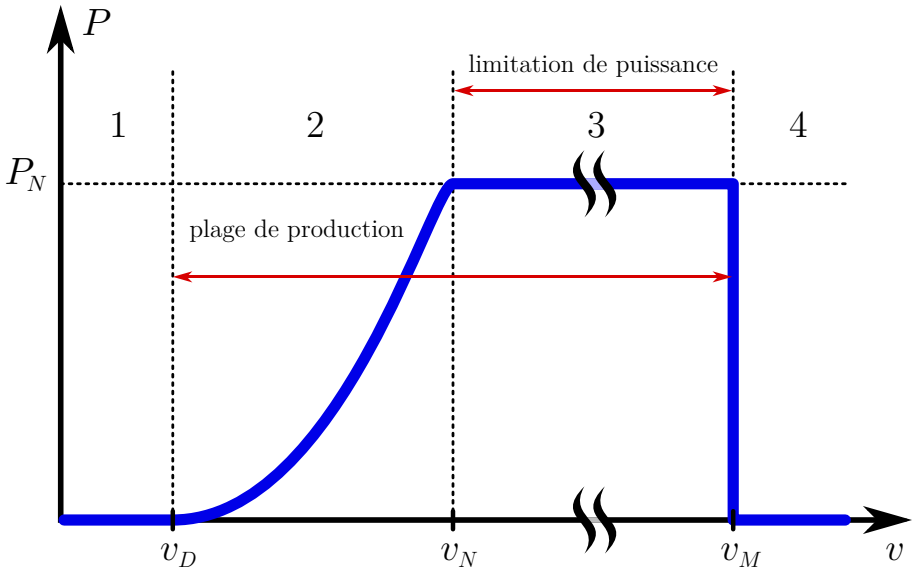

FIGURE 7 – Courbe de puissance d'une éolienne

la déconnecter/reconnecter trop souvent à cause des bourrasques, le vent est moyenné et la commande se fait avec une certaine hystérésis. On a alors $P \approx 0$ W.

2. *production à vitesse variable* : lorsque le vent est modéré ($v_D < v < v_N$) l'éolienne cherche à capturer un maximum de puissance en ajustant sa vitesse de rotation
3. *production écrêtée* : lorsque le vent est fort ($v_N < v < v_M$), la commande va jouer sur l'angle de calage des pâles pour écrêter la puissance mécanique collectée à un niveau acceptable pour la génératrice électrique. On a alors $P \approx P_{nominale} = 1$ MW
4. *éolienne en arrêt de sécurité* : lorsque le vent est trop fort ($v > v_M$) l'éolienne se met en drapeau pour ne pas atteindre une survitesse préjudiciable à la machine.

On remarque que le mode 4 n'a pas été atteint pendant la période de mesure.

Par observation du nuage de point, une méthode empirique simple pour déterminer le mode de fonctionnement est de discriminer le niveau de puissance produite. J'ai choisi, pour avoir une marge de sûreté, de fixer les frontières respectivement à 0,1 % et 98 % de la puissance nominale donnée par le fabricant. L'influence du choix de $Pmin$ et $Pmax$ est étudié à la partie 2.4.1.

Ce faisant, la modélisation du mode 2 se fera avec les couples de données qui appartiennent au domaine

$$\mathcal{D} = \{(P_i,v_i) \in \mathbb{R}^2 \text{ tels que } P_i \in [1\,\text{kW}, 980\,\text{kW}]\} \tag{4}$$

Le domaine $\mathcal{D}$ de fonctionnement à vitesse variable correspond à la tranche horizontale bleue de la figure 8 alors que les domaines où l'éolienne est à l'arrêt ou en production régulée ont été marqués en rouge.

Remarque Cette définition du domaine $\mathcal{D}$ basée sur la puissance (variable expliquée) est une définition *a posteriori*. Elle n'est utilisable qu'une fois la mesure de puissance effectuée. Elle ne rend pas compte du fonctionnement sous-jacent du système de contrôle-commande qui utilise différentes mesures de variables explicatives, en particulier la composante basse

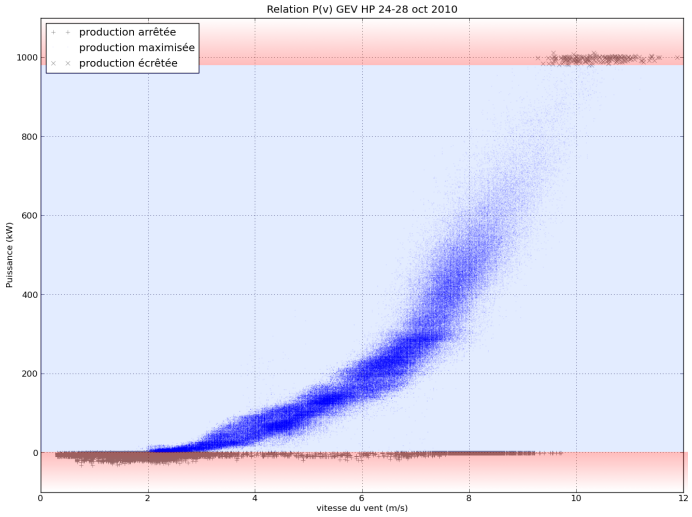

FIGURE 8 – Discrimination des modes de fonctionnements

fréquence de la vitesse du vent. Cependant, ce choix a été fait par raison de simplicité car je ne dispose pas d'une connaissance assez précise de cette loi de commande.

2.3 Régression linéaire, 1er modèle

Pour modéliser le fonctionnement à vitesse variable à l'aide d'une régression linéaire on peut supposer que le lien entre les variables P et v est

$$P = P_0 + \alpha.v^3 + \varepsilon \tag{5}$$

où ε est une variable aléatoire qui représente l'écart au modèle. Ce modèle s'inspire bien de la relation physique (3). De plus, il est *linéaire en les paramètres*, c'est-à-dire un modèle de régression linéaire.

Les paramètres à estimer sont α et P_0. Le terme affine P_0, absent dans (3), est ajouté pour prendre en compte un "offset" sur la puissance produite et l'étude de significativité montrera si sa présence est utile ou non. On va utiliser la méthode d'estimation des *moindres carrés* (MC).

Le critère quadratique à minimiser s'écrit

$$J(\alpha, P_0) = \sum_{(P_i, v_i)\in\mathcal{D}} (P_i - P_0 - \alpha.v_i^3)^2 \tag{6}$$

On définit alors trois hypothèses sur les données qui permettront d'obtenir un certain nombre de propriétés sur les estimateurs. On les vérifiera au moment où elle seront utiles.

- Hypothèse $\mathcal{H}_1$: la variable explicative prend au moins deux valeurs différentes, c'est à dire qu'il existe (i,j) tels que $v_i \neq v_j$.
- Hypothèse $\mathcal{H}_2$: les erreurs ε_i sont centrées, de même variance σ^2 (homoscédasticité) et non corrélées entres elles.
- Hypothèse $\mathcal{H}_3$: les erreurs sont identiquement distribuées selon une loi normale $\mathcal{N}(0, \sigma^2)$ et sont indépendantes ($\mathcal{H}_3$ contient $\mathcal{H}_2$).

On constate sur la figure 6 que $\mathcal{H}_1$ est bien vérifiée : il y a plusieurs valeurs de vent. Cette hypothèse permet d'assurer *l'unicité* des paramètres qui minimisent $J(\alpha, P_0)$.

L'expression analytique des estimateurs des moindres carrés est

$$\hat{\alpha} = \frac{\sum(v_i^3 - \bar{v^3})P_i}{\sum(v_i^3 - \bar{v^3})^2} \tag{7a}$$

$$\hat{P}_0 = \bar{P} - \hat{\alpha}.\bar{v^3} \tag{7b}$$

où $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ est la moyenne empirique d'une série d'observations.

Régression avec R La régression a été effectuée avec la fonction `lm` disponible dans l'environnement de calcul statistique `R` [26]. J'ai repris ici un extrait du diagnostic de régression fourni par la commande `summary`

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.015e+01  1.482e-01  -68.52   <2e-16 ***
I(speed^3)   9.097e-01  4.786e-04 1900.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49 on 315466 degrees of freedom
Multiple R-squared: 0.9197,     Adjusted R-squared: 0.9197
```

La régression donne donc les estimations suivantes des paramètres
- Terme linéaire $\hat{\alpha} = 0{,}9097$
- Terme affine $\hat{P}_0 = -10{,}15\,\text{kW}$
- Écart-type des résidus $\hat{\sigma} = 49\,\text{kW}$.

Par ailleurs, on constate que le test de significativité affirme sans surprise que v^3 est bien une variable explicative de P. De plus P_0, le terme affine, est aussi significatif mais faible par rapport à la puissance nominale.

Qualité de la régression Cette régression permet d'obtenir une valeur ajustée $\hat{P}(v) = \hat{P}_0 + \hat{\alpha}.v^3$ qui est représentée par la ligne rouge sur la figure 9. On peut interpréter le terme P_0 comme contenant la puissance dissipée statiquement par l'éolienne pour son bon fonctionnement. Le constructeur Vergnet déclare une consommation statique plus faible : $P_0 \approx -5\,\text{kW}$.

Sous l'hypothèse $\mathcal{H}_2$ d'homoscédasticité, on dispose d'une estimation de l'écart-type des estimateurs. On constate que l'incertitude relative est de l'ordre de 1 % pour P_0, et encore plus faible pour α. Cette faible incertitude est due au grand nombre de points utilisés dans la régression.

Par ailleurs, on souhaite construire un intervalle de confiance sur la prédiction d'une $(N+1)^{\text{ième}}$ valeur de puissance obtenue par

$$\hat{P}_{N+1} = \hat{P}_0 + \hat{\alpha}.v_{N+1}^3 \tag{8}$$

L'incertitude sur cette prévision a deux origines : *i*) l'erreur d'estimation des paramètres et *ii*) la présence d'un terme d'erreur ε_{N+1}. À cause du grand nombre de points utilisés dans la régression, c'est ce dernier terme qui est prépondérant.

Un tel intervalle d'incertitude s'obtient facilement dans `R` à l'aide de la commande `predict`. Son calcul est basé sur la normalité des résidus (hypothèse $\mathcal{H}_3$), hypothèse plus

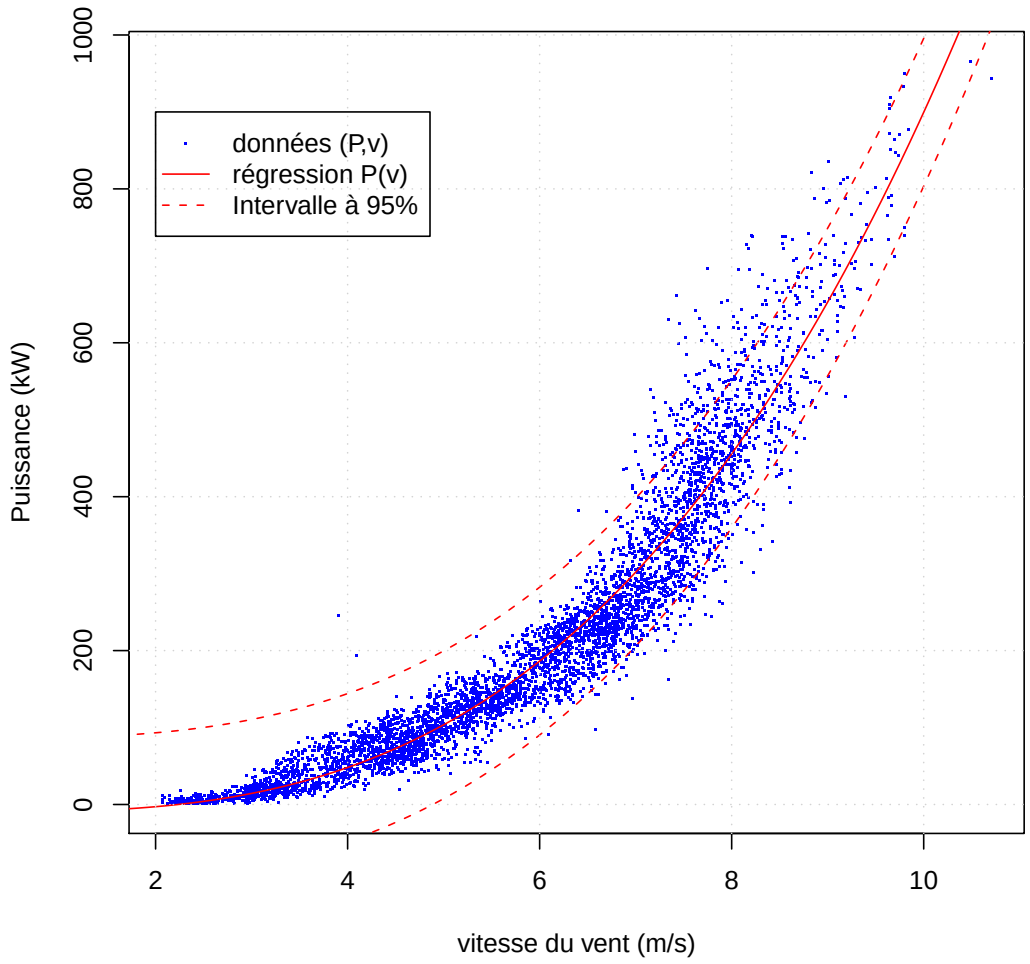

FIGURE 9 – Représentation du modèle (5)

forte que la simple homoscédasticité ($\mathcal{H}_2$). On obtient pour P_{N+1} un intervalle de confiance exact $\mathcal{I}_\alpha$ de niveau $(1-\alpha)$ qui vaut

$$\mathcal{I}_\alpha = \left[\hat{P}_{N+1} \pm t_{N-2}(1-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{N}+\frac{(v^3_{N+1}-\bar{v^3})^2}{\sum(v^3_i-\bar{v^3})^2}}\right] \tag{9}$$

où $t_{N-2}(1-\alpha/2)$ est le quantile de niveau $1-\alpha/2$ d'une loi de Student $\mathcal{T}_{N-2}$ qu'on obtient dans `R` avec la commande `qt(p, df)`. Cet intervalle a été tracé sur les figures 9 et 10.

Vu que N est très grand, sous réserve qu'on ne s'éloigne pas du barycentre du nuage de point, l'IC est approximativement un *ruban autour de la droite de régression* :

$$\mathcal{I}_\alpha = \left[\hat{P}_{N+1} \pm t_{N-2}(1-\alpha/2)\hat{\sigma}\right] \tag{10}$$

De plus, N très grand implique aussi que la loi de Student s'assimile à une normale, et par exemple, pour un IC à 95 % on retrouve le "classique" $t_{N-2}(1-\alpha/2) \approx 1.96$. Donc l'intervalle à 95 % est approximativement $\pm 1{,}96\hat{\sigma} \approx 96$ kW. C'est sur la figure 10, tracée dans l'espace (P, v^3) des variables de la régression linéaire que l'IC apparaît bien comme un ruban d'environ 200 kW de largeur.

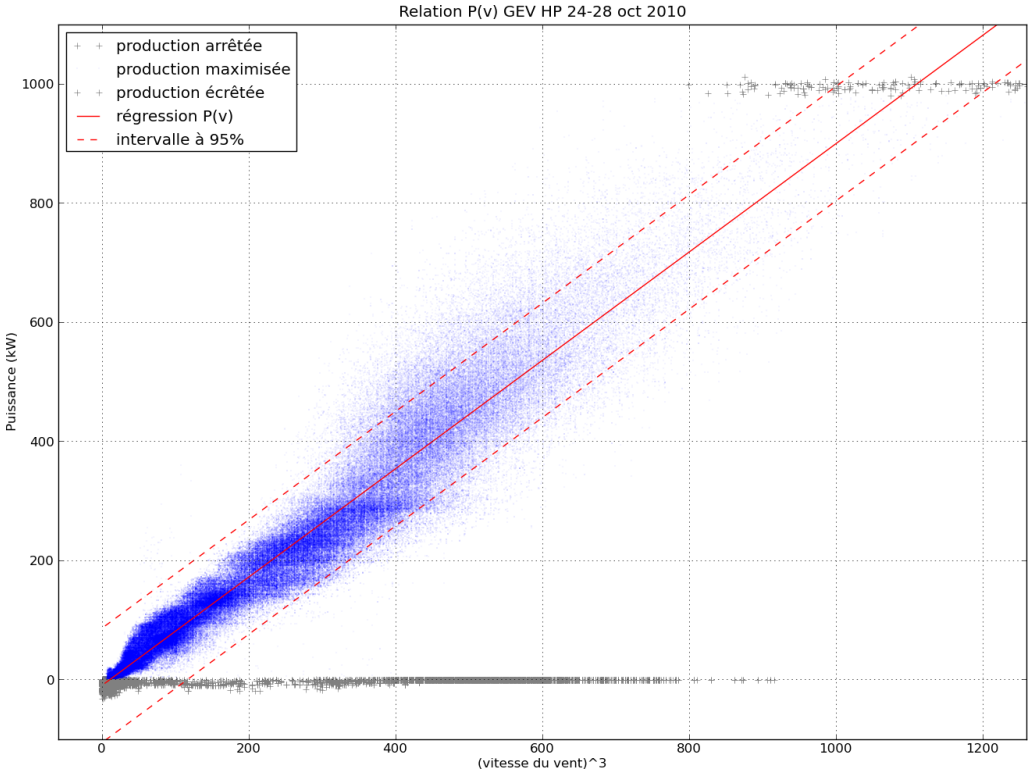

FIGURE 10 – Représentation du modèle (5) dans l'espace des variables de la régression linéaire

On constate alors que ce ruban entoure très mal les données. En effet, le nuage de points présente une *structure en cône*, c'est à dire que la variance des ε_i augmente manifestement avec v_i^3. On a donc une erreur hétéroscédastique conditionnellement à la variable explicative v^3. L'intervalle de confiance donné par (9) étant basé sur des hypothèses non vérifiées, on comprend qu'il ne soit pas satisfaisant.

L'hétéroscédasticité des erreurs constatée ici graphiquement peut aussi être testée numériquement par un test statistique. Cela sera fait après pour le modèle (11) lorsque l'analyse graphique ne permettra pas de trancher de façon aussi claire.

En conclusion, le modèle de régression (5), bien que s'inspirant directement de l'équation physique (3) n'est pas adapté aux données.

2.4 Régression linéaire, un meilleur modèle

On propose alors de se baser sur la même équation (3), mais en se plaçant cette fois dans l'espace des variables $(P^{1/3}, v)$, c'est à dire "l'espace du vent". Le modèle s'écrit

$$P^{1/3} = \beta_0 + \beta_1.v + \varepsilon \tag{11}$$

où ε est à nouveau une variable aléatoire centrée, de variance σ^2 mais dont l'interprétation physique change par rapport à (5) car elle est *homogène à une vitesse*[12] et non plus à une puissance.

12. à la constante β_1 près

Justification du nouveau modèle Le choix de placer le terme d'erreur sur le vent est conforté par une analyse physique du système. En effet la puissance électrique $P(t)$ est mesurée précisément alors que $v(t)$ est mesurée par un anémomètre de précision moindre. De plus, il y a probablement des turbulences dues à la machine qui perturbent la mesure (l'anémomètre se situe en aval des pâles). Enfin, le vent est mesuré en un point alors qu'il n'est pas forcément uniforme sur l'ensemble de la surface balayée par les pâles.

Régression avec R Cette nouvelle régression a été effectuée avec la fonction `lm` et voici un extrait du diagnostic de régression fourni par la commande `summary` :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4176233  0.0027215  -153.5   <2e-16 ***
speed        1.0120306  0.0004466  2266.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4279 on 315466 degrees of freedom
Multiple R-squared: 0.9421,     Adjusted R-squared: 0.9421
```

La régression donne donc les estimations suivantes des paramètres
- Terme linéaire $\hat{\beta}_1 = 1{,}012$
- Terme affine $\hat{\beta}_0 = -0{,}418$
- Écart-type des résidus $\hat{\sigma} = 0{,}428$.

Et les deux paramètres de la régression sont bien significatifs. Par ailleurs, le coefficient R^2 est meilleur que précédemment, mais comme les modèles (5) et (11) ne portent pas sur les mêmes variables on ne peut pas les comparer.

Il est physiquement intéressant de traduire le terme affine β_0 comme un décalage de la mesure de la vitesse v. Un tel "offset" vaudrait $v_0 = -\beta_0/\beta_1$. On obtient ainsi $\hat{v}_0 = 0.413\,\text{m/s}$. De façon similaire, on peut rendre l'écart-type homogène à une vitesse : $\hat{\sigma}/\hat{\beta}_1 = 0{,}423\,\text{m/s}$.

Le tracé de la régression est obtenu par l'estimateur

$$\hat{P}_{N+1} = (\hat{\beta}_1 . v_{N+1} + \hat{\beta}_0)^3 \tag{12}$$

que l'on obtient en faisant l'approximation $\hat{P} = (\widehat{P^{1/3}})^3$. C'est le tracé rouge de la figure 11.

Par ailleurs, on peut construire un intervalle de confiance comme dans la partie 2.3, en replaçant les couples (P, v^3) par $(P^{1/3}, v)$. En particulier, l'expression simplifiée (10) de l'intervalle de confiance $\mathcal{I}_\alpha$ de niveau $(1-\alpha)$ devient :

$$\mathcal{I}_\alpha = \left[\widehat{P^{1/3}}_{N+1} \pm t_{N-2}(1-\alpha/2)\hat{\sigma}\right] \tag{13}$$

La largeur de l'IC à 95%, que l'on visualise le mieux sur la figure 12, est par exemple égale à $2 * 1{,}96\ \hat{\sigma} \approx 1{,}68$. Par rapport à la figure 10, on constate que cet IC encadre nettement mieux le nuage de points.

Correction de la prévision La régression (11) permet d'obtenir la prévision de la variable $Y = P_{N+1}^{1/3}$. Pour obtenir la prévision de $P_{N+1} = Y^3$, il n'est pas exact de prendre simplement le cube de celle de Y car $E(Y^3) \neq E(Y)^3$. On va calculer la "correction"

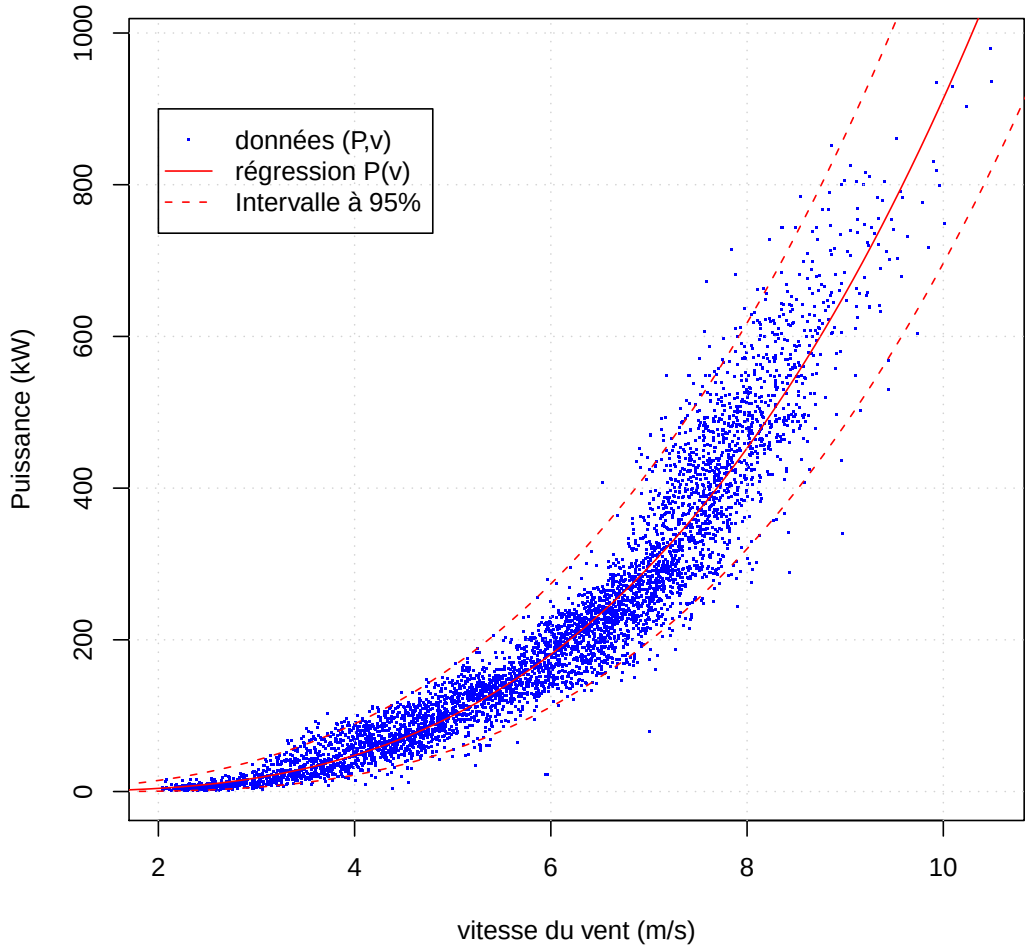

FIGURE 11 – Représentation du modèle (11)

nécessaire en faisant l'hypothèse que les résidus sont gaussiens. Partant de $Y \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu = \widehat{P^{1/3}}_{N+1}$, on obtient :

$$E(P|v) = E(Y^3) = \int (x+\mu)^3 \mathcal{N}_{0,\sigma^2}(x).dx$$

On développe la puissance en somme de quatre termes qui vont donner les différents moments

$$E(Y^3) = \int (x^3 + 3x^2\mu + 3x\mu^2 + \mu^3)\mathcal{N}_{0,\sigma^2}(x).dx$$

Après séparation de l'intégrale en une somme de quatre termes, on constante que le 1$^{\text{er}}$ et le 3$^{\text{ième}}$ sont nuls pour cause d'imparité. Il reste finalement

$$E(P|v) = E(Y^3) = \mu^3 + 3\mu\sigma^2 \tag{14}$$

Cependant, la valeur numérique du terme correctif $3\mu\sigma^2$ est numériquement faible au regard de μ^3 sauf pour de faibles valeurs de μ. Ainsi dans le cas considéré, l'approximation $E(P) \approx \mu^3 = (\widehat{P^{1/3}}_{N+1})^3$ est acceptable.

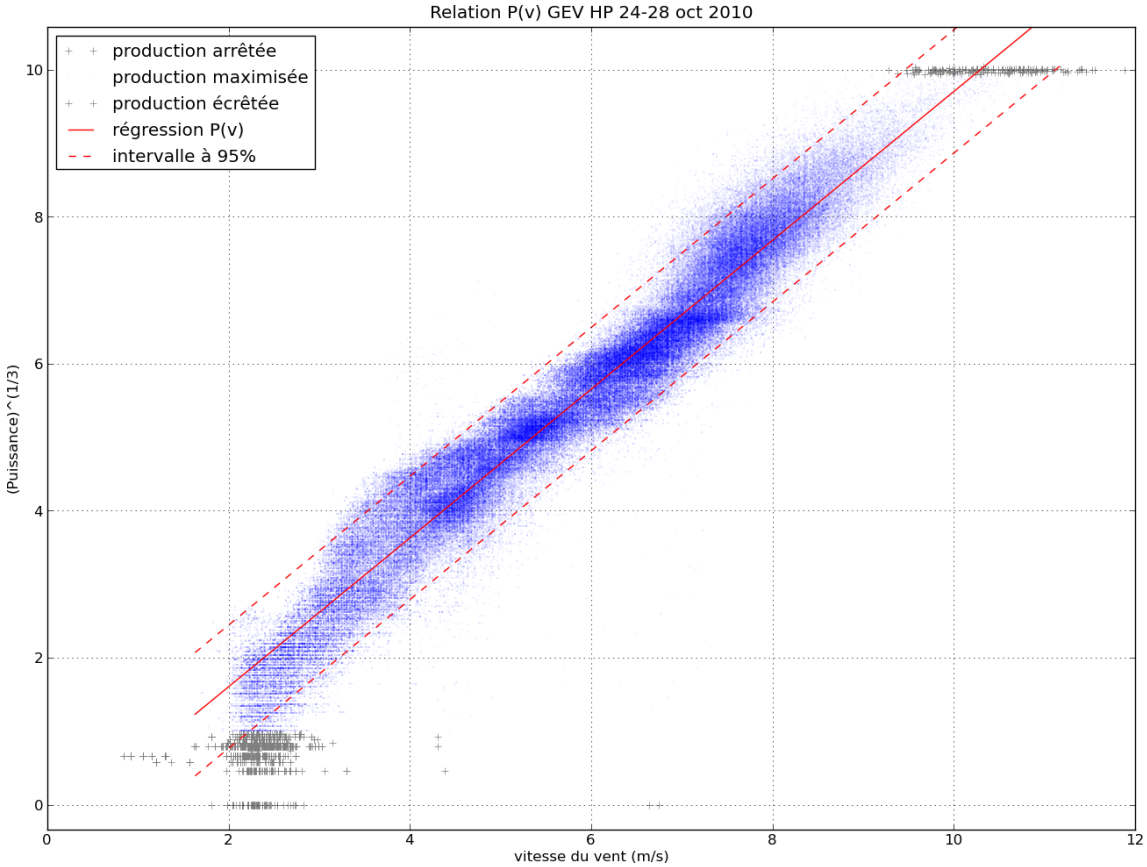

FIGURE 12 – Représentation du modèle (11) dans l'espace des variables de la régression linéaire

Analyse des résidus Pour vérifier la validité de l'IC donné par (13) il faut vérifier l'hypothèse $\mathcal{H}_3$ de normalité des erreurs, ce qui implique en premier lieu homoscédasticité $\mathcal{H}_2$.

Pour commencer avec une analyse graphique, on a représenté sur la figure 13 les résidus du modèle (11) de deux manières différentes : dans le domaine temporel et en fonction de la puissance mesurée. On a également ajouté les lignes $\pm 1{,}96\,\hat{\sigma}$ pour référence.

On constate donc que les résidus sont situés de façon *assez homogène* dans le ruban $\pm 1{,}96\,\hat{\sigma}$ et il n'y a pas de tendance simple donnant une augmentation/diminution de la variance avec P. Cependant on ne peut pas dire que les résidus sont indépendants de P. Par exemple, les résidus semblent décentrés vers $P^{1/3} \approx 6{,}5$ (c'est à dire $P \approx 300\,\text{kW}$).

L'analyse peut aussi être menée numériquement avec un test statistique d'hétéroscédasticité, tel que le test de White ou ou celui de Breusch–Pagan [4][13]. Similaires dans leur principe, ils se basent sur une régression du carré des résidus $\hat{\varepsilon}_i^2$ par les variables explicatives de la première regression. Dans le cas du modèle (11), il s'agit de régresser ε^2 sur v :

$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 . v_i + \eta_i \tag{15}$$

Le test de White suit le même principe, en y ajoutant des variables explicatives supplémentaires tels que les produits croisés. Une fois le modèle posé, il s'agit de voire si

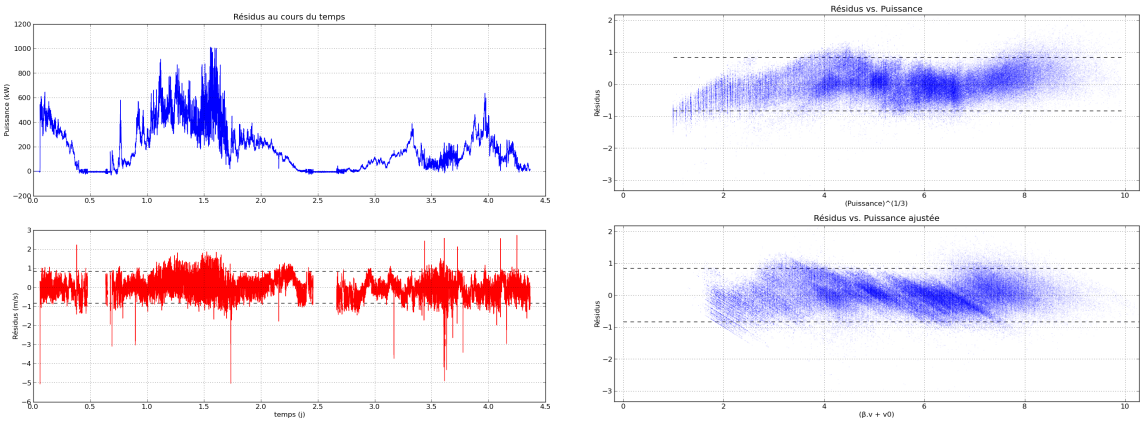

FIGURE 13 – Représentation des résidus, dans le domaine temporel et en fonction de la puissance

l'hypothèse nulle $H_0 : \gamma_1 = 0$ est acceptable. La statistique de test utilisé est $N.R^2$ dont la loi asymptotique est un χ^2 à un degré de liberté [13].

Le test de Breusch–Pagan a été implémenté dans `R` par la fonction `bptest` du paquet `lmtest` [32]. On obtient la statistique de test $BP = 850{,}7$ (version "studentisée") dont la p-value est $< 2{,}2.10^{-16}$ (en dessous de la limite de précision de calcul en virgule flottante). L'homoscédasticité est donc clairement rejetée ! À titre de comparaison, le même test appliqué à la regression précédente (5) génère une statistique encore plus grande : $BP = 46944{,}7$.

Par ailleurs, le calcul de l'autocorrélation estimée des résidus montre que ceux-ci sont corrélés : il n'y a donc pas non plus d'indépendance temporel. Cependant comme il y a eu des soucis au niveau du mécanisme d'acquisition, ce constat n'est pas sereinement interprétable.

Distribution des résidus J'ai également regardé si les résidus estimés sont *distribués normalement*. Cette analyse de la gaussianité peut se faire avec un histogramme, où l'on constate que la distribution empirique des résidus est *très proche d'une loi normale* $\mathcal{N}(0, \sigma^2)$. Cependant, une étude plus attentive montre que la distribution est proche de la normale dans un intervalle $[-3\sigma, +3\sigma]$ (qui contient plus de 99,5 % des données). Par contre en dehors de cet intervalle, il y a beaucoup plus de points extrêmes qu'attendu avec une loi normale. C'est cette présence de valeurs extrêmes qu'illustre la figure 14 où les quantiles empiriques sont tracés en fonction de ceux d'une loi normale $\mathcal{N}(0, 1)$ (fonction `qqnorm` de `R`). On voit que la ligne des quantiles gaussien est bien suivie sur l'intervalle $[-3, +3]$, mais au-delà les résidus extrêmes sont en surnombre.

2.4.1 Robustesse vis à vis du jeu de donnée

On a expliqué dans la partie 2.2.2 que la régression serait faite avec un sous-ensemble du jeu de données $(P_i, v_i) \in \mathcal{D}$, dans l'objectif de ne garder que les points correspondant au fonctionnement *à vitesse variable*. Le domaine $\mathcal{D}$ est défini à l'aide de deux constantes : P_{min} et P_{max}. Leur choix est dicté en partie par des spécifications physique

13. dans le cas général la loi est un $\chi^2(p-1)$ où p est le nombre de paramètres de la régression

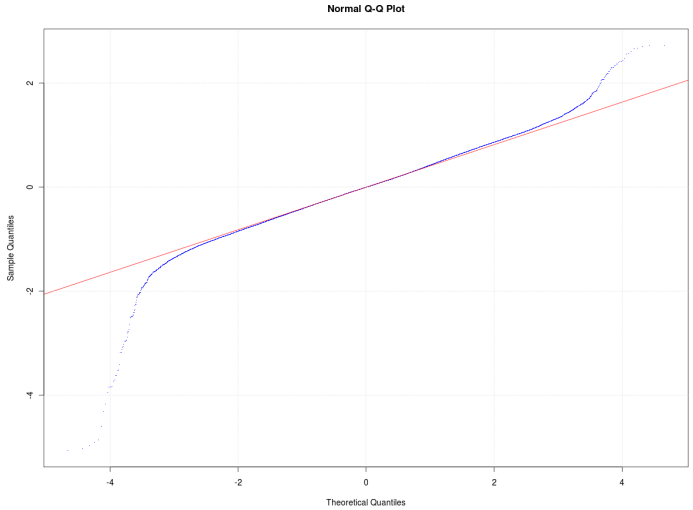

FIGURE 14 – Représentation de la distribution des résidus avec `qqnorm`

($P_{nominale} = 1$ MW) et en partie par des constats empiriques. Dans l'étude qui précède les constantes étaient fixées respectivement à 0,1 % et 98 % de la puissance nominale. On observe maintenant l'impact de ces choix sur l'estimation des coefficients β_0 et β_1 de la régression (11). Les estimées $\hat{\beta}_0$ et $\hat{\beta}_1$ sont tracées sur les figures 16 et 15 en fonction de P_{min} et P_{max} que l'on a fait varier séparément. Pour visualiser le comportement général de la regression, on a aussi tracé les valeurs de $\hat{\sigma}$ et de R^2.

Sensibilité à P_{max} Sur la figure 15, on constate que $\hat{\beta}_0$ et $\hat{\beta}_1$ varient peu. Leur variation se fait *en sens opposés*, ce qui est attendu [14]. Les variations sont *faibles* car le nombre de points en fonctionnement écrêté est faible (cf figure 6) sur la série de données étudiées.

On voit un effet que $\hat{\beta}_1$ (qui est la pente du modèle) baisse progressivement pour $P_{max} > 980$ kW. Cette baisse s'explique par l'incorporation progressive des points correspondants au fonctionnement écrêté, où P est plus faible qu'attendu par le modèle en v^3. On en conclut donc que le choix de P_{max} à 98 % de la puissance nominale est raisonnable puisqu'il élimine juste les points "écrêtés".

Sensibilité à P_{min} Sur la figure 16, on constate que $\hat{\beta}_0$ et $\hat{\beta}_1$ varient beaucoup plus que lorsque P_{max} varie. Cela est s'explique par le beaucoup plus grand nombre de données en jeu.

Le grand saut accompagné d'une chute du R^2 pour $P_{min} < -0{,}25$ kW s'explique par l'incorporation brutale des points où l'éolienne est transitoirement arrêtée alors que le vent est non nul. Comme ces points ont tous une puissance négative on en conclut que P_{min} aurait pu être choisi un peu plus faible, par exemple à 0 %.

Le fait que $\hat{\beta}_0$ et $\hat{\beta}_1$ varient continûment pour $P_{min} > 0$ kW montre que les points où la puissance est faible sont en limite de validité du modèle.

14. dans le cas d'une régression linéaire simple avec où le régresseur est décentré [5, Prop. 1.2]

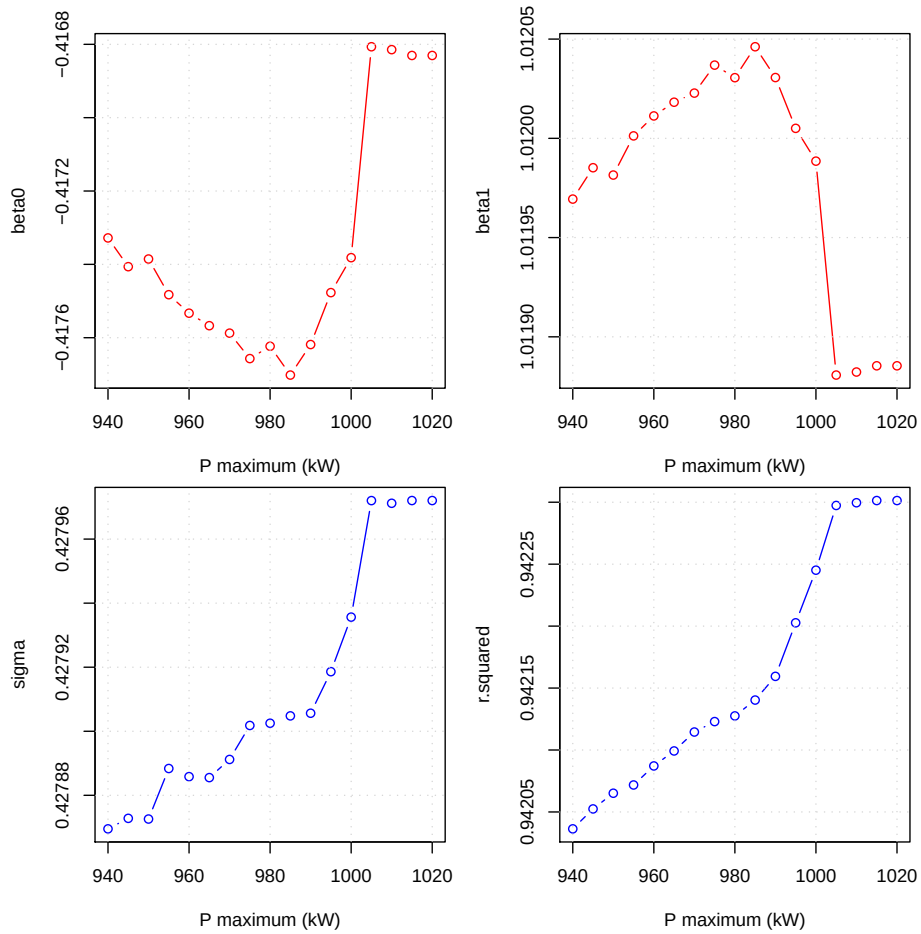

FIGURE 15 – Effet du choix de P_{max}

2.5 Interprétation physique de la régression

La puissance mécanique que peut récupérer une éolienne est classiquement liée à la puissance du vent par une relation issue de la mécanique des fluides [27][3]

$$P_w = \frac{1}{2}\rho A_r c_p(\lambda, \theta) v^3 \tag{16}$$

qui fait intervenir les grandeurs suivantes :

- la vitesse du vent v, supposée uniforme
- la masse volumique de l'air $\rho \approx 1{,}2\,\mathrm{kg/m^3}$
- l'aire du disque balayée par les pâles de la machine $A_r = \pi R^2 \approx 3019\,\mathrm{m^2}$
- le coefficient de puissance $c_p(\lambda, \theta)$ qui caractérise le rendement de la collecte d'énergie cinétique. (où $\lambda = \Omega.R/v$ est la vitesse spécifique et θ l'angle de calage des pâles)

La régression linéaire effectuée au 2.3 permet donc d'identifier le coefficient de puissance. Il faut cependant garder à l'esprit que l'on ne peut obtenir qu'une approximation, probablement sous-évaluée, du c_p car la puissance électrique est égale à la puissance mécanique *minorée des pertes* de conversion électromécanique par exemple. Pour l'identification on utilise les modèles (5) et (11) où l'on ne conserve que le terme variant avec v.

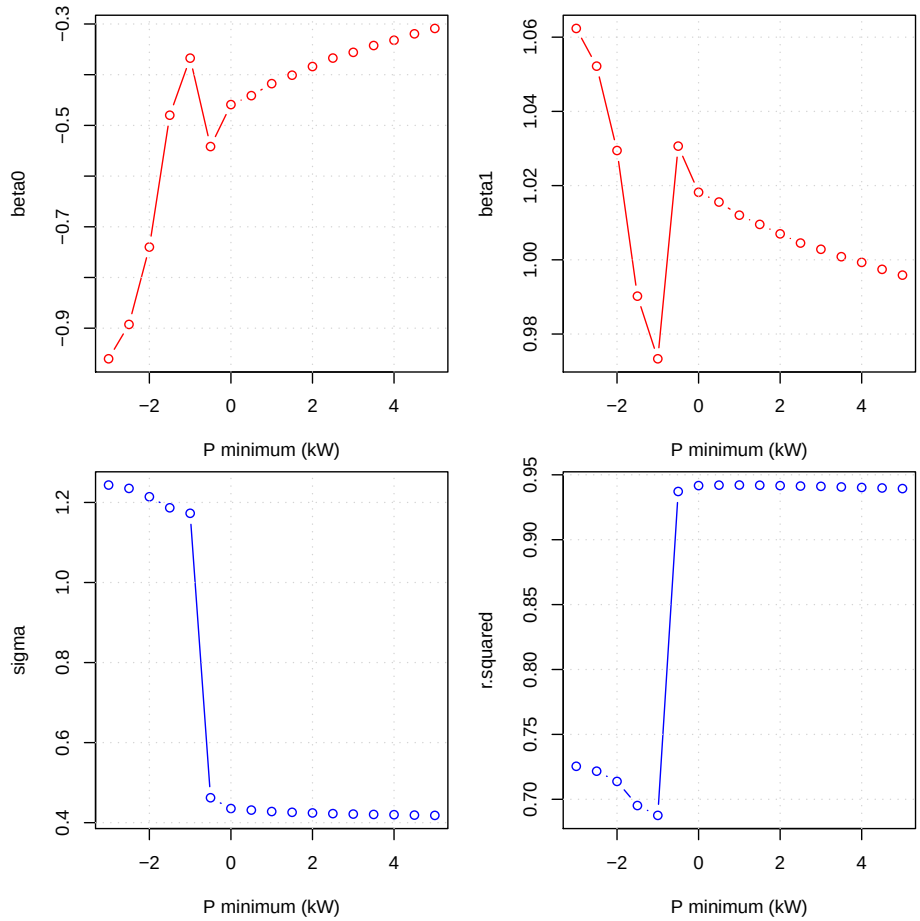

FIGURE 16 – Effet du choix de P_{min}

$$c_p \approx \frac{\alpha}{\frac{1}{2}\rho A_r} \approx \frac{\beta^3}{\frac{1}{2}\rho A_r} \tag{17}$$

Sachant qu'on avait (en unités SI) $\alpha \approx 1000\text{Wm}^{-3}\text{s}^3$, on obtient donc $c_p \approx 0{,}5$ ce qui est conforme à ce qu'on peut attendre d'une éolienne bipale dans sa zone de fonctionnement optimal [27]. On rappelle aussi que Albert Betz a établi en 1920 [3] que la limite théorique, obtenue avec certaines hypothèses idéalisatrices sur l'écoulement de fluide, est de $16/27 \approx 60\%$. Les éoliennes actuelles atteignent 70 à 85 % de cette limite [27] et c'est bien ce qu'on observe avec la GEV HP de Vergnet.

2.6 Vers un modèle plus compliqué ?

La "structure en S" des résidus que l'on observe sur la figure 13 montre que l'explication de P par la variable v^3 est insuffisante. Cette structure a été confirmée par des régressions faite en *moyennant temporellement* les données par paquets quelques milliers de points (quelques dizaines de minutes). On constate que ce moyennage – qui devrait resserrer les résidus autour de zéro si ils étaient centrés et décorrélés – ne fait que renforcer la perception de cette "structure en S".

Pour rendre les résidus plus indépendants de la variable v, on pourrait ajouter au modèle $P(v)$ des termes explicatifs additionnels (v, v^2, ...) mais le choix de ce modèle amélioré serait assez arbitraire et spécifique au type de machine étudié contrairement au v^3 directement inspiré par la physique. On reste donc sur une regression simple, tout en gardant à l'esprit ses limites.

Modélisation temporelle On suppose fortement que les résidus du modèle entre deux instants consécutifs sont *corrélés*. Cependant les premières séries de donnée reçues sont victimes de défauts rédhibitoires dus au mécanisme d'acquisition (cf partie 2.1.1). C'est donc faute de mieux que nous nous sommes tenu à une modélisation *statique*. Nous avons œuvré pour obtenir des nouvelles séries et nous espérons qu'elles permettront de passer à une modélisation temporelle.

3 Données de puissance avec prévision

Cette partie s'appuie sur les archives de production des fermes Aérowatt en outre-mer. On dispose des historiques de production (puissance électrique produite) qui sont accompagnés de *prévision* de cette puissance. La prévision est fournie par la société Metnext, présentée dans la partie 1.3.1. Par rapport à celles la partie 2, ces données sont disponibles sur de bien plus grandes périodes – plusieurs mois – ce qui implique une *meilleure signification climatique*. Par contre la fréquence d'échantillonage est plus faible : une donnée par heure. L'objectif est de *quantifier l'incertitude* qui entoure la prévision.

3.1 Présentation des données

Aérowatt nous a fourni des données de plusieurs types, issues de plusieurs fermes. L'étude présentée ici s'appuie sur les archives des couples (P_i, Q_i) où P_i désigne la production à l'heure i, alors que Q_i désigne la prévision correspondante. Cette dernière est générée quotidiennement, la veille pour le lendemain [15] par le logiciel Metnext. Cette prévision sert actuellement à EDF SEI à optimiser sa planification de production pour la Guadeloupe.

Nous avons utilisées les données de deux fermes :

- *Fonds Caraïbes* (FC) en Guadeloupe : 20 machines 220 kW pour une puissance nominale de 4400 kW
- *Grand Maison* (GM) en Guadeloupe : 5 machines 275 kW pour une puissance nominale de 1375 kW

Les données d'une troisième ferme, La Perrière à La Réunion n'ont pas été exploitées car la puissance produite y était volontairement limitée durant de nombreux mois. En effet, le poste d'interconnexion électrique était un temps sous-dimensionné. En conséquence, la distribution statistique de la puissance produite est très différente de celle d'une ferme habituelle.

Un extrait de deux semaines des données de Grand Maison est proposé en figure 17. Pour rendre le graphique plus lisible, la mesure de puissance a été filtrée par un passe-bas (courbe bleue). La mesure non filtrée est en bleu clair en arrière plan. La prévision est en rouge. On constate bien une ressemblance entre prévisions et mesures et on souhaite quantifier ce degré de ressemblance par un intervalle de confiance.

15. la valeur précise du pas de prédiction n'apparaît malheureusement pas dans les fichiers

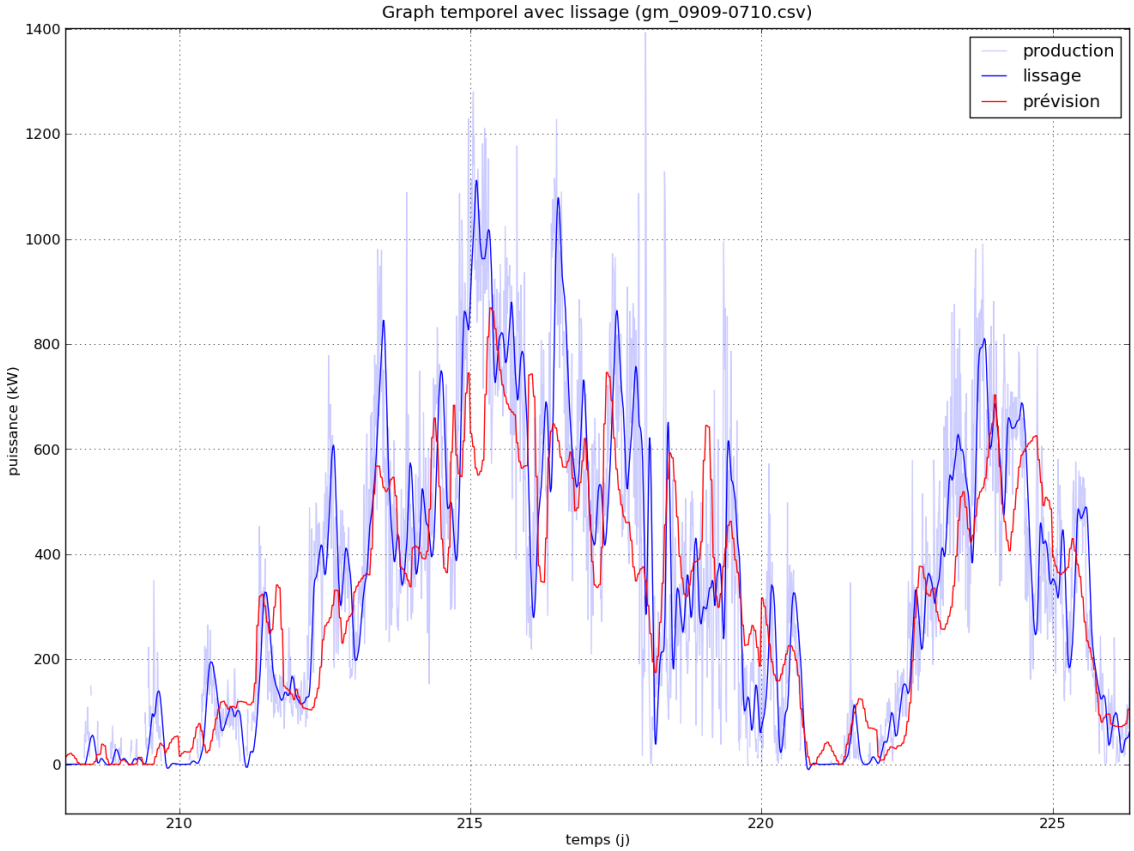

FIGURE 17 – Tracé temporel des puissances mesurée et prévue sur 2 semaines

3.1.1 Agrégation des données

Le fichier utilisé contient en réalité la production moyenne par pas de 10 minutes, alors que la prévision est établie par pas d'une heure (chaque prévision apparaît donc répétées 6 fois de suite). Pour comparer statistiquement production et prévision, ceci n'est pas satisfaisant car les deux séries ne capturent pas les mêmes dynamiques temporelles. J'ai donc reconstitué une moyenne de production sur une heure en utilisant les 6 tranches de 10 minutes appartenant à une même heure. La formulation mathématique correspondante est

$$P_i = \frac{1}{T}\sum_{n=1}^{T} \tilde{P}_{(i-1)T+n} \quad \text{avec } T = 6 \tag{18}$$

où $(\tilde{P}_n)_{n\geqslant 1}$ désigne la série initiale des puissances en moyenne 10 minutes alors que $(P_i)_{i\geqslant 1}$ est une série agrégée des moyennes sur 1 heure.

3.1.2 Nettoyage des données & Points manquants

Les données issues des fermes Aérowatt ne sont pas complètes car la communication entre les machines et le centre de supervision est parfois rompue. Les mesures manquantes se doivent d'être traitées proprement en temps que `NAs` [16]. Malheureusement, dans cer-

16. code pour une donnée non disponible (Not Available)

taines tableaux d'archives, le code "hors-ligne" est utilisé ; d'autres fois les cellules sont laissées vides. C'est pourquoi de nombreuses heures ont été dépensées dans l'écriture de scripts de retraitement pour générer des fichiers formatés de façon plus uniforme qui soit facilement lisibles par `R` ou `Python`.

Plus dangereux, il est apparu quelques problèmes sur le contenu des données :

- *Décalage temporel :* dans une première version des fichiers de données, les puissances et les prévisions semblaient décalées (décalage détecté par le tracé de l'intercorrélation). Il s'est avéré que les données avaient été indexées dans des fuseaux horaires différents ! Ce problème a été résolu dans la version suivante des fichiers.
- *Données aberrantes :* certaines mesures de puissance sont apparues aberrantes, car atteignant des valeurs trop grandes et présentant un aspect trop lisse. Un script *ad hoc* a été écrit pour remplacer ces points par des `NAs`. La détection est basée essentiellement sur le calcul de la différence d'ordre 2 ($\propto$ dérivée seconde). Si la dérivée est trop faible, le point est considéré comme aberrant. Le réglages des seuils a été fait empiriquement avec validation manuelle (méthode *trial & error*).

Ce travail de préanalyse, à défaut d'être très intellectuellement gratifiant, est un préalable absolument nécessaire pour pouvoir démarrer une étude statistique qui n'aboutisse pas sur des conclusions biaisées.

Proportion de points manquants Au final, la quantité de données non disponibles, après le moyennage décrit au paragraphe 3.1.1 est de 24 % pour GM (15 % avant agrégation), et de 50% pour FC (29 % avant agrégation). L'augmentation de la proportion de `NAs` après l'agrégation est liée au caractère "contaminant" des `NAs`. En effet, une moyenne est considérée `NA` dès qu'un des termes de la somme (18) est `NA`.

3.1.3 Normalisation

Pour permettre une meilleure comparaison des résultats entre les fermes, la puissance produite a été normalisée grâce à une transformation de normalisation :

$$P_i \leftarrow P_i / P_{nom} \tag{19a}$$

$$Q_i \leftarrow Q_i / P_{nom} \tag{19b}$$

En conséquence, les données se situent *a priori* dans l'intervalle $[0, 1]$. La valeur de P_{nom} est de 4400 kW pour FC et de 1375 kW pour GM. Cette normalisation facilite aussi la transformation introduite ci-après (partie 3.3).

Après normalisation, voici quelques statistiques sur les données FC.

	Min.	1$^{\text{er}}$ Q.	Méd.	Moy.	3$^{\text{ème}}$ Q.	Max.
Production normalisée	0,000	0,075	0,158	0,218	0,314	0,917
Prévision normalisée	0,000	0,084	0,172	0,205	0,302	0,657

On constate que comme avec les données de la partie 2, la puissance produite par une ferme éolienne est fortement *dissymétrique* et *décentrée* vers la borne gauche de l'intervalle nominal $[0, 1]$.

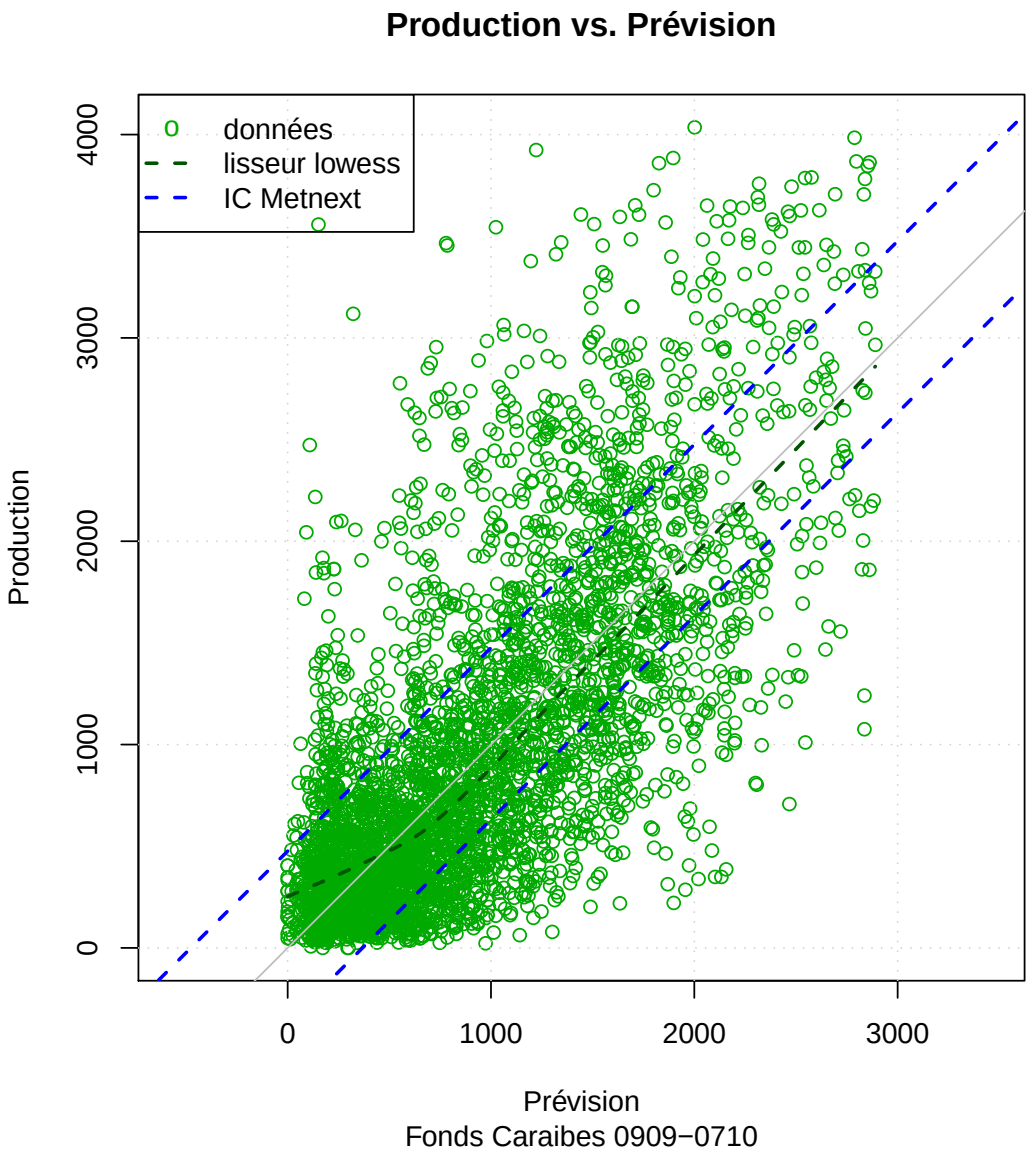

FIGURE 18 – Nuage de point $P_i \sim Q_i$ accompagné de l'IC Metnext

3.2 Un premier tracé

On a tracé en figure 18 le nuage de point $P_i \sim Q_i$ pour la ferme de Fonds Caraïbes. Si la prévision était parfaite, les points seraient alignés sur la droite identité (ligne grise) et ce n'est bien sûr pas le cas. On constate que les écarts peuvent être assez important et c'est ces écarts que nous voulons quantifier.

Hétéroscédasticité Les prévisions Metnext fournies chaque jour à EDF pour la gestion du réseau sont accompagnées depuis le 20 mars 2009 d'un intervalle de confiance. Nous avons constaté qu'il est de la forme

- Fonds Caraïbes : $[Q_i - 99\,\text{kW}, Q_i + 112\,\text{kW}] \cap [0, P_{nom}]$
- Grand Maison : $[Q_i - 368\,\text{kW}, Q_i + 476\,\text{kW}] \cap [0, P_{nom}]$

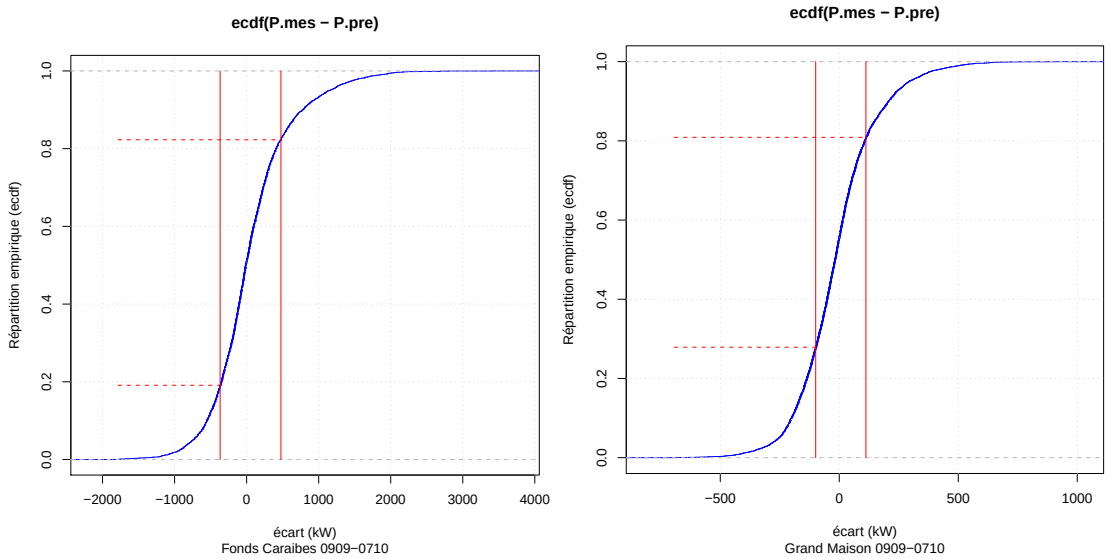

FIGURE 19 – Répartition empirique de l'écart à la prévision $P_i - Q_i$

Ces intervalles forment ce que l'on peut appeler un *ruban d'incertitude*. Il est tracé en bleu sur la figure 18. Sa largeur est de l'ordre de 10 % de la puissance nominale et on constate qu'il est légèrement *dissymétrique*. Nous n'avons malheureusement que peu de détails sur la construction ou sur les propriétés de cet IC. Nous avons donc procédé à une étude rapide de son niveau empirique [17] grâce au calcul de la fonction de répartition empirique (fonction `ecdf` de `R`) des écarts $e_i = P_i - Q_i$. Cette répartition est tracée pour chaque ferme sur la figure 19. Le niveau de l'intervalle pour FC et GM est respectivement évaluée à 63 % et 53 %. Ces niveaux empiriques sont à comparer avec le *niveau nominal* visé par Metnext qui est de 50 %.

La construction d'un intervalle en ruban suppose que l'écart e_i est de variance conditionnelle constante : hypothèse *d'homoscédasticité*. Or on peut imaginer deux causes raisonnables d'hétéroscédasticité conditionnelle :

1. la variance peut dépendre de l'horizon de prédiction et
2. la variance peut dépendre du niveau de puissance prévue.

L'effet de l'horizon de prédiction sur la variance ne nous a *pas paru significatif*. Dans une étude indépendante [19] l'horizon ne semble pas non plus avoir grand effet.

Par contre, l'observation de la figure 18 montre que l'incertitude forme un *cône* qui s'évase lorsque le niveau de puissance prévue Q_i augmente. L'intervalle en ruban n'est donc pas très bien adapté. Il est alors possible de procéder à une modélisation de l'écart e_i dont la variance dépendrait de Q_i. Cependant cela nécessite de choisir la forme de la dépendance de la variance $\sigma^2(Q)$. Nous préférons proposer une méthode qui permet, dans une certaine mesure, de guider ce choix.

17. proportion de données à l'intérieur de l'intervalle

3.3 Transformation des variables

Inspiré par le travail de la partie 2.4 nous proposons de transformer les variables de la régression en utilisant une transformation "gamma [18]" :

$$\Phi_\gamma : x \mapsto x^{1/\gamma} \quad \text{avec } \gamma > 0 \tag{20}$$

Il s'agit d'une transformation non-linéaire instantannée, paramétrée par le réel $\gamma > 0$ qu'il faudra ensuite choisir. On relève que c'est une bijection de $[0, 1]$ sur lui-même, d'où l'intérêt de normaliser les données.

Motivation Dans la partie 2.4 on avait vu que les données $P_i^{1/3}$, qui correspondent au cas $\gamma = 3$, avaient, pour des raisons physiques, une meilleure distribution statistique que les données originales P_i. Or on a montré au 3.1.3 que les données des fermes Aérowatt ont des distributions similaires aux données Vergnet (dissymétriques et décentrées). Il paraît donc raisonnable d'appliquer une transformation similaire. Comme on n'a plus de variable représentant le vent, il n'y a pas de raison de se contraindre à $\gamma = 3$; on peut le prendre positif quelconque. Par contre si l'argumentation physique doit rester pertinente, on devrait trouver une *valeur proche de 3*.

3.3.1 Critère de choix

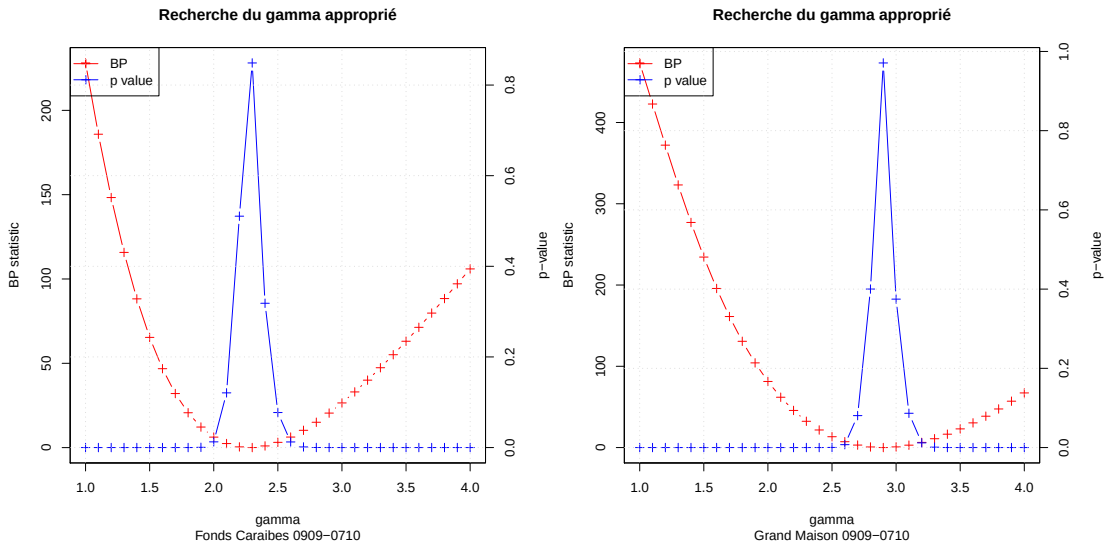

FIGURE 20 – Recherche d'une valeur de γ appropriée

Dans la partie 2.4, l'application de la transformation Φ pour $\gamma = 3$ avait permis d'obtenir une régression où les résidus étaient *plus homoscédastiques* que sans la transformation. Maintenant que nous avons un degré de liberté sur γ, nous allons chercher à obtenir les résidus les *plus homoscédastiques possible*. Il s'agit donc d'un problème d'optimisation dont le critère va se baser sur la statistique du test d'homoscédasticité de Breusch-Pagan [4] déjà utilisé au 2.4.

18. nom lié au *gamma* ou *facteur de contraste* utilisé pour l'enregistrement et la reproduction d'images

3.3.2 Modèle de régression

On utilise un modèle de régression linéaire simple paramétré par γ :

$$P_i^{1/\gamma} = \beta_0 + \beta_1 . Q_i^{1/\gamma} + \varepsilon_i \tag{21}$$

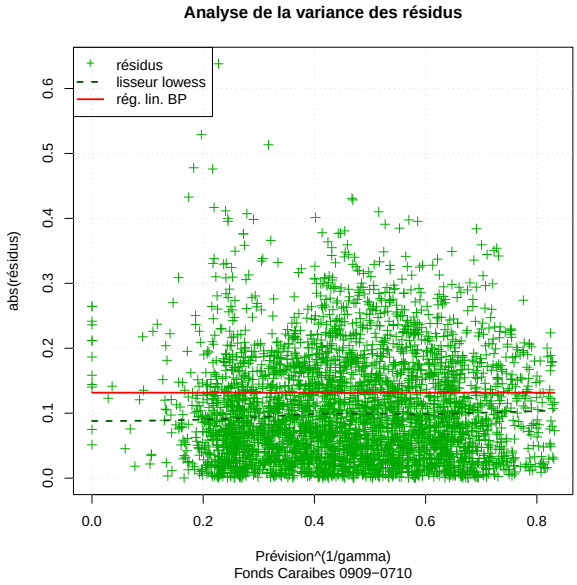

FIGURE 21 – Résidus, en valeur absolue, en fonction de Q_i^{1/γ^*} (transformation optimale)

L'application de la méthode des moindres carrés ordinaires permet d'obtenir $\hat{\beta}_0$ et $\hat{\beta}_1$ et par suite, des résidus estimés $\hat{\varepsilon}_i$. Le test de Breusch–Pagan se base sur le modèle (15) qui régresse les $\hat{\varepsilon}_i^2$ par rapport à la variable explicative $Q_i^{1/\gamma}$. Sous l'hypothèse $\mathcal{H}_0$, la statistique $BP = N.R^2$ a pour loi asymptotique un χ^2 à un degré de liberté. En particulier, elle est de *valeur faible* en espérance, par rapport au cas $\bar{\mathcal{H}}_0$. On va donc se servir de la statistique de test comme critère à minimiser.

$$J(\gamma, \underline{P}, \underline{Q}) = BP\{P_i^{1/\gamma} \sim Q_i^{1/\gamma}\} \tag{22}$$

Le calcul de ce critère se fait donc à l'aide du calcul de deux régressions linéaires. Il est donc relativement rapide à calculer. Vu que les données $\underline{P}$ et $\underline{Q}$ sont fixées, il s'agit d'un problème d'optimisation monovariable, que l'on a résolu numériquement avec la fonction `optimize` qui est dédiée à la recherche monodimensionnelle sur sur un intervalle. Pour illustrer ce travail d'optimisation, on a tracé l'évaluation du critère $J(\gamma)$ pour plusieurs valeurs de γ sur la figure 20 (courbe rouge). Le minimiseur γ^* que l'on trouve est une *fonction des données.*

$$\gamma^*(\underline{P}, \underline{Q}) = \underset{\gamma}{\operatorname{argmin}}\{J(\gamma, \underline{P}, \underline{Q})\} \tag{23}$$

En conséquence, on voit sur la figure 20 que γ^* est différent pour FC et pour GM. On peut critiquer le fait de chercher une seule valeur optimale pour γ alors que la modélisation 21 implique que les puissances P_i sont des *variables aléatoires.* C'est pourquoi, nous avons déterminé un intervalle de γ acceptable, défini par un seuil minimum placé sur la *p-value* de la statistique BP (tracé bleu de la figure 20). On a choisi de conserver les γ tels que :

$$\text{p-value}\{BP(\gamma)\} > 1\,\% \tag{24}$$

Cette limite sur *p-value* est équivalente à $BP(\gamma) > \texttt{qchisq}(0.99, 1) = 6{,}63$. Les résultats de ce travail de choix de transformation sont résumés dans le tableau suivant :

	Valeur optimale γ^*	Intervalle acceptable $[\gamma_{min}; \gamma_{max}]$
Fonds Caraïbes	2,28	[2,0; 2,6]
Grand Maison	2,90	[2,6; 3,2]

Pour la suite de cette étude, on a choisi d'utiliser indépendamment pour chaque ferme la valeur optimale de γ pour appliquer la transformation Φ_γ aux données. Sur la figure 21, le tracé des résidus $\hat{\varepsilon}_i$ en valeur absolue fonction du niveau de puissance prévue permet de vérifier l'efficacité de la transformation optimale. On constate que la régression linéaire sur les $\hat{\varepsilon}_i^2$ (tracé rouge) qui correspond au test de Breusch–Pagan donne une droite parfaitement horizontale.

3.4 Régression et inférence

À présent que la transformation Φ_γ est fixée à Φ_{γ^*}, on peut utiliser le modèle de régression (21) – où l'erreur est maintenant garantie d'être homoscédastique conditionnellement à la prévision – pour inférer un intervalle de confiance sur P^{1/γ^*}. Cet IC sera ensuite transformé grâce à $\Phi_\gamma^{-1} = \Phi_{1/\gamma}$ pour donner un IC sur la puissance mesurée P.

3.4.1 Construction d'un Intervalle de Confiance

À l'aide des résidus estimés $\hat{\varepsilon}_i$, on peut estimer la variance $\hat{\sigma}^2$ du modèle. On procède de la même façon qu'à la partie 2 pour aboutir à l'équation (10).

Sous réserve que les résidus suivent une loi Gaussienne, ce qui sera vu au paragraphe 3.4.2, l'intervalle de confiance exact $\mathcal{I}_\alpha$ de niveau $(1-\alpha)$ pour une nouvelle valeur P_{N+1}^{1/γ^*} vaut :

$$\mathcal{I}_\alpha = \left[\hat{P}_{N+1}^{1/\gamma^*} \pm t_{N-2}(1-\alpha/2)\hat{\sigma}\right] \tag{25}$$

où la valeur ajustée $\hat{P}_{N+1}^{1/\gamma^*}$ vaut :

$$\hat{P}_{N+1}^{1/\gamma^*} = \hat{\beta}_0 + \hat{\beta}_1 . Q_{N+1}^{1/\gamma^*} \tag{26}$$

Et encore une fois, le quantile $t_{N-2}(1-\alpha/2)$ de la loi de Student est sensiblement égal à celui d'une loi Gaussienne. On a tracé cet intervalle en ruban sur la figure 22 pour un niveau de confiance de 70 %, soit une valeur "intermédiaire" comme le conseille Pinson [24] et pour être dans le même ordre de grandeur que le niveau fourni par Metnext.

Une fois l'intervalle $\mathcal{I}_\alpha$ calculé par la formule simple (25), on calcule son image par Φ_γ^{-1} et l'on obtient un IC pour P. Il est tracé en figure 23. Par rapport à l'intervalle Metnext tracé sur la figure 18, notre IC s'élargit lorsque Q augmente. Il y a donc bien prise en compte de l'hétéroscédasticité conditionnelle à Q.

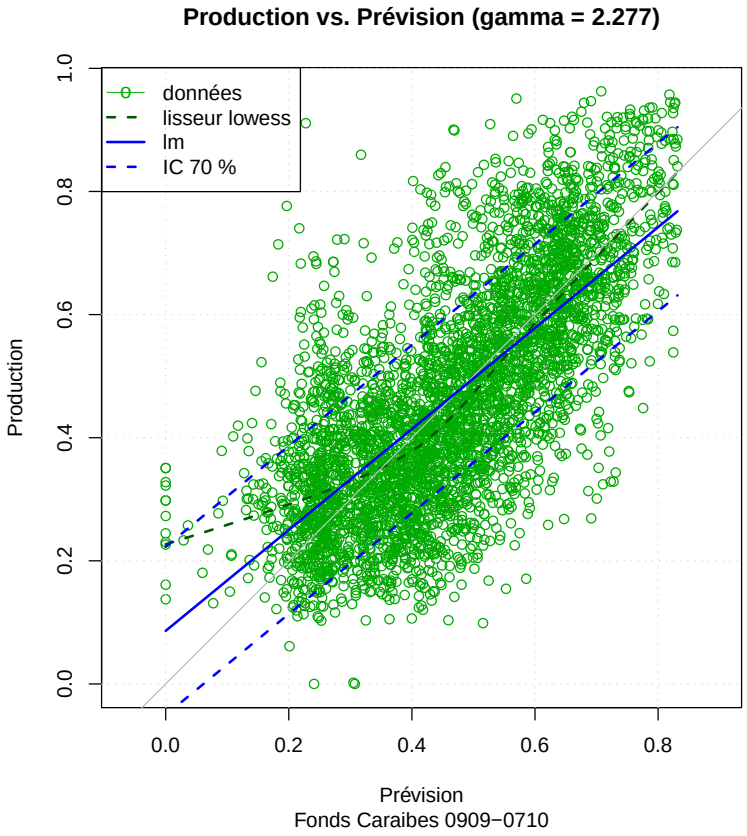

FIGURE 22 – Nuage de point $P_i^{1/\gamma} \sim Q_i^{1/\gamma}$ avec intervalle de confiance

Remarque sur la valeur des estimateurs On peut remarquer sur la figure 22 que le couple d'estimateur $(\hat{\beta}_0, \hat{\beta}_1)$ n'est pas égal à $(0, 1)$: la droite de régression (bleue) n'est pas confondue avec la droite identité (grise). C'est un constat un peu surprenant vu que Q est une prévision de P. Comme il existe une incertitude sur la détermination de ce couple on a tracé en figure 24 l'ellipse de confiance jointe sur leur valeur. On voit que la valeur attendue $(0, 1)$ est nettement en dehors de la région de confiance à 95% mais elle est bien dans la direction de plus grande incertitude.

Puisque $(\hat{\beta}_0, \hat{\beta}_1) \neq (0, 1)$ on peut être tenté de dire que la prévision de Metnext est biaisée, cependant ce jugement serait hâtif, car l'application de la transformation Φ_γ nous fait sortir du cadre des moindres carrés ordinaires sur $P \sim Q$.

De plus, nous soupçonnons que les points de mesures non disponibles (`NAs`) sont préférentiellement des points qui auraient été de valeur faible. Autrement dit les données sur lequel nous travaillons seraient issues d'un échantillonnage préférentiel. Cette supposition reste à vérifier, par exemple en étudiant la moyenne des points qui précèdent un point `NA`.

3.4.2 Analyse des résidus

Distribution des Résidus Pour pouvoir inférer un IC comme en 3.4.1, il faut que les résidus soient distribués selon une loi normale centrée $\mathcal{N}(0, \sigma^2)$. On vérifie sur la figure 25 que la distribution empirique est bien proche d'une gaussienne.

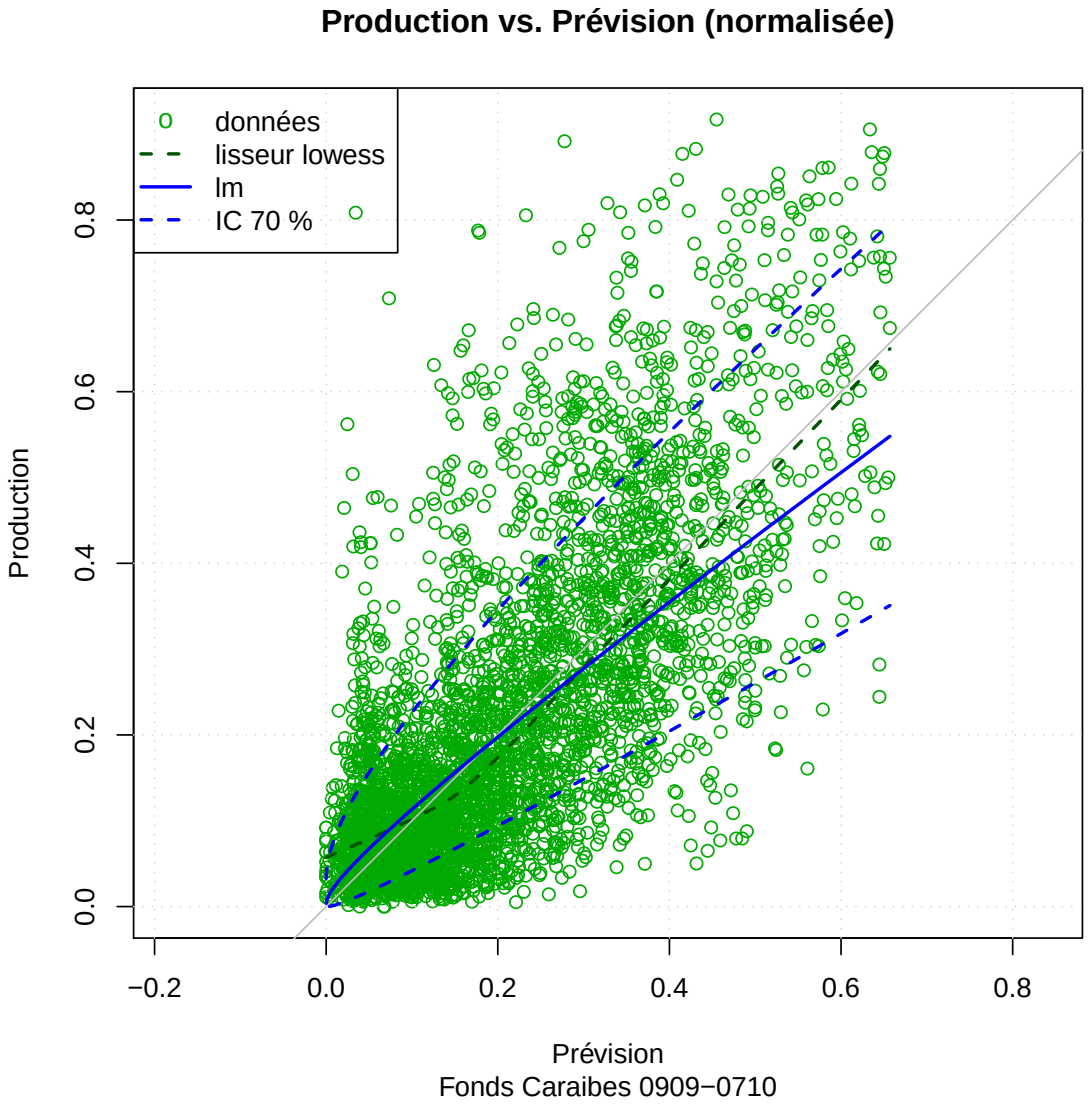

FIGURE 23 – Nuage de point $P_i \sim Q_i$ avec intervalle de confiance (25) retransformé

Corrélation des erreurs La méthode des moindres carrés ordinaires (MCO) ne donne un estimateur sans biais de *variance minimale* que si les résidus sont indépendants (théorème de Gauss–Markov). Or un simple tracé de la fonction d'autocorrélation des $\hat{\varepsilon}_i$ montre qu'elle est *significativement non nulle* jusqu'au rang 20 au moins (décalage équivalent à une journée). La corrélation au rang 1 est d'environ 75 %. Il faudrait donc se diriger vers la méthode des moindres carrés généralisés (MCG).

Pour connaître la covariance des résidus, on s'est intéressé une modélisation ARMA. La sélection des ordres (p, q) du modèle peut se faire grâce au critère d'information AIC. La fonction `auto.arima` du paquet `forecast` [12] automatise ce processus de sélection. On aboutit à un simple modèle AR(1) où le coefficient AR vaut $\Phi \approx 0{,}75$.

Alors que l'étude temporelle des résidus est encore en cours, on s'est posé la question de la stationnarité de l'erreur. Pour cela, on a découpé les résidus en 13 fenêtres de 600 points où l'on appliqué la modélisation AR(1) grâce à la fonction `arima(e, order=c(1,0,0))`.

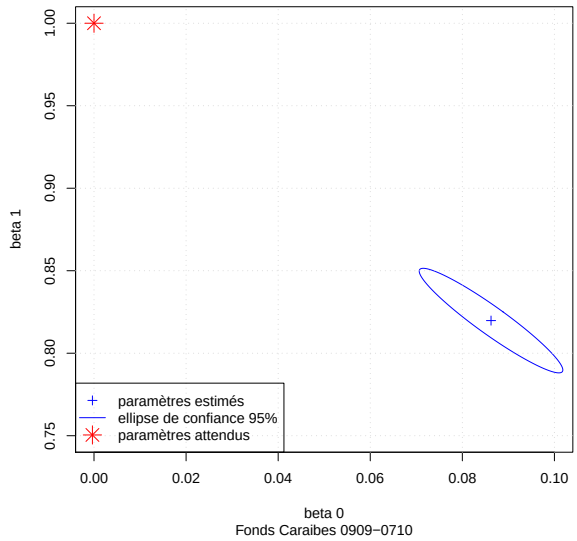

FIGURE 24 – Ellipse de confiance des estimateurs du modèle (21) pour $\gamma = 2{,}277$

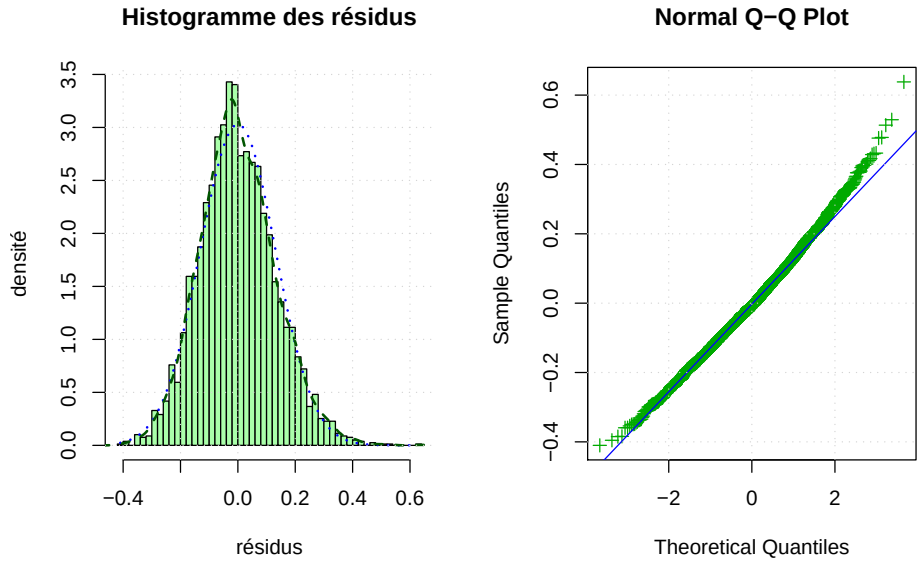

FIGURE 25 – Distribution des résidus de la régression (21)

On a cherché à savoir si les variations de l'estimée $\hat{\Phi}$ est liée à une variabilité normale ou bien à une non stationnarité. On a donc fait la comparaison avec des données AR(1) simulée grâce à `arima.sim`, sur lesquelles on a appliqué la même méthode de modélisation. La comparaison des résultats est donnée sur la figure 26. On constate que les données réelles présentent plus de disparités que les données simulées, mais les variations semblent encore acceptable.

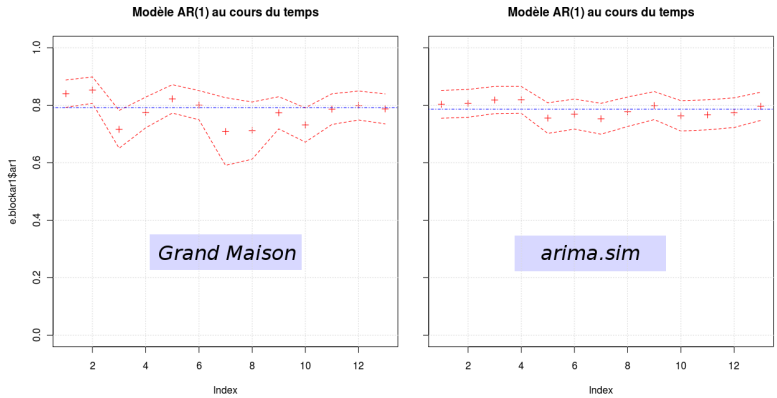

FIGURE 26 – Modélisation AR(1) sur des fenêtres temporelles pour étudier la stationnarité des résidus

4 Conclusion

Les études des parties 2 et 3 ont montré que la construction d'un intervalle de confiance (IC) sur des données de production éolienne doit se faire en prenant particulièrement soin du fait que la *variance conditionnelle dépend du niveau de puissance* (mesurée ou prévue). Nous avons proposé une transformation "gamma" simple – inspirée par la physique – qui transforme les variables en des quantités dont la variance conditionnelle est constante (résidus *homoscédastiques*). Cette transformation a permis de construire des IC grâce à l'hypothèse de gaussianité qui était raisonnable pour chacun des deux jeux de données étudiés dans ce stage.

Remerciements Je tiens à remercier Pascal Bondon pour m'avoir accueilli au LSS dans de très bonnes conditions et pour m'avoir judicieusement guidé dans ce stage alors que le sujet, parfois mouvant, se situe à la limite mal définie de plusieurs domaines a priori éloignés – génie électrique, statistiques & traitement du signal.

Je veux également remercier Bernard Multon, mon futur directeur de thèse, qui a rendu ce stage possible en guidant la collaboration entre le SATIE et le LSS et qui m'a fait bénéficier de ses conseils, malgré la distance.

Je remercie Stéphane Lascaud de EDF R&D. Sa large connaissance de l'état de l'art du stockage électrique sur réseau [15] m'a grandement aidé à appréhender le sujet. Je regrette de n'avoir pas pu étudier plus en détail la ferme de La Perrière qui lui tient particulièrement à cœur, mais ce n'est que partie remise !

Je remercie aussi les sociétés Vergnet et Aérowatt pour avoir accepté de partager les données qui sont à la base de ces travaux de stage et pour m'avoir aimablement accueilli en avril 2011, alors que le stage débutait.

4.1 Travaux futurs

Les modélisations effectuées dans ce stage sont exclusivement à caractère statique : le temps n'est pas pris en compte et les effets temporels ne sont qu'esquissés (partie 3.4.2). Comme l'a très bien souligné Pierre Pinson [23], il est crucial de prendre en compte ces effets, en particulier la corrélation de l'erreur, pour dimensionner un système de stockage.

Dans les semaines qui viennent, on souhaite procéder aux premiers essais de simulation temporelles d'un système de stockage. Ces simulation pourront utiliser un *rejeu des données*. Cependant, si l'on dispose d'une modélisation temporelle assez fiable, on peut imaginer simuler des données pour procéder à des simulations du type Monte Carlo qui pourraient aboutir à des estimations de meilleure précision que le simple rejeu.

Références

[1] Abbey, C., and Joos, G. Sizing and power management strategies for battery storage integration into wind-diesel systems. In *Industrial Electronics, 2008. IECON 2008. 34th Annual Conference of IEEE* (nov. 2008), pp. 3376–3381.

[2] Aubry, J., Bydlowski, P., Multon, B., Ben Ahmed, H., and Borgarino, B. Energy Storage System Sizing for Smoothing Power Generation of Direct Wave Energy Converters. In *3rd International Conference on Ocean Energy* (2010).

[3] Betz, A. Das Maximum der theoretisch möglichen Ausnützung des Windes durch Windmotoren. *Zeitschrift für das gesamte Turbinenwesen 26* (1920), 307–309.

[4] Breusch, T., and Pagan, A. A simple test for heteroscedasticity and random coefficient variation. *Econometrica : Journal of the Econometric Society* (1979), 1287–1294.

[5] Cornillon, P.-A., and Matzner-Løber, É. *Régression : théorie et applications*. Collection Statistique et probabilités appliquées. Springer-Verlag France, Paris, 2006.

[6] Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., and Feitosa, E. A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews 12*, 6 (2008), 1725–1744.

[7] Daniélo, O. El Hierro, l'Île Électrique. *Systèmes Solaires, le Journal des Énergies Renouvelables*, 201 (Jan. 2011), 88–97.

[8] Daniélo, O. La Norvège, future "Batterie de l'Europe" ? *Systèmes Solaires, le Journal des Énergies Renouvelables*, 201 (Jan. 2011), 98–99.

[9] Ewing, B. T., Kruse, J. B., and Schroeder, J. L. Time series analysis of wind speed with time-varying turbulence. *Environmetrics 17*, 2 (2006), 119–127.

[10] Faulstich, M., et al. Wege zur 100 % erneuerbaren Stromversorgung. Tech. rep., Sachverständigenrat für Umweltfragen (SRU), Luisenstraße 16, 10117 Berlin, Jan. 2011.

[11] Giebel, G., Brownsword, R., Kariniotakis, G. N., Denhard, M., and Draxl, C. The state-of-the-art in short-term prediction of wind power : A literature overview. Tech. rep., ANEMOS.plus, 2011.

[12] Hyndman, R. J. *forecast : Forecasting functions for time series*, 2011. R package version 2.19.

[13] Koenker, R. A note on studentizing a test for heteroscedasticity. *Journal of Econometrics 17*, 1 (1981), 107–112.

[14] Koenker, R., and Hallock, K. Quantile Regression. *The Journal of Economic Perspectives 15*, 4 (2001), 143–156.

[15] Lascaud, S. Accumulateurs/piles à combustibles, applications réseaux. In *Chaire du Développement durable - Environnement, Énergie et Société* (Mar. 2011), Collège de France.

[16] Li, Q., Choi, S., Yuan, Y., and Yao, D. On the Determination of Battery Energy Storage Capacity and Short-Term Power Dispatch of a Wind Farm. *Sustainable Energy, IEEE Transactions on 2*, 2 (april 2011), 148–158.

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

[18] Martí, I., Kariniotakis, G. N., Pinson, P., Sanchez, I., Nielsen, T. S., Madsen, H., Giebel, G., Usaola, J., Palomares, A., Brownsword, R., et al. Evaluation of advanced wind power forecasting models–results of the Anemos project. In *Proc. of EWEC* (2006).

[19] Nielsen, H. A., Madsen, H., and Nielsen, T. S. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy 9*, 1-2 (2006), 95–108.

[20] Pinson, P. *Estimation of the uncertainty in wind power forecasting.* PhD thesis, Ecole des Mines de Paris, Paris, 2006.

[21] Pinson, P., Chevallier, C., and Kariniotakis, G. N. Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power. *Power Systems, IEEE Transactions on 22*, 3 (aug. 2007), 1148–1156.

[22] Pinson, P., and Kariniotakis, G. N. On-line assessment of prediction risk for wind power production forecasts. *Wind Energy 7*, 2 (2004), 119–132.

[23] Pinson, P., Madsen, H., Nielsen, H. A., Papaefthymiou, G., and Klöckl, B. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy 12*, 1 (2009), 51–62.

[24] Pinson, P., Nielsen, H. A., Møller, J. K., Madsen, H., and Kariniotakis, G. N. Non-parametric probabilistic forecasts of wind power : required properties and evaluation. *Wind Energy 10*, 6 (2007), 497–516.

[25] Pinson, P., Papaefthymiou, G., Klöckl, B., and Verboomen, J. Dynamic sizing of energy storage for hedging wind power forecast uncertainty. In *Power Energy Society General Meeting, 2009. PES '09. IEEE* (july 2009), pp. 1–8.

[26] R Development Core Team. *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[27] Rapin, M., and Noël, J.-M. *Énergie éolienne. Principes. Études de cas.* Collection technique & ingénierie. Série Environnement & sécurité. Dunod, Paris, 2010.

[28] Sánchez-Squella, A., Ortega, R., Griñó, R., and Malo, S. Dynamic Energy Router. *Control Systems, IEEE 30*, 6 (dec. 2010), 72–80.

[29] Teleke, S., Baran, M., Bhattacharya, S., and Huang, A. Optimal Control of Battery Energy Storage for Wind Farm Dispatching. *Energy Conversion, IEEE Transactions on 25*, 3 (sept. 2010), 787–794.

[30] Teleke, S., Baran, M., Bhattacharya, S., and Huang, A. Rule-Based Control of Battery Energy Storage for Dispatching Intermittent Renewable Sources. *Sustainable Energy, IEEE Transactions on 1*, 3 (oct. 2010), 117–124.

[31] Teleke, S., Baran, M., Huang, A., Bhattacharya, S., and Anderson, L. Control Strategies for Battery Energy Storage for Wind Farm Dispatching. *Energy Conversion, IEEE Transactions on 24*, 3 (sept. 2009), 725–732.

[32] Zeileis, A., and Hothorn, T. Diagnostic Checking in Regression Relationships. *R News 2*, 3 (2002), 7–10.