ENSAE ParisTech
École nationale de la statistique et de l'administration économique

Année scolaire 2014-2015

Séminaire de Modélisation Statistique

Etude de la prédiction du prix de l'électricité en Allemagne avec régression LASSO

Encadrants
M.Vincent Cottet
M.Pierre Alquier
M.Mohamed Hebiri

Élèves
M.Alexandre de Larrard
M^{lle} Romane Persch
M.Alexandre de la Morinerie

8 janvier 2017

Sommaire

1 Résumé **3**

1.1 Contextualisation . . . 3

1.1.1 Présentation des données . . . 3

1.1.2 Les défis posés par la modélisation des prix de l'électricité . . . 4

1.2 Modélisation . . . 5

1.2.1 Cadre général : modèle autorégressif vectoriel (VAR) . . . 5

1.2.2 Gestion de la périodicité à l'aide de B-splines . . . 6

1.2.3 Modélisation de l'effet de levier à l'aide d'un modèle TARCH . . . 8

1.2.4 Résolution d'un problème avec une pénalité de type L1 (LASSO) afin de restreindre le nombre de paramètres . . . 9

1.3 Résultats obtenus . . . 10

1.3.1 Un modèle encore trop peu dépendant du passé . . . 10

1.3.2 Une dépendance quotidienne et hebdomadaire importante . . . 11

1.3.3 Un effet de levier correctement modélisé par le modèle auto-régressif TARCH . . . 11

1.3.4 Le prix de l'électricité diminue lorsque la quantité d'électricité produite par l'éolien et le solaire augmente . . . 12

1.3.5 Prévision du prix de l'électricité . . . 13

2 Confrontation avec l'article de Keles **14**

2.1 Des points de divergence fondamentaux . . . 14

2.2 Une approche sensiblement différente, axée sur l'estimation de séries temporelles . . . 14

2.3 Plus simpliste, l'approche de D. Keles obtient des résultats convaincants . . . 15

3 Modélisation du prix spot de l'électricité en Allemagne à partir de la production d'électricité d'énergies éoliennes **16**

3.1 Présentation des données et statistiques descriptives . . . 16

3.2 Le modèle AR(p) . . . 16

3.2.1 Présentation du modèle . . . 16

3.2.2 Choix du paramètre de pénalisation λ dans la régression LASSO . . . 17

3.2.3 Résultats de prédiction des 480 heures futures . . . 17

3.2.4 Comparaison des prédictions . . . 17

3.2.5 Comparaison avec une estimation par MCO . . . 18

3.3 Le modèle VAR(p) . . . 18

3.3.1 Présentation du modèle . . . 18

3.3.2 Choix des paramètres de pénalisation λ dans la régression LASSO . . . 19

3.3.3 Résultats obtenus . . . 19

3.3.4 Comparaison des modèles AR(p) et VAR(p) sur les 4 dernières semaines . . . 20

4 Annexe **22**

Abstract

Le marché de l'électricité s'est récemment libéralisé en Allemagne, tout comme dans de nombreux pays. La prévision des prix sur marché spot est ainsi devenue un sujet de recherche actif. Nous avons essayé ici d'en comprendre les principaux enjeux. Dans cette optique, le rapport s'articule en trois parties. Nous présentons en premier lieu l'article de F. Ziel qui propose un modèle de prédiction des prix spot de l'électricité en Allemagne dont l'originalité est l'utilisation du LASSO et la prise en compte de l'information disponible sur la production d'énergies renouvelables, de plus en plus importante dans ce pays. Dans un second temps, nous avons comparé cette approche avec celle de D. Keles, de façon à prendre plus de recul sur les solutions proposées par F. Ziel. Enfin, nous avons cherché à appliquer nous-mêmes dans ses grandes lignes la méthode établie par F. Ziel. La troisième partie nous a à ce titre permis de comprendre plus concrètement certains détails techniques, parfois peu explicités dans l'article. Elle met également en lumière l'intérêt d'avoir introduit une telle complexification.

Chapitre 1

Résumé

1.1 Contextualisation

La production d'électricité connaît aujourd'hui des mutations importantes qui affectent significativement ses prix. On peut souligner, entre autres, l'apparition d'autres sources de production telles que l'énergie solaire et éolienne, ainsi que l'introduction de prix négatifs sur le marché spot depuis 2008. La libéralisation des marchés de l'électricité ces dernières années a permis d'augmenter les volumes d'électricité vendus. Eu égard à la dépendance des ménages et des entreprises aux prix de l'électricité (chauffage, climatisation, appareils ménagers, production etc.), l'étude de ces derniers est devenue d'autant plus stratégique que les modèles actuellement mis en place insistent surtout sur la prédiction de la charge du réseau ou encore de la consommation en électricité et non sur celle des prix. Les transformations structurelles des marchés de l'énergie incitent à renouveler les modélisations statistiques en place pour prédire les prix de l'électricité, et ce particulièrement en Allemagne, en pleine phase de transition énergétique. Les énergies renouvelables occupent une place toujours plus prépondérante dans ce pays, comme le souligne la décision politique du gouvernement allemand d'arrêter progressivement le nucléaire à l'horizon 2022. Dans cette optique, l'étude propose de construire un modèle des prix de l'électricité en tenant compte à la fois des caractéristiques de l'électricité, de l'introduction des énergies renouvelables et des données de consommation. L'intégration de ces éléments permet de construire un modèle plus complet et plus précis que ceux précédemment élaborés.

L'article entend ainsi répondre aux principaux enjeux suivants :

- Intégrer les différents facteurs jouant sur les prix de l'électricité, tels que la saisonnalité (jours de la semaine, saison etc.) pour améliorer la qualité du modèle statistique
- Mettre en place un modèle flexible qui tient compte d'une saisonnalité variable d'une année sur l'autre (jours fériés)
- Détecter la présence d'un éventuel effet de levier sur les prix de l'électricité
- Démontrer l'impact de l'énergie éolienne et solaire sur les prix de l'électricité
- Obtenir un modèle plus précis et plus efficace de prévision des prix de l'électricité

1.1.1 Présentation des données

Le modèle sera ajusté sur les prix horaires d'électricité de l'European Power Exchange (EPEX) entre le 28 Septembre 2010 et le 1er Mai 2014. L'ensemble des données mises à disposition sont contenues dans trois tables distinctes :

- La première regroupe les prix spot horaires de l'électricité en Allemagne et en Autriche tirées de l'EPEX [1].

1. URL www.epexspot.com, valide le 11/05/2015

- La seconde recense les données de charge du réseau électrique en Allemagne, fournies par l'European Network of Transmission System Operators for Electricity (ENTSOE [2]).
- La troisième comprend les données d'alimentation horaire d'électricité en énergie solaire et éolienne, issues de la Transparency Platform of the European Energy Exchange (EEX [3]).

1.1.2 Les défis posés par la modélisation des prix de l'électricité

Impossibilité de stockage et inélasticité de la demande en électricité

Deux types de produits sont échangés : les **produits spot**, achetés dans le but d'être livrés le jour même ou dans un horizon proche, et les **produits à terme**, livrés par définition à plus long terme. Ce sont les premiers qui posent le plus problème. Les producteurs et distributeurs doivent être prêt à produire et livrer à tout instant de l'électricité, à un prix qui peut varier à une échelle horaire. Ce dernier reflète l'équilibre offre-demande à court terme. Les prix de court terme sont en effet soumis à une forte volatilité et connaissent des phénomènes de pics (*price spikes*).

Ceci est lié à deux caractéristiques originales qui doivent être absolument soulignées : **l'absence de capacité technologique de stockage de l'électricité** ainsi que **l'inélasticité de la demande d'électricité.** La première a pour conséquence de créer des pics de consommation, et donc des pics de prix sur les marchés. Une fois que cette dernière est produite, il n'est pas possible de la stocker quelque part, de façon à éviter les gaspillages. Il s'agit d'une source de risque singulière et non-partagée par d'autres sources d'énergie. La seconde vient du fait que la plupart des consommateurs ne sont pas en mesure d'observer ni de réagir aux évolutions des prix en temps réel : la sensibilité de la consommation aux prix est donc très faible. La demande de court terme est donc inélastique au prix.

On peut aussi évoquer comme source de volatilité et de phénomènes de pics **l'influence des conditions climatiques** fortement variables sur la production et consommation d'électricité (l'absence de vent provoque une chute de la production éolienne en Allemagne, des vagues de froid entraînent une augmentation de la demande en électricité pour se chauffer) ainsi que **l'impact des décisions politiques** (abandon du nucléaire en Allemagne se traduisant par une transition énergétique nécessaire ayant nécessairement un impact sur la production d'électricité).

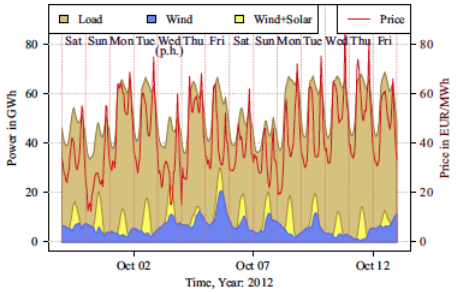
L'impact des énergies renouvelables sur les prix d'électricité est majeur

Une mesure importante prise par le gouvernement allemand et qui mérite d'être soulignée est la Erneuerbare-Energien-Gesetz (EEG). Elle fait la promotion des énergies renouvelables, incitant les producteurs à augmenter la part de ces dernières dans la production totale d'électricité à travers notamment des subventions, ou des primes à la vente sur les marchés financiers. L'effet attendu est double : d'une part une augmentation significative de la production d'électricité via les énergies vertes (il y a généralement plus de vent en hiver, l'énergie éolienne fonctionne donc bien durant cette saison, et répond mieux aux pics de demande en électricité durant cette période), d'autre part une tendance à la baisse des prix de l'électricité. De fait, le coût marginal des énergies éoliennes et solaires est proche de zéro contrairement aux autres sources de production de l'électricité. Par conséquent, les énergies renouvelables, moins chères, remplacent progressivement les autres technologies qui produisent à des coûts marginaux plus élevés, entraînant ainsi une baisse à plus ou moins long terme des prix de gros de l'électricité.

Cela implique que la modélisation de l'impact des énergies renouvelables en Allemagne est nécessaire, vu que la part croissante de ces nouvelles sources d'énergie dans la production d'électricité conduira toujours à une structure des coûts marginaux modifiée, et donc à une structure des prix modifiée (selon le modèle de tarification au coût marginal, énoncé plus haut).

2. URL www.entsoe.eu, valide le 11/05/2015.
3. URL www.transparency.eex.com, valide le 11/05/2015.

Le graphique 1.1 illustre le phénomène décrit ci-dessus : les prix de l'électricité (en rouge) semblent difficilement prévisibles, étant très volatiles, et ont tendance à descendre lorsque la production d'énergie solaire et éolienne (en jaune) augmente dans la production totale d'électricité (en marron). On peut également observer que la série des prix de l'électricité est hétéroscédastique : elle présente une forte volatilité variable selon les jours de la semaine, avec de forts pics de tailles inégales au cours du temps.



Graphique 1.1 – Structure des prix et la production d'électricité en Allemagne

Impact de la saisonnalité sur les prix de l'électricité

Les prix de l'électricité sont fortement dépendants du jour de la semaine, des saisons annuelles ainsi que des jours fériés, plus variables. La première raison est que la production et la consommation d'énergie solaire et éolienne dépendent toutes deux du climat, et donc des saisons. Ainsi, la demande atteindra un pic annuel en hiver, puisque c'est à cette période que les besoins en consommation d'électricité sont les plus forts et les plus imprévisibles à cause du froid et de la plus grande volatilité des températures. La production sera de même plus élevée durant cette saison. La seconde raison est que la demande d'électricité dépend des jours de travail, et donc des jours de congés. La demande d'électricité, et en particulier la demande commerciale, sera plus élevée les jours de la semaine où la population active travaille puisque les entreprises ont à ce moment là de forts besoins en énergie. Le graphique 1.1 souligne cette corrélation entre jours de congés et consommation d'électricité : le 3 Octobre est le jour de l'Unité allemande, durant lequel personne ne travaille. On observe aisément que les prix ce jour là sont significativement moins élevés que les autres jours de la semaine, et ressemble plus à un jour dominical.

Effet de levier et prix négatifs

Comme sur les marchés financiers, il semble que la volatilité réagit ici de façon différente à une forte augmentation et à une forte baisse des prix. C'est ce qu'on appelle l'effet de levier. **L'effet de levier "traditionnel"** se définit comme une volatilité accrue à la suite de chocs négatifs (par rapport à la volatilité observée à la suite de chocs positifs). **L'effet de levier inverse** se définit, par opposition, comme une volatilité accrue cette fois à la suite de chocs positifs. Les prix de l'électricité semblent plutôt connaître ce dernier phénomène, ce que l'article s'attache à démentir par la suite.

Les prix négatifs, eux, ont été autorisés en Allemagne depuis 2008 sur les marchés de l'énergie. Le problème provient du fait que de telles données empêchent certaines transformations des séries temporelles, comme la transformation logarithmique, et qui empêchent ainsi l'utilisation de certaines modélisations classiques. Un prix négatif des prix spot correspond à une situation où l'acheteur reçoit à la fois de l'argent et de l'électricité des vendeurs. Cela peut se produire lorsqu'il y a une forte production non-flexible d'électricité (énergie solaire, éolienne), couplée avec une faible demande. Le surplus est écoulé notamment *via* des prix négatifs. Une telle situation arrive fréquemment en Allemagne du fait de la part croissante des énergies renouvelables dans sa production totale d'électricité.

1.2 Modélisation

Pour répondre aux différents objectifs évoqués ci-dessus, les auteurs de l'article utilisent une modélisation combinant trois méthodologies classiques de façon à flexibiliser le modèle autorégressif vectoriel (VAR).

1.2.1 Cadre général : modèle autorégressif vectoriel (VAR)

Un aspect essentiel de l'article est d'intégrer dans la modélisation des prix des informations sur la quantité totale d'électricité présente sur le réseau (*load*) et sur la quantité produite d'électricité d'origine solaire et éolienne. Les auteurs choisissent de formaliser ceci en cherchant à prédire le processus multivarié $(Y_t^{\mathscr{P}}, Y_t^{\mathscr{L}}, Y_t^{\mathscr{R}}) \in \mathbb{R}^3$ indiquant respectivement le prix de l'électricité, la quantité d'électricité présente sur le réseau et la quantité d'électricité

d'origine solaire ou éolienne à l'heure t. Le prix de l'électricité n'occupe donc pas de place spécifique dans la modélisation par rapport aux variables indiquant la quantité d'électricité.

Le modèle autorégressif vectoriel (VAR) classique est alors considéré, à la différence près que les coefficients ϕ et μ sont dépendants du temps t :

$$Y_t^i = \mu^i(t) + \sum_{j=\mathscr{P},\mathscr{L},\mathscr{R}} \sum_{k\in I_{i,j}} \phi_k^{i,j}(t) Y_{t-k}^j + \epsilon_t^i$$

où $i = \mathscr{P},\mathscr{L},\mathscr{R}$ et $I_{i,j}$ indique les lags considérés [4].

L'idée du modèle VAR classique est simplement de prédire la quantité Y à la date t à partir des quantités strictement antérieures. Le prix à la date t est donc ici prédit à partir des prix passés mais aussi des quantités d'électricité antérieures (totale et renouvelable). Néanmoins, en aucun cas le prix à la date t n'est prédit en utilisant les quantités produites d'électricité à la date t : ceci ne serait pas applicable dans la réalité.

Il faut noter que le choix des lags pris en compte (ensembles $I_{i,j}$) se fait ici de façon plus ou moins arbitraire : il est fondé sur l'observation des données à l'aide d'outils de statistiques descriptives simples (graphiques, étude des autocorrélations etc) et sur la lecture d'autres articles à ce sujet. Il est nécessaire de faire une sélection préalable car sinon le modèle contiendrait trop de paramètres. Un très grand nombre de paramètres dans un modèle à objectif prédictif peut en effet créer des problèmes de surapprentissage : une trop grande complexité de la modélisation conduit à une estimation des coefficients correspondant presque à un "apprentissage par coeur" des variables à prédire. Lorsque le modèle est alors appliqué sur de nouvelles données, la prédiction devient très mauvaise. Le modèle peut également devenir trop lent à estimer.

1.2.2 Gestion de la périodicité à l'aide de B-splines

Une approche très intuitive pour gérer la périodicité hebdomadaire des prix de l'électricité aurait par exemple été de n'utiliser que les lag de type $t - 168k$ au lieu de chercher à faire dépendre les coefficients ϕ du temps t. On aurait également pu introduire des indicatrices du jour de la semaine devant le coefficient μ. L'article introduit néanmoins une modélisation plus complexe pour répondre aux trois contraintes majeures évoquées plus haut :

- Les périodicités sont différentes, selon que l'on s'intéresse au prix de l'électricité, à la quantité totale d'électricité ou à la quantité d'origine renouvelable.
- Plusieurs types de périodicités se superposent : périodicité journalière, hebdomadaire et annuelle. Le motif observé à une date est donc le résultat de la combinaison de différents motifs récurrents.
- Il n'y a pas dans la réalité de périodicité parfaite à cause des jours fériés et des vacances, dont la date change chaque année.

Pour répondre à ces enjeux, les auteurs de l'article utilisent ainsi des B-splines uniformes cubiques [5] dont les noeuds ont des distances différentes : par exemple, 4 heures pour gérer la périodicité **hebdomadaire et journalière** du prix $\mathscr{P}$ et de la quantité d'électricité totale $\mathscr{L}$ et 1460.96 heures pour gérer leur périodicité **annuelle**.
Pour chaque type de périodicité, le choix de telles distances (par exemple, le fait de choisir 4 heures et non pas 2 heures) résulte d'un compromis entre le nombre de paramètres et le niveau de précision de la mesure des effets temporels : plus la distance est faible, plus la précision est grande mais plus le nombre de paramètres augmente. Le fait de combiner des B-splines uniformes de distances extrêmement différentes entre les noeuds permet quant à lui de capter à la fois des effets de longue période (annuelle) et des effets de période très courte (journalière) comme cela sera explicité plus bas.

4. Par exemple, dans l'article $I_{\mathscr{P},\mathscr{R}} = 1,...,49$. Cela signifie qu'on suppose que la quantité d'électricité d'origine renouvelable n'a pas d'impact durant plus de 2 jours sur le prix.
5. Pour la définition mathématique précise, voir http://www.math.u-psud.fr/~pansu/web_maitrise/bsplines.pdf

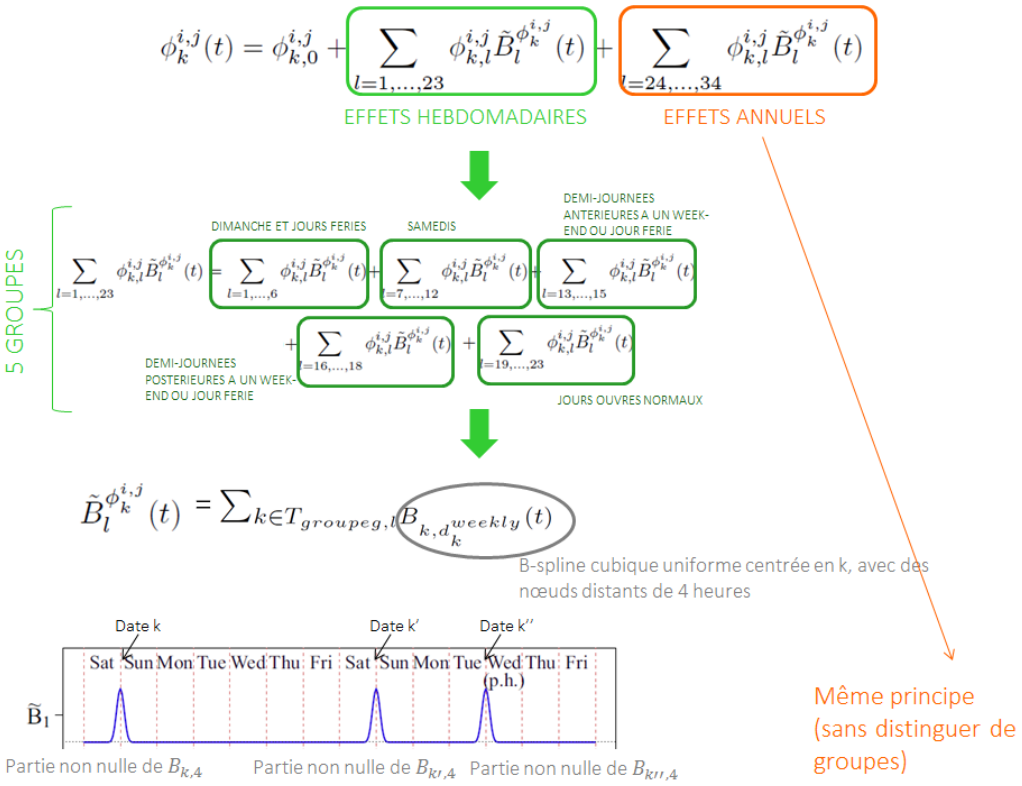
Les B-splines cubiques sont en effet traditionnellement utilisées pour approximer une courbe (*Curve fitting*) en trouvant les coefficients α_i qui résolvent le problème de minimisation $Min \sum_t (y(t) - \sum_i \alpha_i B_{i,3}(t))^2$. Dans le cas présenté ici, elles permettent donc en un sens **d'approcher la "tendance"** et **d'approcher les effets du temps sur le niveau d'impact des différents lags** (les quantités Y_{t-k}^j). Le coefficient μ et les coefficients ϕ sont donc décomposés ainsi pour $i = \mathscr{P}, \mathscr{L}$:

$$\mu^i(t) = \mu_0^i + \mu_{lin}^i + \sum_{l=1,\dots,23} \mu_l^i \tilde{B}_l^{\mu^i}(t) + \sum_{l=24,\dots,34} \mu_l^i \tilde{B}_l^{\mu^i}(t)$$

$$\phi_k^{i,j}(t) = \phi_{k,0}^{i,j} + \sum_{l=1,\dots,23} \phi_{k,l}^{i,j} \tilde{B}_l^{\phi_k^{i,j}}(t) + \sum_{l=24,\dots,34} \phi_{k,l}^{i,j} \tilde{B}_l^{\phi_k^{i,j}}(t)$$

Néanmoins, ici les $\tilde{B}_l$ sont **des sommes de B-splines et non des B-splines simples. Ceci permet de modéliser la périodicité.** Les $\tilde{B}_l$ pour **l=1,...,23** sont des sommes de B-splines servant à capter la périodicité **hebdomadaire/journalière** et les $\tilde{B}_l$ pour **l=24,...,34**[6] sont des sommes de B-splines servant à capter la périodicité **annuelle**.

Graphique 1.2 – Schéma de la décomposition des fonctions de base



Pour l allant de 1 à 23, la décomposition des $\tilde{B}_l$ en somme de B-splines permet alors de faire en sorte que par exemple tous les dimanches et jours fériés 8h aient le même coefficient $\phi_{k,l}^{i,j}$. Le schéma 1.2 détaille la procédure utilisée dans l'article. Ainsi, tous les dimanches et jours fériés 8h ont le même début du coefficient $\phi_k^{i,j}(t)$ et le même début de la partie non linéaire de $\mu(t)$. De même, la fin de ces coefficients sera identique pour tous les temps t' correspondant à la même période de l'année (même heure même date) de façon à prendre en compte la récurrence annuelle d'un motif. On voit en observant la fonction $\tilde{B}_1$ sur le schéma que les jours fériés sont biens détectés avec ce système.

6. Notons une coquille dans l'article : il y a 12 fonctions de base annuelles mais, comme pour les fonctions de base hebdomadaires, il est nécessaire de supprimer la dernière car sinon leur somme vaudrait constamment 1. Cela poserait ensuite des problèmes de singularité dans l'estimation (hypothèse d'absence de multicolinéarité parfaite entre les variables non vérifiée).

Un exemple pour mieux comprendre : Le coefficient $\phi_k^{i,j}$(dimanche 23 septembre 2012 8h) vaut $\phi_{k,0}^{i,j}+\sum_{l=1,...,23}\phi_{k,l}^{i,j}\tilde{B}_l$(dimanche 23 septembre 2012 8h) $+\sum_{l=24,...,34}\phi_{k,l}^{i,j}\tilde{B}_l$(dimanche 23 septembre 2012 8h).
-Pour $l=1,...,23$: $\tilde{B}_l$(dimanche 23 septembre 2012 8h) a la même valeur pour tous les dimanches ou jours fériés 8h. Ainsi, $\sum_{l=1,...,23}\phi_{k,l}^{i,j}\tilde{B}_l$(dimanche 23 septembre 2012 8h) est identique pour tous les dimanches ou jours fériés 8h.
-Pour $l=24,...,34$: $\tilde{B}_l$(dimanche 23 septembre 2012 8h) a, une valeur identique pour tous les 23 septembre 8h. Ainsi, $\sum_{l=24,...,34}\phi_{k,l}^{i,j}\tilde{B}_l$(dimanche 23 septembre 2012 8h) est identique pour tous les 23 septembres 8h.

On réussit ainsi bien à modéliser la superposition de motifs se répétant à des intervalles de temps de longueur différente, et ce de façon flexible, c'est-à-dire en détectant les jours spéciaux comme les jours fériés. Nous pouvons alors noter que ce modèle présente un énorme avantage dans la réalité pour les prévisions de court terme (par exemple, si une entreprise comme EDF souhaitait prédire les prix) : il suffit, lorsqu'un jour spécial est anticipé, même en cas d'événement extraordinaire (par exemple, un couvre-feu obligeant les gens à rester chez eux comme un dimanche), de placer les incides correspondant aux 24h de cette journée dans le groupe le plus proche (dimanche, samedi, jour ouvré normal etc) et les prédictions peuvent être rapidement réajustées.

On procède de même pour tenir compte de la périodicité journalière et annuelle de l'énergie renouvelable. Il faut également noter qu'un ajustement des fonctions de base est effectué pour tenir compte des 2 changements d'heures annuels sur la quantité d'électricité d'origine renouvelable. Ils ont en effet un forte impact sur la production d'energie solaire : la "périodicité" est translatée.

Le modèle finalement obtenu est [7] :

$$Y_t^i = \mu_0^i + \mu_{lin}^i + \sum_{l=1,...,M(\mu^i)} \mu_l^i \tilde{B}_l^{\mu^i}(t) + \sum_{j=\mathscr{P},\mathscr{L},\mathscr{R}} \sum_{k\in I_{i,j}} \phi_{k,0}^{i,j} Y_{t-k}^j + \sum_{j=\mathscr{P},\mathscr{L},\mathscr{R}} \sum_{k\in L_{i,j}} \sum_{l=1,...,M(\phi_k^{i,j})} \phi_{k,l}^{i,j} \tilde{B}_l^{\phi_k^{i,j}}(t) Y_{t-k}^j + \epsilon_t^i$$

Le modèle se réécrit donc sous forme matricielle de la façon suivante :

$$Y^i = X^i\theta^i + \epsilon^i$$

où $i=\mathscr{P},\mathscr{L},\mathscr{R}$.

1.2.3 Modélisation de l'effet de levier à l'aide d'un modèle TARCH

Comme expliqué dans la partie 1, l'électricité semble connaître des effets de leviers inverses : l'envolée des prix est plus forte lors d'un choc positif que la baisse lors d'un choc négatif. L'un des objectifs de l'article est ainsi d'en démontrer ou d'en infirmer l'existence. Les auteurs utilisent pour cela un modèle TARCH. Ces modèles sont initialement utilisés en finance pour modéliser les effets de leviers que connaît le prix des actifs sur les marchés financiers.

Pour modéliser cette hétéroscédasticité, le terme d'erreur est décomposé comme $\epsilon_t^i = \sigma_t^i Z_t^i$ où les $(Z_t^i)_t$ sont i.i.d de moyenne nulle et de variance constante 1. Les $(Z_t^i)_t$ sont donc des résidus homoscédastiques. La volatilité σ_t^i est alors modélisée en fonction des précédents chocs, positifs et négatifs, en permettant des coefficients différents devant les chocs positifs et devant les chocs négatifs de manière à laisser la possibilité d'un effet de levier :

$$\sigma_t^i = \alpha_0^i(t) + \sum_{k\in J_i} \alpha_k^{+,i}\epsilon_{t-k}^{+,i} + \alpha_k^{-,i}\epsilon_{t-k}^{-,i}$$

où $\epsilon_{t-k}^{+,i} = max(\epsilon_{t-k}^i, 0)$ (intensité des chocs positifs) et $\epsilon_{t-k}^{-,i} = max(-\epsilon_{t-k}^{-,i}, 0)$ (intensité des chocs négatifs). Cela permet, après estimation des coefficients, de voir si les chocs positifs ont plus d'impact que les chocs négatifs. $\alpha_0^i(t)$ est ici décomposé avec les mêmes fonctions de bases que $\mu(t)$ [8] pour tenir compte des effets de périodicité.

7. Il faut noter que pour limiter le nombre de paramètres de type ϕ les auteurs utilisent des ensembles restreints $L_{i,j} \subseteq I_{i,j}$.
8. mais sans la tendance linéaire

Après réécriture, il est possible d'estimer les coefficients α en régressant la valeur absolue des résidus ϵ_t^i sur les parties positives et parties négatives des lags ϵ_{t-k}^i estimés à l'aide du modèle précédent.
La critique généralement adressée lors de l'utilisation de modèles de type ARCH/GARCH en finance de marché, notamment à la suite de la crise financière de 2008, reste ici valable : l'idée que la volatilité future s'explique principalement par le passé semble peu probable. Néanmoins, dans le cas de l'électricité, l'objectif n'est pas uniquement de prédire la volatilité future, il est surtout de démontrer l'existence d'un effet de levier afin d'alimenter la théorie économique sur le comportement de ce marché. Cette modélisation permet alors bien de répondre à cette question : il devient en effet alors possible de tester mathématiquement l'hypothèse d'absence d'effet de levier en testant l'hypothèse nulle $\mathbb{H}_0 : \sum_{j=1}^{k}(\alpha_j^{+,i} - \alpha_j^{-,i}) = 0$.

1.2.4 Résolution d'un problème avec une pénalité de type L1 (LASSO) afin de restreindre le nombre de paramètres

Même si un certain nombres de lags ont été mis de côté de façon plus ou moins arbitraire, le modèle possède toujours un très grand nombre de paramètres : environ 3500. Il est ainsi très probable qu'une partie non négligeable de variables n'aient pas d'impact significatif, autrement dit qu'elles ne soient pas "utiles" à la prédiction. L'introduction d'une pénalité de type L1 dans le problème de minimisation des moindres carrés ordinaire permet alors de forcer les coefficients placés devant les variables sans impact significatif à prendre la valeur exacte 0. Il s'agit là d'un modèle appelé LASSO (*Least Absolute Shrinkage and Selection Operator*).

Néanmoins, dans le modèle LASSO classique, les résidus sont supposés homoscédastiques, ce qui n'est pas le cas ici comme expliqué ci-dessus. En effet, on a : $Y^i = X^i\theta^i + \sigma_t^i Z_t^i$ où σ_t^i dépend de t. Le modèle LASSO est donc combiné à un modèle des moindres carrés pondérés (*Weighted Least Squares*). L'idée est de réécrire le modèle sous la forme : $\frac{1}{\sigma_t^i}Y_t^i = \frac{1}{\sigma_t^i}X_t^i\theta^i + Z_t^i$. On a alors bien Z_t^i homoscédastique. Le problème de minimisation considéré est donc le suivant [9] :

$$\hat{\theta}_i = argmin_{\theta\in\mathbb{R}^{p_i}} \sum_{t=1}^{n} w_t^i \left(Y_t^i - X_t^i\theta^i\right)^2 + \lambda_{i,n}\sum_{j=1}^{p_i}|\theta_j|$$

où $(w_1^i, ..., w_n^i)$ est un vecteur de poids.

Comme on ne connaît pas les volatilité σ_t^i, on ne peut pas utiliser directement les poids $(\frac{1}{(\sigma_1^i)^2}, ..., \frac{1}{(\sigma_n^i)^2})$. Ainsi, les poids prennent initialement tous la valeur 1 : il s'agit de la modélisation LASSO classique dans laquelle les résidus ϵ_t^i sont supposés homoscédastiques. Les volatilités σ_t^i sont alors estimées. Les poids sont donc réajustés en fonction de cette estimation, et les coefficients réestimés à l'aide de ces nouveaux poids, et ainsi de suite. L'algorithme formalisé est visible en Figure 1.3.
L'algorithme s'arrête lorsque pour tout $i = \mathscr{P}, \mathscr{L}, \mathscr{R}$, $\frac{1}{n}||\sigma_{K-1}^i - \sigma_K^i||_1 \leq 0.001$, autrement dit lorsque la volatilité estimée a convergé vers une certaine valeur. $K = 4$ itérations suffisent alors en pratique.

Les auteurs de l'article mettent en évidence un inconvénient majeur de leur modèle : la convergence et la normalité asymptotique des estimateurs des coefficients θ dans un modèle LASSO à la fois auto-régressif et hétéroscédastique n'a pas été démontrée dans la littérature. Certains papiers s'attachent néanmoins à étudier séparément le LASSO auto-régressif, et d'autres séparément le LASSO hétéroscédastique. De plus, **dans un objectif pur de prévision, cette absence de justification théorique pose moins problème** : les résultats pratiques sont ce qui importe le plus dans la décision d'utiliser le modèle ou non.

Enfin, notons que **les auteurs de l'article explicitent peu la méthode utilisée pour choisir le terme de pénalisation** $\lambda_{i,n}$. Il s'agit probablement d'une méthode de validation croisée comme cela est souvent le cas, mais

9. Notons ici aussi une erreur de frappe de l'article : le vecteur de poids W se place avant la parenthèse, il pondère non seulement X mais aussi Y.

Graphique 1.3 – Algorithme utilisé

1) Set the initial $d \times n$ dimensional weight matrix $\mathbf{W} = (1, \ldots, 1)$ and the iteration parameter $K = 1$.
2) Estimate Eq. (1) using LARS-lasso method with weights $\mathbf{W}$.
3) Estimate σ_t by Eqs. (5) and (6) with $|\hat{\varepsilon}_t|$ as absolute residuals from 2) using the NNLS algorithm.
4) Redefine $\mathbf{W} = (\mathbf{w}^1, \ldots, \mathbf{w}^d)$ by $\mathbf{w}^i = \left(\left(\hat{\sigma}_1^i\right)^{-2}, \ldots, \left(\hat{\sigma}_n^i\right)^{-2}\right)$ with $\hat{\boldsymbol{\sigma}}_t = \left(\hat{\sigma}_t^1, \ldots, \hat{\sigma}_t^d\right)$ as fitted values from 3).
5) Stop the algorithm, if a stopping criteria is satisfied, otherwise $K = K + 1$ and go back to 2).

l'article n'indique pas quelles sous-bases sont comparées, ni le critère de comparaison utilisé. Comment sont-réalisées les prédictions ? Quelle erreur est comparée ? Il faut à ce titre bien remarquer que qu'une validation croisée de type K-Fold "traditionnelle" (telle qu'elle peut être précodée dans un package) fondée sur l'erreur quadratique moyenne (ou *Mean Squared Error*) ne fonctionne pas ici. D'une part, il serait en effet un non sens de répartir les lignes de la base de données au hasard dans K groupes de même taille car il s'agit d'une série temporelle : les lignes ne sont pas i.i.d. De plus, même si l'on divisait la base en K groupes en conservant l'ordre des données, de nombreuses colonnes de la base de test X sont des lags de la variable à prédire Y. Il faut donc absolument réaliser des prédictions de manière "emboîtée" [10] comme le font les auteurs de l'article dans leur partie 6, et en aucun cas considérer les colonnes de la base de test X comme des données connues. On peut néanmoins raisonnablement supposer que les auteurs utilisent les mêmes N = 506 échantillons qu'en partie 6 [11] pour réaliser une version adaptée de la "N-Fold Cross Validation" dans laquelle les prédictions sont effectuées de façon emboîtée comme indiqué dans l'article et où l'erreur comparée entre les différents λ testés est $\frac{1}{Nh}\sum_{j=1}^{h}\sum_{k=1}^{N}|Y_{j,k} - \hat{Y}_{j,k}|$. Enfin, remarquons que le même λ semble utilisé pour toutes les itérations K, mais cela n'est pas précisé explicitement.

1.3 Résultats obtenus

1.3.1 Un modèle encore trop peu dépendant du passé

L'étude des autocorrélations des résidus et de leur valeur absolue, notés $\hat{Z}$, montre que ceux lié à la charge ne semblent pas stationnaires. Cela peut s'expliquer soit par la présence d'une dépendance temporelle non linéaire pour la charge, soit par la présence d'une dépendance temporelle plus importante avec son passé que celle prise en compte dans le modèle. En effet, les niveaux d'auto-corrélations des résidus $\hat{Z}_t^i$ tel que $\sigma_t^i \hat{Z}_t^i = \epsilon_t^i$ pour la modélisation de la charge d'électricité sont significativement non nuls pour des lags supérieurs à 300, ce qui signifie que les résidus ne sont pas stationnaires. Cependant, il semblerait que l'hypothèse selon laquelle les résidus sont stationnaires semble globalement vérifiée pour les résidus liés au prix d'électricité et à la production d'électricité par énergie renouvelables comme le montre le fait que les amplitudes des auto-corrélations restent majoritairement dans l'intervalle de confiance de celui d'un bruit blanc [12].

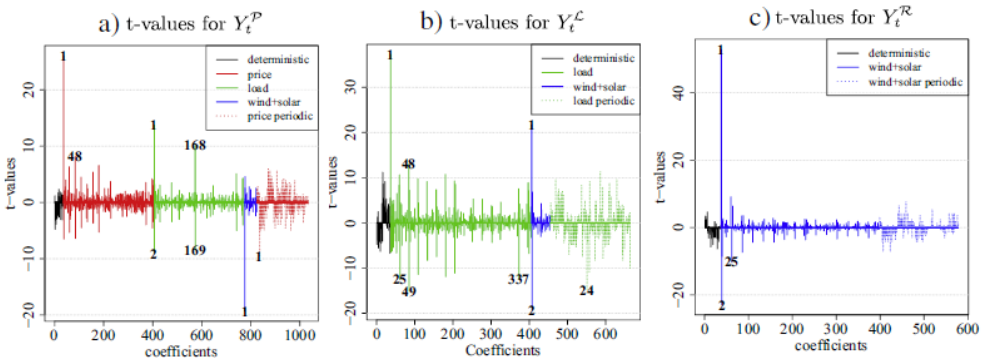
De plus, le fait que la valeur absolue des résidus aient des auto-corrélations relativement significativement nulles conforte le fait que le modèle est homoscédastique. En effet, cela montre que la variance des résidus ne semblent pas dépendre du temps. Ce résultat valide le modèle TARCH appliqué pour modéliser la dépendance temporelle de la volatilité pour le prix spot d'électricité, sa charge et la quantité d'électricité produite par voie solaire et éolienne.

10. La prévision de Y_{t+h} à partir de Y_t se fait étape par étape tel que : $\hat{Y}_{n+h} = g(\hat{Y}_{n+h-1}, \hat{Y}_{n+h-2}, \ldots)$
11. divisés de la même manière qu'en partie 6 en base de test / base d'apprentissage
12. IC $= \frac{\pm 1.96}{\sqrt{n}}$ pour un niveau de significativité de 5%

1.3.2 Une dépendance quotidienne et hebdomadaire importante

L'étude de la valeur des coefficients estimés par la régression LASSO permet de valider les lags sélectionnés en amont pour chaque série temporelle. Afin d'observer la non nullité de l'ensemble des coefficients estimés, la représentation des t-value [13] semble être un choix judicieux. Ainsi plus la t-value d'un lag sera élevée, plus le coefficient sera significatif.

Graphique 1.4 – Représentation des t-values de chaque coefficient par série temporelle



On observe alors sur le graphique 1.2 que :

- **Le prix** à l'instant t est fortement influencé par celui aux mêmes horaires dans les jours passés. On observe également que le prix est dépendant de la charge d'électricité une semaine avant l'instant t. Le prix semble également posséder une dépendance forte avec l'énergie solaire et éolienne produite mais uniquement à court terme.
- **La charge électrique** semble dépendre, tout comme pour le prix, des valeurs prises quelques heures auparavant et à la même heure pour les jours précédents. Cette dépendance périodique semble s'estomper rapidement lorsqu'on avance dans le passé. Tout comme pour le prix, il semblerait que la charge sur le réseau électrique dépende à très court terme de l'électricité éolienne et solaire, de l'ordre d'une et 2 heures avant t.
- **L'électricité produite par l'éolien et le solaire** à l'instant t semble dépendre de celle produite les 2 heures avant t puis quotidiennement pour ces mêmes horaires, avec une dépendance s'estompant rapidement. Cette dépendance à très court terme du passé semble logique étant donné que la météo n'est que peu modifiée sur deux heures d'intervalles mais varie rapidement d'un jour à l'autre. Météo France effectue notamment ses prévisions à 10 jours près car des prévisions trop lointaines dans le temps sont trop peu précises [14].

En définitive, la régression LASSO semble avoir très peu pénalisé les coefficients, étant donné que 82.2% des paramètres liés à la modélisation du prix de l'électricité sont non nuls, 92.8% pour la charge et 76.6% pour la production d'énergie renouvelable. Cela nous pousse à nous demander quels sont les avantages d'une régression LASSO qui semble peu pénaliser le grand nombre de variables prises en compte dans le modèle. Il aurait notamment été intéressant de comparer les résultats trouvés par la régression LASSO à ceux obtenus par MCO ou régression Ridge afin de justifier le choix de la méthode de régression pris par les auteurs de l'article.

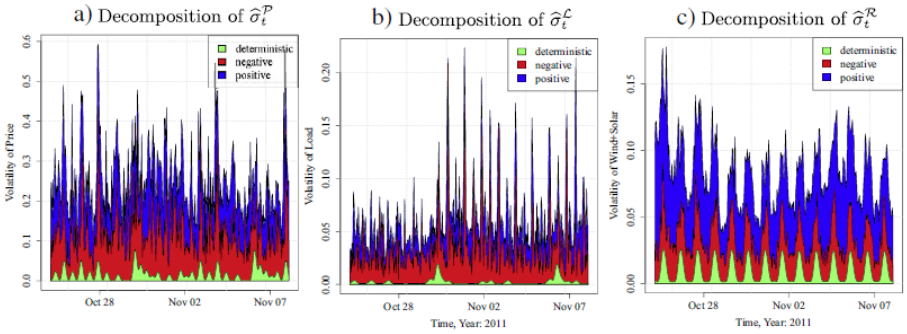
1.3.3 Un effet de levier correctement modélisé par le modèle auto-régressif TARCH

Comme introduit lors de la présentation du modèle, la volatilité σ_t a été modélisée par un modèle TARCH prenant en compte ce phénomène d'effet de levier [15]. Pour s'en rendre compte, la volatilité a été décomposée en 3 composantes,

13. la t-value d'un coefficient $\theta \in \mathbb{R}$, lorsqu'on teste la nullité de ce coefficient, se définit de la manière suivante : $tvalue = \frac{\hat{\theta}}{s\hat{d}(\theta)}$
14. Cf. modèles de prévision ARPEGE utilisé par météo France
15. Cf. partie précédente

l'une déterministe, une seconde traduisant la composante positive de la volatilité et une dernière faisant état de la composante négative de la volatilité. Puis les valeurs des composantes ont été représentées pour les données temporelles s'étalant du 28/10/2011 au 07/11/2011.

Graphique 1.5 – Décomposition de la volatilité temporelle en 3 composantes pour le prix, la charge et les énergies renouvelables



On observe tout d'abord sur le graphique 1.4 que la composante déterministe possède une forte saisonnalité surtout pour ce qui est des énergies renouvelables (soleil + vent). Nous observons ensuite que la seule série temporelle ayant un effet de levier important est celle représentant la charge puisque la composante négative de la volatilité semble atteindre des amplitudes plus importantes que pour la composante positive.

Tableau 1.1 – Test de l'effet de levier $H_0 : A_{i,k=max} = 0$

	$A_{i,k=max}$	$\sigma(A_{i,k=max})$	t-value	p-value
Prix	-0.0561	0.0540	-1.0394	0.2986
Charge	-0.1780	0.0417	-4.2666	0.0000
Soleil + Vent	0.8427	0.0635	13.2689	0.0000

Le test statistique d'hypothèse nulle $\mathbb{H}_0 : A_{i,k} = \sum_{j=1}^{k} \alpha_j^{+,i} - \alpha_j^{-,i} = 0$ [16] avec k le nombre de lags permet de statistiquement confirmer la présence d'un effet de levier ou non. Le tableau 1.1 permet de rejeter fortement au niveau 1% l'absence d'effet de levier pour le soleil + vent et la charge mais pas pour le prix. De plus, l'effet de levier pour le soleil et le vent est inverse à la charge d'électricité. La volatilité est d'autant plus importante que la quantité d'électricité produite d'origine renouvelable augmente. Ce phénomène est notamment illustré par le signe de $A_{i,k}$ positif pour le vent + soleil et négatif pour la charge.

1.3.4 Le prix de l'électricité diminue lorsque la quantité d'électricité produite par l'éolien et le solaire augmente

L'effet marginal d'une augmentation de charge d'1GWh semble avoir un effet négatif sur le prix de -0.108($\pm$0.499) $\frac{EUR}{MWh}$ bien que l'intervalle de confiance de cet impact soit large (la probabilité pour que l'impact d'une plus grande charge impacte positivement le prix est non nulle). En outre, l'effet marginal d'une augmentation de volume d'électricité produit par l'éolien et le solaire d'1GWh semble diminuer le prix de -2.031 ($\pm$ 0.375) $\frac{EUR}{MWh}$.
L'impact de la production d'énergie solaire et éolien semble donc réduire de manière importante le prix à court terme. L'impact de la charge semble nettement plus faible même s'il permet également de réduire le prix de l'électricité.

16. $\mathbb{H}_0$:absence d'effet de levier

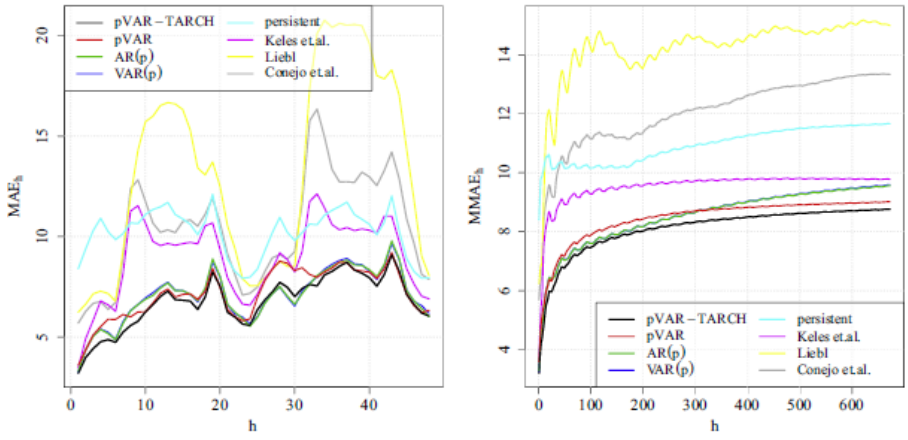
1.3.5 Prévision du prix de l'électricité

Etant donné le modèle construit dans cet article, il est aisé de prédire les valeurs futures du prix à partir des paramètres estimés du modèle et des valeurs passées. La prévision de Y_{n+h} à partir de Y_n se fait de manière imbriquée telle que $\hat{Y}_{n+h} = g(\hat{Y}_{n+h-1}, \hat{Y}_{n+h-2}, ...)$.
Afin de pouvoir évaluer la qualité des prédictions, il est également bon d'avoir les véritables valeurs de la série étudiée. Ainsi, à partir de la moitié de la série temporelle des prix de l'électricité (soit 18481 valeurs, donc 2 ans et une heure), les auteurs ont cherché à estimer les 4 semaines futures. En outre, afin de se rendre compte de la performance du modèle introduit dans cet article, il est utile de comparer les taux d'erreur du modèle présenté face aux autres modèles principaux présents dans la littérature. Deux termes d'erreur sont alors étudiés :

- l'erreur moyenne de la prévision de Y_h pour un h donné sur l'ensemble des individus. Celle-ci se définit par MAE_h. Par exemple si h=24, alors MAE_{24} correspond à l'erreur du modèle pour prédire la valeur de $Yn + 24$ à partir de Yn et des valeurs passées de Yn.
- l'erreur moyenne sur l'ensemble des prévisions jusqu'en h définie par $MMAE_h$.

$$MAE_h = \frac{1}{N}\sum_{k=1}^{N} |Y_{h,k} - \hat{Y}_{h,k}| \quad et \quad MMAE_h = \frac{1}{Nh}\sum_{j=1}^{h}\sum_{k=1}^{N} |Y_{j,k} - \hat{Y}_{j,k}|$$

Graphique 1.6 – Erreurs de prédiction en EUR/MWh par modèle pour les h heures futures



Le graphique 1.6 donne alors l'évolution des erreurs suivant l'heure future h prédite. On observe alors que le modèle pVaR-TARCH de cet article possède un taux d'erreur de prévision minimal pour presque toutes les heures futures de prédiction et une erreur globale des 672 heures futures nettement plus faible que celle des autres modèles. On observe également que le modèle simple d'autorégression AR(p) (avec p de l'ordre de 800) fonctionne presque aussi bien que le modèle pVaR-TARCH pour les 300 premières heures futures. Le modèle autorégressif vectoriel ayant les composantes prix et charge noté pVAR (avec p = 400) semble également très bien fonctionner, bien que légèrement moins bien que le modèle pVaR-TARCH.

Le modèle présenté par Florian Ziel, Rick Steinert et Sven Husmann semble donc être très performant aussi bien pour modéliser la saisonnalité que pour prendre en compte l'effet de levier et l'hétéroscédasticité des séries temporelles du prix, de la charge et de la production d'électricité par voie solaire et éolien. Ce modèle a également pu mettre en évidence l'impact négatif sur le prix de l'électricité de la production d'électricité par l'éolien et le solaire. Il semble cependant sous estimer la dépendance temporelle de la charge d'électricité et avoir des résultats très proches, bien que meilleur, d'un modèle autorégressif AR simple.

Chapitre 2

Confrontation avec l'article de Keles

Dogan Keles, Massimo Genoese, Dominik Möst et Wolf Fichtner ont ont eux aussi proposé dans un article des méthodes de prédictions des prix de l'électricité. Leur article est par ailleurs cité par Florian Ziel à titre de comparaison lorsque ce dernier fait le point sur les approches déjà suivies pour modéliser les prix de l'électricité. La procédure suivie par D. Keles est de fait la suivante :

- Identifier les caractéristiques des prix de l'électricité, à savoir la tendance, les cycles annuels, hebdomadaires et journaliers, la saisonnalité, les prix négatifs, les pics et sauts de prix
- Estimer différents modèles de séries temporelles (modèles ARMA, ARIMA, GARCH et de Mean Reversion)
- Comparer la précision de chacun des modèles prédictifs, en utilisant les critères d'erreur MAPE et RMSE
- Etudier l'impact des prix négatifs ainsi que de l'approche de changement de régime pour pour simuler les pics que connaissent fréquemment les prix de l'électricité, afin d'améliorer la prédiction

2.1 Des points de divergence fondamentaux

L'article de D. Keles présente des différences essentielles avec celui de F. Ziel. **La première concerne les objectifs**. Si l'objectif principal, celui d'obtenir des prédictions plutôt de court-terme satisfaisantes, est commun, l'article de F. Ziel se donne pour objectif explicite de démontrer empiriquement une théorie économique : l'influence des énergies renouvelables sur les prix. **La seconde concerne la modélisation**. D. Keles utilise une approche classique des séries temporelles consistant à transformer la série initiale jusqu'à ce qu'elle vérifie les hypothèses nécessaires des modèles classiques (ARMA, GARCH etc). Il s'inspire également explicitement des méthodes utilisées en finance de marché et en propose une application différente (MR, *Regime-Switching*). Au contraire, F. Ziel propose une approche mixte combinant des méthodes de séries temporelles classiques (VAR, TARCH) avec des méthodes plutôt spécifiques à l'apprentissage automatique (*machine learning*) comme l'introduction de fonctions de bases et l'utilisation du LASSO. **Notons néanmoins que ces deux articles se posent les mêmes contraintes, fortes** : il ont la spécificité de prendre en compte les jours fériés ainsi que les prix négatifs.

2.2 Une approche sensiblement différente, axée sur l'estimation de séries temporelles

2.2.1 Approche générale

Celle-ci se fait en plusieurs temps. Comme dans un modèle d'estimation de séries temporelles classique, on extrait de la série temporelle préalablement modifiée via une transformation logarithmique la tendance ainsi que les cycles journaliers, hebdomadaires et annuels.

La transformation logarithmique n'est pas simple puisque depuis 2008, les prix de l'électricité peuvent être négatifs en Allemagne. D. Keles propose de modifier ces valeurs en prix positifs, selon une approche axée sur la loi de probabilité empirique des prix négatifs. Les modèles ARMA, ARIMA, GARCH et MR peuvent ensuite être estimés sur la partie de la série restante. En ajoutant les différentes caractéristiques qui avaient été enlevées telles que la tendance et la saisonnalité, en repassant à l'exponentielle, puis en réintroduisant les prix négatifs, des prédictions complètes peuvent être obtenues.

2.2.2 Présentation et justification des modèles testés

Les modèles présentés et utilisés par D. Keles (MR, ARMA-ARIMA et GARCH) dans son article sont explicités en Annexe 4.1.

2.2.3 Gestion des pics et sauts d'électricité

Il convient également de souligner l'effort de D. Keles pour modéliser les pics et sauts d'électricité selon une approche différente de F. Ziel. Le modèle de changement de régime (*Regime-Switching*) a été utilisé afin de simuler les transitions entre les deux types de régimes définis par D. Keles, à savoir le régime de base (*base regime*), et le régime de saut (*jump regime*) des prix de l'électricité. Le premier est une série lissée et qui correspond au niveau des prix en temps normal. Le second correspond à un niveau inhabituel et temporaire des prix, faisant par exemple suite à un choc. Dans son article, D. Keles cherche à simuler ces deux états.

2.3 Plus simpliste, l'approche de D. Keles obtient des résultats convaincants

Les résultats indiquent que l'approche en deux temps (opérations préliminaires sur les séries brutes, puis modélisation de la composante stochastique) de D. Keles est intéressante. Un des premiers éléments à retenir est que le modèle ARIMA n'est pas nécessaire. Les opérations successives de différenciation sur les log-séries des prix pour enlever la tendance ainsi que la saisonnalité n'a pas un apport prépondérant. L'approche de D. Keles d'un changement de régime pour la modélisations des pics et sauts de prix a, elle, induit un apport conséquent. Les processus stochastiques ne sont de fait pas à même de simuler parfaitement ces derniers, pas même le modèle GARCH qui cherche à résoudre le problème d'hétéroscédasticité. Le modèle introduit axé sur ce changement de régime parvient mieux à générer des sauts et pics, dont la structure correspond assez bien avec ceux issus de la série historique. Pour finir, les processus MR (*Mean Reversion*) et ARMA(1,5) donnent des résultats convaincants dans l'évaluation des cycles journaliers et hebdomadaires, ainsi que de la volatilité stochastique. En revanche, le modèle GARCH génère des prix plus élevés que ceux issus de la série historique. Les indicateurs d'erreur RMSE et MAPE sont d'ailleurs beaucoup plus faibles pour les modèles ARMA et MR que GARCH, soulignant la meilleure qualité des deux premiers par rapport au troisième. Il est enfin à noter que l'approche cherchant à évaluer les prix négatifs a un apport significatif sur la diminution de l'erreur.

Conclusion

Avec une modélisation plus classique, et seulement univariée, D. Keles semble donc arriver à des résultats relativement satisfaisants même s'ils semblent l'être moins que ceux du modèle proposé par F. Ziel comme l'indique la comparaison effectuée par ce dernier. L'article a néanmoins le mérite de proposer les fondements d'une méthode réutilisable dans tous types de modèles pour traiter les prix négatifs et de proposer une alternative aux modèles de type ARCH / GARCH pour traiter l'hétéroscédasticité. Deux éléments doivent cependant être signalés. D'une part les données des prix ne sont pas les mêmes pour les deux articles puisqu'elles sont toutes deux issues d'une période distincte (2008 - 2010 pour D. Keles, 2010 - 2014 pour F.Ziel). D'autre part les indicateurs d'erreur utilisés par chacun des article ne sont pas similaires (MAPE, RMSE pour D. Keles, MAE et MMAE pour F. Ziel). Il reste donc possible que le modèle proposé par F. Ziel appliqué à la période étudiée par Keles et évalué selon les critères d'erreurs MAPE et RMSE soit moins performant. Les résultats doivent donc être considérés avec précaution.

Chapitre 3

Modélisation du prix spot de l'électricité en Allemagne à partir de la production d'électricité d'énergies éoliennes

Après avoir étudié les articles de F. Ziel et de D. Keles, nous proposons dans cette partie une courte mise en application afin de mieux comprendre les différences entre leurs approches et l'utilité des complexifications qu'ils introduisent. L'idée est de voir quels sont les résultats obtenus à l'aide de modèles simples, et d'en comprendre les faiblesses. Nous avons cherché à expliquer et prévoir le prix de l'électricité en utilisant des b-splines pour gérer la saisonnalité ainsi que les données de production d'électricité éolienne. Nous avons dans un premier temps réalisé un modèle AR(p) puis un VAR(p) en y intégrant des B-splines comme dans l'article de F. Ziel.

3.1 Présentation des données et statistiques descriptives

Les données utilisées dans les applications à venir sont les données horaires du prix spot d'électricité du lundi au vendredi entre le 01/01/2010 et le 11/06/2014 ainsi que les données de la production d'énergie éolienne pour ces mêmes dates et avec le même pas. Pour plus de précisions sur l'origine des données traitées ainsi que les statistiques descriptives les caractérisant, nous vous invitons à vous reporter en Annexe de ce rapport.

3.2 Le modèle AR(p)

3.2.1 Présentation du modèle

Ce modèle s'inspire de l'approche de l'article par son utilisation des B-splines et du LASSO. Il est néanmoins largement simplifié. Les deux différences majeures d'avec l'article sont :

- La non prise en compte des données de production et de consommation d'électricité : il s'agit d'une modélisation univariée (AR) et non multivariée (VAR).
- La non prise en compte des jours fériés et la non distinction des différents jours de la semaine (le lundi et le vendredi sont considérés comme similaires aux autres jours ouvrés)

Le modèle utilisé est alors le suivant, avec Y_t le prix spot à la date t :

$$Y_t = \mu(t) + \sum_{k=1}^{1000} \phi_k Y_{t-k} + \epsilon_t$$

où la tendance prend en compte la périodicité annuelle et la périodicité journalière à l'aide de B-splines cubiques :

$$\mu(t) = \mu_0 + \mu_{lin} + \sum_{l=1:5} \mu_l \tilde{B}_l(t) + \sum_{l=1:11} \mu_j \tilde{B}_j(t)$$

Les 5 premières fonctions de base (l = 1,..., 5) $\tilde{B}_l$ sont des sommes de b-splines cubiques uniformes dont les noeuds sont espacés de 4 heures et centrées sur les mêmes heures que dans l'article (minuit, 4 heures du matin, etc.). Il s'agit donc dans notre cas, où les jours fériés ne sont pas distingués, de 5 fonctions parfaitement périodiques d'une période de 24 heures. La 6ème fonction de base est supprimée, comme dans l'article, pour éviter des problèmes de multicolinéarité parfaite des variables [1]. De la même façon, on utilise 11 fonctions de base pour modéliser la périodicité annuelle, exactement comme dans l'article [2]. La période utilisée ici n'est pas la même que celle de l'article (elle est de 6264 heures contre 8765.76 dans l'article), car nous ne disposons dans la base de données que de 5 jours de la semaine [3]. Dans un premiers temps les lags n'ont pas été sélectionnés, contrairement à ce qui a été fait dans l'article : le LASSO pénalisera les 1000 lags proposés, pour ne garder que ceux utiles au modèle [4].

3.2.2 Choix du paramètre de pénalisation λ dans la régression LASSO

Nous effectuons une validation croisée de type K-Fold proche de celle probablement effectuée par l'article [5] : les prédictions sont bien des prédictions emboîtées et l'erreur comparée entre les différents λ proposés est l'erreur MMAE. Si l'article semblait utiliser 506 échantillons différents pour comparer les paramètres λ, nous n'en utilisons que 5, pour des raisons de lenteur [6]. Ces 5 partitions sont tirées au hasard sur différents échantillons et estiment les coefficients à partir de 110 semaines + 1 heure. L'erreur de prédiction a été testée pour l'équivalent de 4 semaines de prédictions (480 heures dans notre cas). Le paramètres λ obtenu est de 0.08.

3.2.3 Résultats de prédiction des 480 heures futures

Les résultats de la régression Lasso sur les 1018 coefficients à estimer sont alors les suivants :

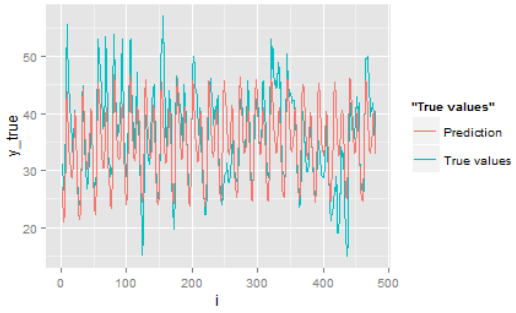
- Seuls 59 sont non nuls. Il semble donc qu'un tri préalable aurait pu permettre d'améliorer le modèle.
- Aucune fonction de base permettant de modéliser la périodicité annuelle n'est retenue : la tendance estimée est donc identique en hiver et en été.
- Le coefficient μ_{lin} est nul. Notons que lorsqu'on utilise un paramètre de pénalisation λ légèrement plus faible (0.05 au lieu de 0.08), ce coefficient devient non nul et il est bien négatif comme dans l'article, laissant présager d'une baisse des prix sur le long-terme due aux énergies renouvelables.

3.2.4 Comparaison graphique des prédictions pour les 4 dernières semaines disponible de l'année 2014 (fin mai-début juin)

On remarque tout d'abord que le modèle a du mal à capter les très grands pics que connaissent parfois les prix. Ceci est probablement dû en grande partie à l'hypothèse implicite d'homoscédasticité des résidus du modèle LASSO, relâchée dans l'article à l'aide d'un modèle TARCH. L'article de Keles propose également une autre méthode pour traiter ce phénomène (*Regime switching*).
On remarque sur le graphique 3.1, qui détaille les prédiction du premier jour et de la première semaine, que les principaux motifs sont néanmoins bien reproduits, même à l'aide de ce modèle très simplifié.

1. En effet, comme évoqué plus haut, la somme des 6 fonctions de base est constamment égale à 1.
2. La 12ème étant supprimée pour la même raison que la périodicité quotidienne
3. Cette approximation ne tiens pas compte des années bissextiles comme les auteurs de l'article le font, 6264 étant le nombre d'heures présentes dans notre base de données du 1er janvier au 31 décembre.
4. Cela nous permet d'une part un gain de temps : comment les choisir ? mais également d'analyser les lags choisis automatiquement et de les comparer à ceux utilisés dans l'article.
5. Voir partie 1
6. Il aurait été possible d'améliorer le code pour que son exécution soit plus rapide.

Graphique 3.1 – Prédiction des 4 dernières semaines

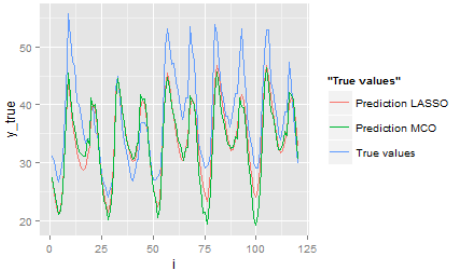


3.2.5 Comparaison avec une estimation par MCO

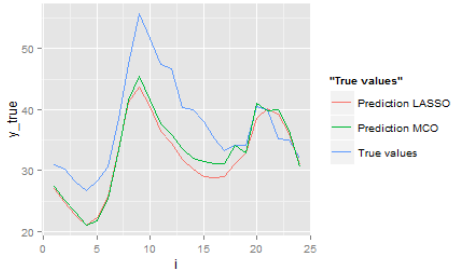
Les coefficients sont estimés par MCO sur les mêmes données de façon à mettre en évidence l'utilité du LASSO, non démontrée par l'article à l'aide d'une expérimentation. On aurait pu également, si le temps l'avait permis, comparer à une estimation par la méthode Ridge. On remarque que sur la première journée, l'estimation par MCO est meilleure, mais les prédictions se dégradent rapidement, comme le montre notamment le calcul de l'erreur $\frac{1}{h}\sum_{j=1}^{h}|Y_j - \hat{Y}_j|$ (erreur MMAE avec N = 1). Notons que comparer les deux méthodes sur les prédictions d'une seule période ne permet pas d'affirmer de façon entièrement fiable que le modèle LASSO est meilleur, il aurait fallu comparer sur N échantillons différents et calculer l'erreur MMAE.

Tableau 3.1 – Comparaison des erreurs MMAE (cas où N =1)

	1er jour	1ère semaine	4 semaines
LASSO	5.808371	**4.669686**	**4.673473**
MCO	**5.05077**	5.250896	5.276853



(a) Comparaison LASSO/MCO sur la 1ère semaine



(b) Comparaison LASSO/MCO sur le 1er jour

Graphique 3.2 – Comparaison LASSO/MCO

3.3 Le modèle VAR(p)

3.3.1 Présentation du modèle

Nous avons alors cherché à affiner le modèle AR(p) en incluant les données liées à la production d'énergie éolienne. Le modèle de prévision étant emboîté, il a été nécessaire de réaliser à la fois un modèle de prévision du prix mais

aussi un modèle de prévision du vent. Le modèle VAR(p) a alors été réalisé sur le même ensemble de données que le modèle AR(p) avec cependant une dépendance avec le passé plus ciblée. Les modèles utilisés sont alors les suivants :

$$\begin{cases} Y_t^{Prix} = \mu^{Prix}(t) + \sum_{k=1}^{I} \phi_k Y_{t-k}^{Prix} + \sum_{k=1}^{49} \psi_k Y_{t-k}^{Vent} + \sum_{k=1}^{I} \sum_{l=1:5} \alpha_{k,l} Y_{t-k}^{Prix} \tilde{B}_l(t) + \sum_{j=1:11} \beta_{k,j} Y_{t-k}^{Prix} \tilde{B}_j(t) + \epsilon_t \\ Y_t^{Vent} = \mu^{Vent}(t) + \sum_{k=1}^{J} \phi_k Y_{t-k}^{Vent} + \sum_{k=1}^{J} \sum_{l=1:5} \alpha_{k,l} Y_{t-k}^{Vent} \tilde{B}_l(t) + \sum_{j=1:5} \beta_{k,j} Y_{t-k}^{Vent} \tilde{B}_j(t) + u_t \end{cases}$$

$$\begin{cases} \mu^{Prix}(t) = \mu_0 + \mu_{lin} + \sum_{l=1:5} \mu_l \tilde{B}_l(t) + \sum_{l=1:11} \mu_j \tilde{B}_j(t) \\ \mu^{Vent}(t) = \mu_0 + \mu_{lin} + \sum_{l=1:5} \mu_l \tilde{B}_l(t) + \sum_{l=1:5} \mu_j \tilde{B}_j(t) \end{cases}$$

avec $I = \{1:40, 48, 72, 96, 120, 240, 360, 480, 600, 720, 840\}$ et $J = \{1:48, 49, 72, 73, 96, 97, 120, 121, 240, 360, 480, 600, 720, 840\}$. Les lags ont été présélectionnés à partir des résultats du modèle utilisé par l'article. De plus, les fonctions b-splines sont les mêmes que celles introduites dans le modèle AR(p), à l'exception près que seules 5 b-splines sont utilisées pour la périodicité annuelle du vent, comme l'article le préconise. Les principales différences d'avec le modèle AR(p) sont celles qui suivent.

- Le prix de l'électricité dépend de la production d'électricité d'énergie éolienne ce qui fait que la prédiction du prix de l'électricité dépendra de la prédiction de la production du vent. Les prédictions doivent donc se faire de manière simultanée pour les deux variables ;
- Une dépendance périodique quotidienne et annuelle des prix passés est introduite grâce au produit des b-splines et des lags du prix. Le même procédé est également présent pour prédire les données du vent.

3.3.2 Choix des paramètres de pénalisation λ dans la régression LASSO

Afin d'obtenir le meilleur modèle prédictif possible, nous avons réalisé une pseudo validation croisée [7] pour ce second modèle. En effet, deux paramètres de pénalisation doivent être trouvés simultanément, le premier pour le modèle de prédiction du prix, et le second pour le modèle de prédiction du vent. Sachant que seule la prédiction du prix nous intéresse, nous avons comparé les erreurs MMAE (avec N=1) pour la prédiction du prix avec plusieurs paramètres λ_{prix}, paramètre de pénalisation de la régression LASSO du prix. Pour cela, nous avons fixé arbitrairement celui de la régression de la production d'énergie d'origine éolienne. Puis une fois que le λ_{prix} optimal a été trouvé, on a cherché le λ_{vent} de la régression de la production d'électricité d'origine éolienne qui minimise l'erreur MMAE du prix sur les 480 heures futures (avec N=1). Nous avons alors trouvé $\lambda_{prix} = 0.25$ et $\lambda_{vent} = 1.02$.

3.3.3 Résultats obtenus

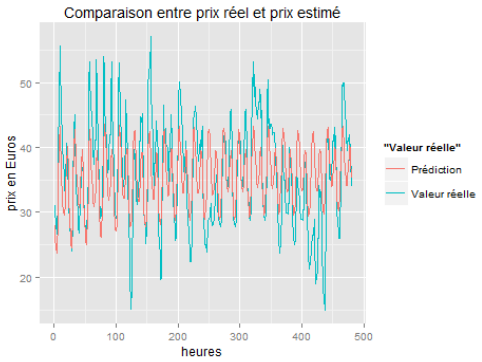
Les résultats obtenus sont alors très proches de ceux de l'AR(p). On observe que :

- sur les 1095 coefficients permettant de modéliser le prix, seuls 18 sont non nuls ;
- La production d'électricité d'énergie éolienne semble peu influer sur l'évolution des prix dans ce modèle puisque seuls les lags 1, 19 et 23 du vent sont non nuls ;
- sur les 672 coefficients permettant de modéliser l'évolution de la production d'énergie éolienne, seuls 84 sont non nuls.

De plus, on remarque sur le graphique 3.3 qui représente la prédiction du prix spot d'électricité et de la production d'énergie d'origine éolienne, sur les 480 heures futures, que les grandes fluctuations sont toujours mal modélisées dû au fait qu'on ne modélise pas le phénomène d'hétéroscédasticité. De plus on observe que la production d'électricité

7. Nous n'avons pas réalisé de validation rigoureuse à la fois par manque de temps, mais aussi parce que celle-ci aurait demandé un temps d'exécution très long puisqu'il faut trouver deux paramètres de pénalisation optimaux de manière simultanée.

d'origine éolienne semble relativement mal modélisée par le modèle, erreur qui se propage dans la prévision du prix de l'électricité.



(a) Prédictions des 480 heures futures pour le prix spot

(b) Prédictions des 480 heures futures pour la production d'électricité éolienne

Graphique 3.3 – Prédictions par le modèle VAR(p)

3.3.4 Comparaison des modèles AR(p) et VAR(p) sur les 4 dernières semaines

On remarque que le modèle VAR(p) semble être meilleur au cours des premières 24 heures et 120 heures (semaine de 5 jours) que le modèle AR(p), alors que sur 480 heures le modèle AR(p) semble l'emporter de peu. Les modèles AR(p) et VAR(p) semblent ainsi avoir globalement les mêmes caractéristiques de prédiction alors que le modèle multivarié semble plus complexe et prendre plus d'information en compte. Ce résultat confirme bien les conclusions de l'article qui soulignaient le fait que le modèle AR(p) fonctionnait presque aussi bien que le modèle mis en place par les auteurs.

Tableau 3.2 – Comparaison des erreurs MMAE (cas où N =1)

	1er jour	1ère semaine	4 semaines
AR(p)	5.808371	4.669686	**4.673473**
VAR(p)	**5.065224**	**4.528606**	4.917084

Conclusion

Le rapport a présenté et appliqué des modèles de prédiction pour les prix horaires de l'électricité en Allemagne. Une double approche a de fait été mise en place *via* les modèles AR(p) et VA(p), en s'inspirant des spécificités de l'article théorique de Florian Ziel telles que les B-splines et la régression de type LASSO, et l'intégration de la production d'électricité d'énergie éolienne. Les résultats de ces modèles simplifiés ne sont bien sûr pas parfaits.
Les points suivants peuvent en définitive être soulignés :

- F. Ziel a monté avec un succès un modèle prédictif à la fois très technique et efficace, en intégrant les données de consommation d'électricité et de production d'électricité grâce à l'énergie solaire et éolienne
- L'approche n'est pas pour autant la seule efficace puisque celle proposée par D. Keles montre des résultats concluants dans un contexte différent
- La régression LASSO semble obtenir de meilleurs résultats qu'une régression MCO classique
- Si les modèle AR(p) et VAR(p) donnent des résultats globalement similaires, le premier semble plus convaincant que le second, notamment parce qu'il est moins compliqué à mettre en place.

Chapitre 4

Annexe

Confrontation avec l'article de Keles

Présentation et justification des modèles testés

Modèle MR

Le modèle MR (Mean Reversion) est surtout utilisé en finance. Il s'appuie sur l'hypothèse que le prix d'une action est proche de son prix moyen sur le long-terme . Le processus MR peut être formulé par l'équation différentielle suivante :

$$dX_t = \kappa(\mu - X_t)dt + dW_t \tag{4.1}$$

En utilisant le Lemme d'Ito, une formule de X_{t+1} en fonction de X_t peut être obtenue :

$$X_{t+1} = X_t.e^{-\kappa\delta} + \mu(1 - e^{-\kappa\delta}) + \sigma\sqrt{\frac{1 - e^{-2\kappa\delta}}{2\kappa}}.\epsilon_t \tag{4.2}$$

En utilisant le maximum de vraisemblance, les différents paramètres inconnus peuvent être estimés.

Modèle ARMA-ARIMA

Les processus ARMA et ARIMA constituent une autre possibilité d'estimation des résidus stochastiques. Sa modélisation est la suivante, le terme d'erreur devant être un bruit blanc :

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} + \epsilon_t \tag{4.3}$$

Une des hypothèses sur laquelle s'appuie le modèle ARMA est la stationnarité faible de la série temporelle. Celle-ci n'est généralement pas vérifiée, il faut donc transformer la série pour arriver à un processus la respectant. Une des stratégies communément employées est la différenciation successive de la série jusqu'à ce que la tendance disparaisse. Ici, la tendance a déjà été supprimée au préalable, l'opération de différenciation n'est donc *a priori* pas nécessaire ici. Néanmoins, la log-série désaisonnalisée et dont la tendance a été enlevée peut toujours comprendre des composantes déterministes. C'est pourquoi le modèle ARIMA est lui aussi appliqué sur les log-séries brutes, de façon à le comparer avec le modèle ARMA.

Modèle GARCH

L'approche GARCH est justifiée par le potentiel caractère hétéroscédastique des résidus de la série temporelle. Ce terme désigne le fait que la variance des erreurs du processus ne soit pas la même pour toutes les observations, ce

qui contredit l'hypothèse de stationnarité. Dans le cadre des prix de l'électricité, la série est très volatile, alternant les pics plus ou moins élevés. La conséquence est que les prix peuvent devenir rapidement imprévisibles lorsque leur volatilité est forte : les marchés sont plus nerveux, les pics des prix sont plus imprévisibles et fréquents. Les modèles GARCH permettent de modéliser un tel comportement, selon la formule suivante :

$$\sigma_t^2 = \omega + \sum_{z=1}^{p} \alpha_z \omega_{t-z}^2 + \sum_{z=1}^{q} \beta_z \epsilon_{t-z}^2 \tag{4.4}$$

Présentation des données utilisées en partie 3

Données des prix spot

La base de données comprenant les prix spot de l'électricité en Allemagne a été récupérée auprès d'une connaissance (celle-ci étant payante sur le site de EEX). Elle contient les prix horaires sur une période similaire à celle prise par l'article : entre 1er janvier 2010 et le 12 juin 2014. Malheureusement, elle ne contient pas les week-ends. Nous disposons ainsi d'une série temporelle de 27 816 données horaires soit 1159 jours.

Prix négatifs

Comme le mettait fortement en évidement l'article de D. Keles, l'existence de prix négatifs empêche, entre autres, d'utiliser une transformation logarithmique des prix, permettant souvent de réduire l'hétéroscédasticité et de se ramener à une série stationnaire. D'autres méthodes sont donc nécessaires.

Graphique 4.1 – Evolution du prix d'électricité d'1MWh en centime d'euro entre le 01/06/2013 et le 12/06/2014 pour 1h, 2h et 3h du matin

Comme nous pouvons le constater sur le graphique ci-dessus, le prix spot prend des valeurs négatives à instants très particuliers. Le 28 octobre 2013, un fort orage s'est déclaré en Allemagne [1] ; le 24 décembre 2013, il y a généralement une réduction forte de la consommation d'électricité étant donné que c'est un jour fériés commun pour tous ; le 17 mars 2013, un fort orage s'est déclaré en Autriche. Les prix négatifs sont également observables durant les jours fériés communs à l'ensemble de la population allemande : par exemple entre le 25 décembre et le 1er janvier.

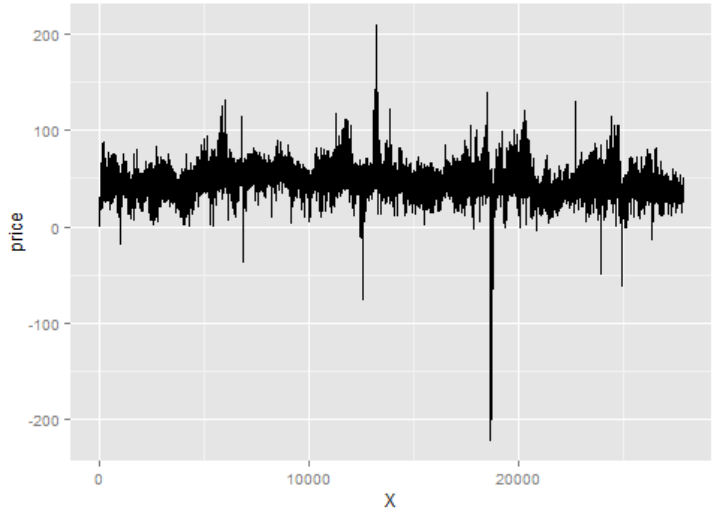
Statistiques descriptives

N	moyenne	Ecart-type	Minimum	Maximum	skewness	kurtosis
27816	41.014	20.406	-221.99	183.49	-0.904	6.907

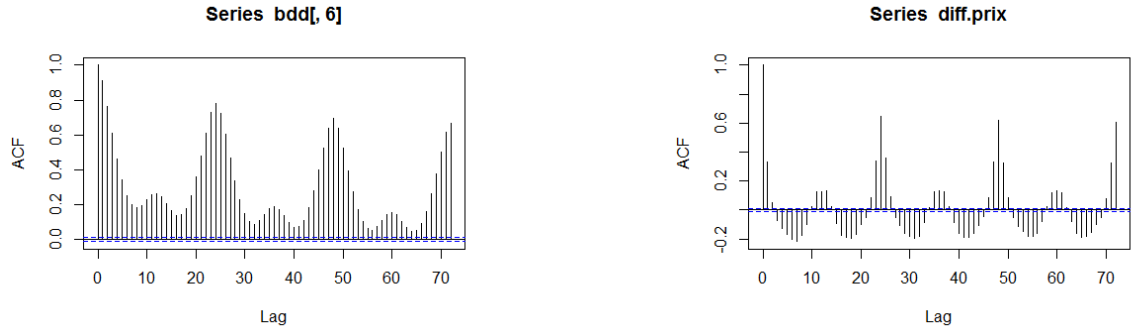
Tableau 4.1 – Statistiques descriptives du prix spot en euro

1. http ://www.metoffice.gov.uk/about-us/who/how/case-studies/st-judes-day-storm-oct-2013

Comme l'indique la table 3.1, les données temporelles du prix présentent un indice de kurtosis relativement élevé et une asymétrie de distribution proche de -0.9 (skewness) ce qui laisse penser que la distribution du prix aura des queues lourdes.



La figure ci-dessus ne permet pas de déceler une quelconque tendance croissante au cours du temps. La saisonnalité constitue donc le principal problème à gérer afin d'obtenir une série stationnaire.



(a) Série avant différenciation

(b) Série après différenciation

Graphique 4.2 – Autocorrélations empiriques des prix spot de l'électricité en Allemagne

Comme le montre la figure 3.2, la série temporelle du prix de l'électricité possède une saisonnalité, qui reste difficile à déterminer. Cependant, lorsque le série temporelle est différenciée, on observe clairement l'existence d'une saisonnalité dans les données du prix. Dans les deux cas, la série possède des autocorrélations non nulles pour les lags autres que zéro ce qui indique que ces deux séries ne sont pas stationnaires.

Les données de production d'électricité d'origine éolienne

Les données sur la production d'énergies renouvelables en Allemagne telles que l'énergie solaire ou éolienne sont accessibles sur l'ensemble des sites internet des sociétés appartenant à TSOs, entreprise allemande de production d'énergie renouvelable. Sur les 4 producteurs d'énergie éolienne en Allemagne (TenneT TSO, Amprion, 50Hertz et TransnetBW), nous avons récupéré les bases de données mensuelles auprès des 3 plus importants couvrant près de 98% de la production.

Préparation de la base de donnée

Les données étant récupérées pour un pas d'un quart d'heure, nous avons dans un premier temps fait la moyenne de la production d'électricité pour TenneT, Amprion et 50Hertz tous les 4 quarts d'heure afin d'obtenir des données horaires.
Certaines données étaient manquantes. Les techniques usuelles de gestion de données manquantes n'ont pas pu être mises en pratique : les données manquantes étaient en effet en très faible nombre. Nous avons donc pris la décision de remplacer ces valeurs manquantes par une moyenne sur les deux jours précédents et les deux jours suivants.
Certaines données étaient également manquantes à cause des changements d'heure s'opérant à la fin des mois de mars (passage à l'heure d'été) et d'octobre (passage à l'heure d'hiver). Nous avons alors complété les données manquantes en moyennant sur les données des autres années pour le même jour et le même quart d'heure. En effet, cette méthode fonctionne bien pour les années qui ont été prises en comptes car les journée de changement d'heure n'étaient que très rarement les mêmes d'une année sur l'autre.

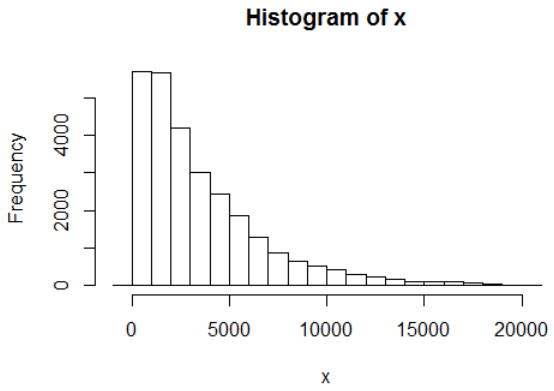
Statistiques descriptives

N	moyenne(MWh)	Ecart-type	Minimum	Maximum	skewness	kurtosis
27816	3596	3292	-310.25	20 485	1.677	6.127

Tableau 4.2 – Statistiques descriptives du prix spot en euro

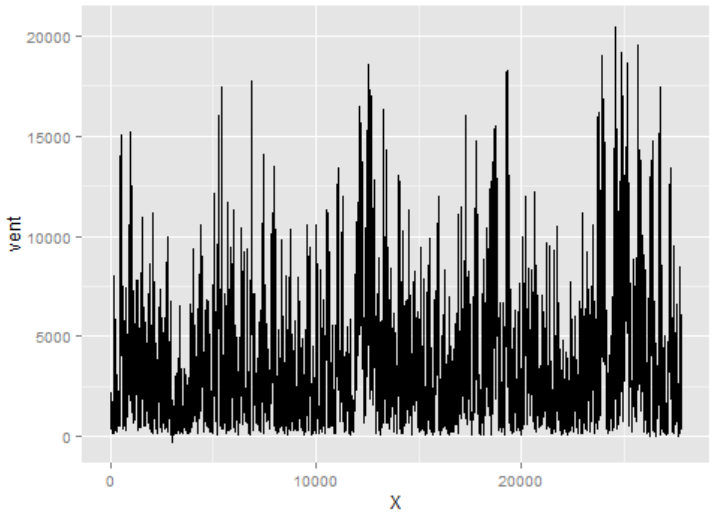
Comme l'indique la table 3.2, les données temporelles du vent présentent un indice de kurtosis relativement élevé et une asymétrie de distribution de 1.68 (skewness) ce qui laisse penser que la distribution de la production du vent aura des queues lourdes et une asymétrie de distribution prononcée. Cette intuition est notamment confirmée par l'observation de son histogramme.

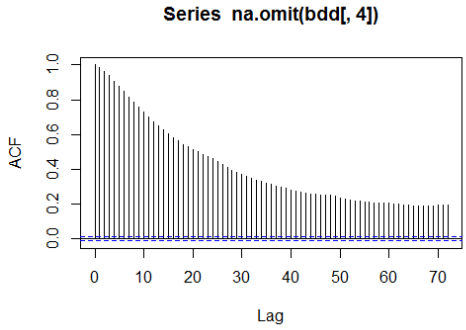
Graphique 4.3 – Production d'électricité éolienne de 2010 à 2014 en MWh



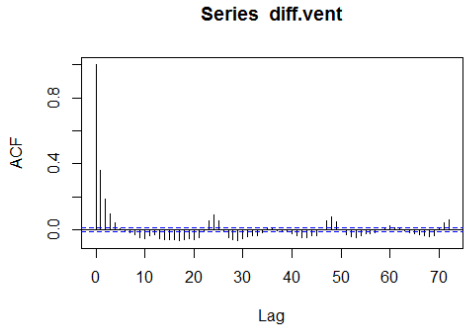
De même que pour les prix, la figure ci-dessous ne permet pas de déceler une quelconque tendance croissante au cours du temps. Cependant une certaine saisonnalité est dors et déjà observable. En effet, on observe 5 pics annuels de production d'électricité d'origine éolienne correspondant aux périodes où le vent est en général le plus violent, c'est-à-dire au début de l'année.
Comme le montre la figure 3.5, les séries temporelles du vent et du vent différencié possèdent une saisonnalité et une tendance importante, tout la série temporelle du prix. Dans les deux cas, la série possède des autocorrélations non nulles pour les lags autres que zéro ce qui montre que ces deux séries sont non stationnaires.

Graphique 4.4 – Production électricité éolienne de 2010 à 2014 en MWh





(a) Série avant différenciation



(b) Série après différenciation

Graphique 4.5 – Autocorrélations empiriques de la production d'électricité d'origine éolienne

Résultats de l'application AR(p)

Tableau 4.3 – A METTRE EN ANNEXES

Tendance	1ère semaine	2ème semaine	3/4èmes semaines	Suite
"(Intercept)"	"lag1"	"lag143"	"lag264"	"lag494"
"bspline_hebdo1"	"lag10"	"lag144"	"lag266"	"lag527"
"bspline_hebdo3"	"lag23"	"lag145"	"lag359"	"lag590"
	"lag24"	"lag158"	"lag362"	"lag599"
	"lag25"	"lag168"	"lag398"	"lag614"
	"lag26"	"lag169"	"lag431"	"lag626"
	"lag47"	"lag170"	"lag455"	"lag638"
	"lag48"	"lag182"	"lag458"	"lag719"
	"lag49"	"lag191"	"lag479"	"lag746"
	"lag72"	"lag193"	"lag482"	"lag838"
	"lag73"	"lag199"		"lag839"
	"lag96"	"lag206"		"lag960"
	"lag97"	"lag216"		
	"lag98"	"lag217"		
	"lag120"	"lag218"		
	"lag121"	"lag239"		
		"lag240"		
		"lag241"		
		"lag242"		

Code R

Code modèle AR(p) avec B-splines et régression LASSO

```
library(questionr)
table <- read.csv("bdd_coma.csv", header = TRUE, sep = ",")
table<-rename.variable(table,"X", "obs")

library(splines)
library(pbs)
library(pomp)
library(glmnet)
library(ggplot2)

lagpad <- function(x, k) {
  if (!is.vector(x))
    stop('x must be a vector')
  if (!is.numeric(x))
    stop('x must be numeric')
  if (!is.numeric(k))
    stop('k must be numeric')
  if (1 != length(k))
    stop('k must be a single number')
  c(rep(NA, k), x)[1 : length(x)]
}

dim(table)
```

```
perio = periodic.bspline.basis(table$obs, nbasis = 6, degree = 3, period = 24, names = NULL)
plot(perio[1:48,1], type="l")
perio = perio[,1:5] #On supprime la dernière pour éviter des pbs de singularité
#vient du fait que la somme de b splines unif est cste égale à 1
perio[1:48,1]+ perio[1:48,2]+perio[1:48,3] + perio[1:48,4] + perio[1:48,5] + perio[1:48,6]


perio_annuel  = periodic.bspline.basis(table$obs, nbasis = 12, degree = 3, period = 6264, names = NULL)
plot(perio_annuel[,1], type="l")
plot(perio_annuel[,12], type="l")
perio_annuel = perio_annuel[,1:11] #Idem suppression de la dernière

bdd_lasso = cbind(table$price,perio,perio_annuel)

for (u in 1:1000){
  new_vect = lagpad(table$price, k = u)
  bdd_lasso = cbind(bdd_lasso, new_vect)
}

df_lasso = data.frame(bdd_lasso)

names <- c("price")
for (k in 1:5){
  nm <- paste("bspline_hebdo",k,sep="")
  names <- c(names,nm)
}


for (k in 1:11){
  nm <- paste("bspline_annuel",k,sep="")
  names <- c(names, nm)
}

for (u in 1:1000){
  nm <- paste("lag",u,sep="")
  names <- c(names,nm)
}


colnames(df_lasso) <- names

#unit = rep(1,27816)
#df_lasso$intercept = unit
df_lasso$trend = table$obs

#http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

############################## PREDICTION ################
```

```
#On supprime les valeurs des 480 dernières heures (4 dernières semaines) pour apprendre
x = as.matrix(df_lasso[1001:27336,2:1018])   # Pour nous 1 semaine = 120 heures, 4 semaines = 480 heures =
dim(x)
y = df_lasso$price[1001:27336]


# Avec une méthode de crros-validation "fausse" dans le cas de séries temporelles, on peut quand même l'ut
#cvfit = cv.glmnet(x, y, intercept = TRUE)
#plot(cvfit)
#cvfit$lambda.min
#coeffs_lasso = coef(cvfit, s = "lambda.min")
#coeffs_lasso = as.matrix(coeffs_lasso) #On a un coeff devant trend qui est négatif comme l'article

#Avec le paramètre lambda choisi plus bas à l'aide d'une cross-validation adaptée à des séries temporelles
fit = glmnet(x, y, intercept = TRUE, lambda = c(0.08))
print(fit) # On ne garde que 59 variables, la variable trend a un coef nul ..
coeffs_lasso = as.matrix(coef(fit))

#Nombre de coeffs non nuls
coef_nonnul = row.names(coeffs_lasso)[coeffs_lasso != 0] #59 coeffs non nuls et Trend est nul...
plot(coeffs_lasso[1:19], type = "l")
plot(coeffs_lasso[20:1017], type = "l")


y_predicted = y[1:26336]
for (h in 1:480){
  new_x = df_lasso[26336+1000+h,2:17]

  for (u in 1:1000){
    lag_p = y_predicted[26336+h-u]
    new_x = cbind(new_x, lag_p)
  }
  new_x = cbind(new_x, df_lasso[26336+1000+h,1018])
  new_x = cbind(1, new_x)
  y_predicted[26336+h] = as.matrix(new_x) %*% coeffs_lasso
}


#En utilisant des MCO normaux, qu'est ce que cela donne ?
model_mco = lm(y ~ x)
coeffs_mco = as.matrix(coef(model_mco)) #En enlevant les dernière B splines périodiques, on a permis l'est

y_predicted_mco = y[1:26336]
for (h in 1:480){
  new_x = df_lasso[26336+1000+h,2:17]

  for (u in 1:1000){
    lag_p = y_predicted_mco[26336+h-u]
    new_x = cbind(new_x, lag_p)
  }
```

```
  new_x = cbind(new_x, df_lasso[26336+1000+h,1018])
  new_x = cbind(1, new_x)
  y_predicted_mco[26336+h] = as.matrix(new_x) %*% coeffs_mco
}

#Comparaison graphique des prédictions / réalité sur les 4 semaines
y_true = df_lasso[27337:27816,1]
y_pred = y_predicted[26337:26816]
y_pred_mco = y_predicted_mco[26337:26816]
graph_comp = data.frame(y_true,y_pred)
graph_comp$i <- 1:nrow(graph_comp)
ggplot(data=graph_comp,aes(x=i)) + geom_line(aes(y=y_true,colour="True values")) + geom_line(aes(y=y_pred,
#Erreur absolue obtenue :
err_abs_4sem = sum(abs(y_true - y_pred))/480 #4.673473
err_abs_4sem_mco = sum(abs(y_true - y_pred_mco))/480 #5.276853

#Comparaison graphique des prédictions / réalité sur les 24 premières heures prédites
y_true = df_lasso[27337:27360,1]
y_pred = y_predicted[26337:26360]
y_pred_mco = y_predicted_mco[26337:26360]
graph_comp = data.frame(y_true,y_pred)
graph_comp$i <- 1:nrow(graph_comp)
ggplot(data=graph_comp,aes(x=i)) + geom_line(aes(y=y_true,colour="True values")) + geom_line(aes(y=y_pred,
#Erreur absolue obtenue :
err_abs_24h = sum(abs(y_true - y_pred))/24 #5.808371
err_abs_24h_mco = sum(abs(y_true - y_pred_mco))/24 # 5.05077

#Comparaison graphique des prédictions / réalité sur les 120 premières heures prédites (première semaine)
y_true = df_lasso[27337: 27456,1]
y_pred = y_predicted[26337: 26456]
y_pred_mco = y_predicted_mco[26337: 26456]
graph_comp = data.frame(y_true,y_pred)
graph_comp$i <- 1:nrow(graph_comp)
ggplot(data=graph_comp,aes(x=i)) + geom_line(aes(y=y_true,colour="True values")) + geom_line(aes(y=y_pred,
#Erreur absolue obtenue :
err_abs_1sem = sum(abs(y_true - y_pred))/120 #4.669686
err_abs_1sem_mco = sum(abs(y_true - y_pred_mco))/120 #5.250896



#Analyse de la tendance
#Du 12 sept au 16 sept 2011

mu_total = x[10585:10704,1:16]%*%coeffs_lasso[2:17,1] + x[10585:10704,1017]*coeffs_lasso[1018,1] + rep(1,1

plot(mu_total, type="l")

#du 5 dec au dec 2011

mu_total1 = x[12025:12144,1:16]%*%coeffs_lasso[2:17,1] + x[12025:12144,1017]*coeffs_lasso[1018,1] + rep(1,
```

```
plot(mu_total1, type="l")

#du 18 juillet au juillet 2011

mu_total2 = x[9625:9744,1:16]%*%coeffs_lasso[2:17,1] + x[9625:9744,1017]*coeffs_lasso[1018,1] + rep(1,120)

plot(mu_total2, type="l")

graph_comp = data.frame(mu_total1,mu_total2)
graph_comp$i <- 1:nrow(graph_comp)
ggplot(data=graph_comp,aes(x=i)) + geom_line(aes(y=mu_total1,colour="Décembre 2011")) + geom_line(aes(y=mu

#On ne trouve pas de saisonnalité annuelle car les bsplines annuelles ne sont pas prises en compte (coeff


################# Validation croisée utilisant l'erreur MMAE et des prédictions emboitées

#Après calculs, en suivant la procédure de l'article on peut comparer ici 547 échantillons => trop long co
#On en sélectionne donc 5 au hasard parmis les 547 et on fait des predictions sur les 480 heures suivantes
#Ceci en utilisant à chaque fois les 110 semaines + 1h précédentes : 110*120 + 1 = 13 201 heures
#La 13 201ème heure correspond toujours à la 23ème heure d'une journée
#Les heures possibles pour la 13201ème heures vérifient donc : 13201 <= 24x + 23
#et : 24x + 23 + 480 <= 26 816 (car nous avons 268016 obs complètes sans NA)
#on trouve : x compris entre 550 et 1096

x = as.matrix(df_lasso[1001:27336,2:1018])   # Pour nous 1 semaine = 120 heures, 4 semaines = 480 heures =
dim(x)
y = df_lasso$price[1001:27336]


#ATTENTION : "Supply instead a decreasing sequence of lambda values." pour que glmnet fit aille vite
cross_validation <- function(lambda_vect, x, y){
  lamb_leng = length(lambda_vect)
  v_poss = sample(550:1096,5)
  errors = matrix(data = rep(0,lamb_leng*5), ncol = 5)

  for (v in 1:5){
    l = 24*v_poss[v] + 23
    x_l = x[(l-13200):l,1:1017]
    y_l = y[(l-13200):l]
    fit_l = glmnet(x_l, y_l, intercept = TRUE, lambda = lambda_vect)
    coef_l = as.matrix(coef(fit_l))

    for (lamb in 1:lamb_leng){
      y_l_predicted = y_l
      for (h in 1:480){
        new_x = df_lasso[l+1000+h,2:17]

        for (u in 1:1000){
```

```
          lag_p = y_l_predicted[13201+h-u]
          new_x = cbind(new_x, lag_p)
        }
        new_x = cbind(new_x, df_lasso[l+1000+h,1018])
        new_x = cbind(1, new_x)
        y_l_predicted[13201+h] = as.matrix(new_x) %*% coef_l[,lamb]
      }
      calc = abs(y_l_predicted[13202:(13201+480)]- y[13202:(13201+480)])
      errors[lamb,v] = sum(calc)
    }
  }
  return(errors)
}

MMAE <- function(errors){
  MMAE_vect = rep(0,nrow(errors))
  for (lamb in 1:nrow(errors)){
    MMAE_vect[lamb] <- sum(errors[lamb,1:5])
  }
  MMAE_vect = MMAE_vect/(5*480)
  return(MMAE_vect)
}

### 1ère cross-vali : déterminer l'ordre de grandeur de lambda #
lambda_vect = c(10,1,0.1,0.01)
err = cross_validation(lambda_vect, x, y) #On trouve comme MMAE : 11.10298 11.05866 10.84570 12.53014
#0.1 semble donc être le bon ordre de grandeur

### 2ème cross-vali : affiner légèrement la valeur du paramètre lambda #
lambda_vect = c(0.5,0.2,0.1,0.05)
err = cross_validation(lambda_vect, x, y) #11.18410 10.70296 10.34403 10.35549

### 3ème cross-vali : affiner légèrement la valeur du paramètre lambda #
lambda_vect = c(0.15,0.1,0.05,0.01,0.005)
err = cross_validation(lambda_vect, x, y) #11.33090 11.09232 11.27518 13.69349 14.55084


lambda_vect = c(0.1,0.09,0.08,0.07,0.06,0.05,0.04)
err = cross_validation(lambda_vect, x, y) #10.45552 10.38233 10.34671 10.34917 10.40863 10.50540 10.66764

#0.08 semble donc être ici le meilleur lambda, choisissons le

### Attention le temps de calcul est très long, c'est pour cela qu'on ne teste pas tant de valeurs de lamb
```

Code modèle VAR(p) avec B-splines et régression LASSO

```
install.packages("glmnet")
install.packages("pbs")
install.packages("pomp")
library(questionr)
```

```
library(ggplot2)
library(glmnet)
library(splines)
library(pbs)
library(pomp)
library(moments)

setwd("C:/Users/Alexandre/Documents/projet electricite/")
bdd<-read.csv("C:/Users/Alexandre/Documents/projet electricite/prix spot germany/bdd_finales.csv",sep=";",
bdd<-bdd[,-1]
bdd$obs<-1:nrow(bdd)
bdd[,5]<- as.numeric(gsub(",",".",bdd[,5]))
bdd[,3]<- as.numeric(gsub(",",".",bdd[,3]))

acf(bdd[,6], lag= 72)
acf(na.omit(bdd[,4]), lag= 72)
diff.prix <- diff(bdd[,6], lag = 1)
diff.vent <- diff(na.omit(bdd[,4]), lag = 1)
acf(diff.vent, lag = 72)
acf(diff.prix, lag = 72)

x<- na.omit(bdd[,4])
skewness(x)
kurtosis(x)
sd(x)
hist(x)

plot(bdd$obs, bdd[,3], type="l")

qplot(bdd[,1], diff.prix)+ geom_line(size=0.2)
qplot(bdd[1:240,5], diff.vent[1:240])+ geom_line(size=0.2)

ggplot(bdd, aes(x=X, y=vent) , xlab = c("2010","2011","2012","2013","2014"),shape=Cond, color=blue) + geom



#################################### etude via b-splines ####################
############################"" Prix ##########"#####

perio = periodic.bspline.basis(bdd[,6], nbasis = 6, degree = 3, period = 24, names = NULL)
#perio = perio[,1:5]

perio_annuel  = periodic.bspline.basis(bdd[,6], nbasis = 12, degree = 3, period = 6264, names = NULL)
#perio_annuel = perio_annuel[,1:11]

bdd_lasso = cbind(bdd[,5], 1:nrow(bdd), perio,perio_annuel)
bdd_lasso= as.ts(bdd_lasso)
#34 coeff
lags= c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,48,49,50,72,73,74
```

```
for (u in lags){
  new_vect = lag(bdd[,5], -u)
  bdd_lasso = cbind(bdd_lasso, new_vect)
}

#2 j +1h comme ds l'article
for (u in 1:49){
  new_vect = lag(bdd[,3],  -u)
  bdd_lasso = cbind(bdd_lasso, new_vect)
}

#bspline quotidiens*lag prix
for (u in lags){
  for(i in 1:6)
  {
  new_vect = lag(bdd[,5], -u)*perio[,i]
  bdd_lasso = cbind(bdd_lasso, new_vect)
  }
}

#bspline annuels*lag prix
for (u in lags){
  for(i in 1:12)
  {
    new_vect = lag(bdd[,5], -u)*perio_annuel[,i]
    bdd_lasso = cbind(bdd_lasso, new_vect)
  }
}


df_lasso = data.frame(bdd_lasso)


names <- c("price","temps")
for (k in 1:6){
  nm <- paste("bspline_hebdo",k,sep="")
  names <- c(names,nm)
}


for (k in 1:12){
  nm <- paste("bspline_annuel",k,sep="")
  names <- c(names, nm)
}

for (u in lags){
  nm <- paste("lag",u,sep="")
  names <- c(names,nm)
}
```

```
for (u in 1:49){
  nm <- paste("lag_vent",u,sep="")
  names <- c(names,nm)
}

for(u in lags)
 {for (i in 1:6){
  nm <- paste("lag",u, "_bs_quot", i,sep="")
  names <- c(names,nm)
 }
}


for(u in lags)
{for (i in 1:12){
  nm <- paste("lag",u, "_bs_ann", i,sep="")
  names <- c(names,nm)
}
}

colnames(df_lasso) <- names

df_lasso<-na.omit(df_lasso)
x_prix =as.matrix(df_lasso[,2:ncol(df_lasso)])
y_prix =as.matrix(df_lasso[,1])

#write.csv(x_prix,"C:/Users/Alexandre/Documents/projet electricite/prix spot germany/x_prix.csv")
#write.csv(y_prix,"C:/Users/Alexandre/Documents/projet electricite/prix spot germany/y_prix.csv")
#x_prix<-read.csv("C:/Users/Alexandre/Documents/projet electricite/prix spot germany/x_prix.csv",sep=",",
#y_prix<-read.csv("C:/Users/Alexandre/Documents/projet electricite/prix spot germany/y_prix.csv", sep=",",
#x_prix = x_prix[,-1]
#y_prix = y_prix[,-1]
#####################  vent ################################

perio = periodic.bspline.basis(bdd[,6], nbasis = 6, degree = 3, period = 24, names = NULL)
perio = perio[,1:5]

perio_annuel  = periodic.bspline.basis(bdd[,6], nbasis = 6, degree = 3, period = 6264, names = NULL)
perio_annuel = perio_annuel[,1:5]

bdd_lasso =NULL
bdd_lasso = cbind(bdd[,3], 1:nrow(bdd), perio, perio_annuel)
bdd_lasso= as.ts(bdd_lasso)
#36 lags
lags= c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36

for (u in lags){
  new_vect = lag(bdd[,3], -u)
  bdd_lasso = cbind(bdd_lasso, new_vect)
}
```

```
#bspline quotidiens*lag prix
for (u in lags){
  for(i in 1:5)
  {
    new_vect = lag(bdd[,3], -u)*perio[,i]
    bdd_lasso = cbind(bdd_lasso, new_vect)
  }
}

#bspline annuels*lag prix
for (u in lags){
  for(i in 1:5)
  {
    new_vect = lag(bdd[,3], -u)*perio_annuel[,i]
    bdd_lasso = cbind(bdd_lasso, new_vect)
  }
}


df_lasso = data.frame(bdd_lasso)
##names

names <- c("vent","temps")
for (k in 1:5){
  nm <- paste("bspline_hebdo",k,sep="")
  names <- c(names,nm)
}


for (k in 1:5){
  nm <- paste("bspline_annuel",k,sep="")
  names <- c(names, nm)
}


for (u in lags){
  nm <- paste("lag",u,sep="")
  names <- c(names,nm)
}

for(u in lags)
{for (i in 1:5){
  nm <- paste("lag",u, "_bs_quot", i,sep="")
  names <- c(names,nm)
}
}

for(u in lags)
{for (i in 1:5){
```

```
  nm <- paste("lag",u, "_bs_ann", i,sep="")
  names <- c(names,nm)
}}

colnames(df_lasso) <- names

df_lasso<-na.omit(df_lasso)
x_vent =as.matrix(df_lasso[361:nrow(df_lasso),2:ncol(df_lasso)])
y_vent =as.matrix(df_lasso[361:nrow(df_lasso),1])

#write.csv(x_vent,"C:/Users/Alexandre/Documents/projet electricite/prix spot germany/x_vent.csv")
#write.csv(y_vent,"C:/Users/Alexandre/Documents/projet electricite/prix spot germany/y_vent.csv")

#x_vent<-read.csv("C:/Users/Alexandre/Documents/projet electricite/prix spot germany/x_vent.csv",sep=",", h
#y_vent<-read.csv("C:/Users/Alexandre/Documents/projet electricite/prix spot germany/y_vent.csv", sep=",",
#x_vent = x_vent[,-1]
#y_vent = y_vent[,-1]
##### evaluation des coeff de régression pour prix
x_prix_coeff = as.matrix(x_prix[1:26496,])
y_prix_coeff = as.matrix(y_prix[1:26496])

cvfit = cv.glmnet(x_prix_coeff , y_prix_coeff, intercept = TRUE)
#plot(cvfit)
#cvfit$lambda.min
coeffs_prix = coef(cvfit, s = 0.245) ## lambda = 0.003451849 -> 2.124466e-05
coeffs_prix = as.matrix(coeffs_prix) #On a un coeff devant trend qui est négatif comme l'article
#write.csv(coeffs_prix,"C:/Users/Alexandre/Documents/projet electricite/prix spot germany/coeff_prix.csv")

count =0
for(i in 1:nrow(coeffs_prix))
{
  if(coeffs_prix[i,1]>0){count=count+1}
}
show(count)

##### evaluation des coeff de régression pour vent
x_vent_coeff = as.matrix(x_vent[1:26496,])
y_vent_coeff = as.matrix(y_vent[1:26496])

cvfit = cv.glmnet(x_vent_coeff, y_vent_coeff, intercept = TRUE)
#plot(cvfit)
#cvfit$lambda.min
coeffs_vent = coef(cvfit, s = 1.015) #### 0.36 ok car 0.3592449/sqrt(26400)->0.002211 graph
coeffs_vent = as.matrix(coeffs_vent) #On a un coeff devant trend qui est négatif comme l'article
#write.csv(coeffs_vent,"C:/Users/Alexandre/Documents/projet electricite/prix spot germany/coeff_vent.csv")

#pourcentage non nul
count =0
for(i in 1:nrow(coeffs_vent))
{
```

```
  if(coeffs_vent[i,1]>0){count=count+1}
}
show(count)
############################ PREDICTION ################
#coeffs_prix<-read.csv("C:/Users/Alexandre/Documents/projet electricite/prix spot germany/coeff_prix.csv",
#coeffs_vent<-read.csv("C:/Users/Alexandre/Documents/projet electricite/prix spot germany/coeff_vent.csv",
#coeffs_prix= as.matrix(coeffs_prix[,-1])
#coeffs_vent= as.matrix(coeffs_vent[,-1])

x_prix = as.matrix(x_prix)
x_vent = as.matrix(x_vent)
lags1= c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,48,49,50,72,73,74
lags2= c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36

#lags1= c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,48,72,96,120,240,360,480,600,720,8
#lags2= c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,48,49,72,73,96,97,120,121,240,3

y_predicted_prix = as.matrix(y_prix[1:26496])
y_predicted_vent =  as.matrix(y_vent[1:26496])

for (h in 1:480){

  new_prix_x = matrix(x_prix[26496 + h,1:19], nrow=1)
  new_vent_x = matrix(x_vent[26496+ h,1:11], nrow=1)

  for (u in lags1){
    lag_p = y_predicted_prix[26496+h-u]
    new_prix_x = cbind(new_prix_x, lag_p)
  }

  for (u in 1:49){
    lag_p = y_predicted_vent[26496+h-u]
    new_prix_x = cbind(new_prix_x, lag_p)
  }

  for (u in lags1){
    for( k in 1:6)
    {
    lag_p = y_predicted_prix[26496+h-u]*new_prix_x[k+1]
    new_prix_x = cbind(new_prix_x, lag_p)
    }
  }

  for (u in lags1){
    for( k in 1:12)
    {
      lag_p = y_predicted_prix[26496+h-u]*new_prix_x[k+7]
      new_prix_x = cbind(new_prix_x, lag_p)
    }
  }
```

```
  ###### vent

  for (u in lags2){
    lag_p = y_predicted_vent[26496+h-u]
    new_vent_x = cbind(new_vent_x, lag_p)
  }

  for (u in lags2){
    for( k in 1:5)
    {
      lag_p = y_predicted_vent[26496+h-u]*new_vent_x[k+1]
      new_vent_x= cbind( new_vent_x, lag_p)
    }
  }

  for (u in lags2){
    for( k in 1:5)
    {
      lag_p = y_predicted_vent[26496+h-u]*new_vent_x[k+5]
      new_vent_x= cbind( new_vent_x, lag_p)
    }
  }

  new_prix_x = cbind(1, new_prix_x)
  new_vent_x = cbind(1, new_vent_x)
  y_predicted_prix[26496+h] = as.matrix(new_prix_x) %*% coeffs_prix
  y_predicted_vent[26496+h] = as.matrix(new_vent_x) %*% coeffs_vent
}

#y_predicted_prix[26401:26880]
#y_predicted_vent[26401:26880]

y_vrai = as.matrix(y_prix[26497:26976])
y_pred = as.matrix(y_predicted_prix[26497:26976])

Y_test = data.frame(y_pred,y_vrai)
ggplot(data=Y_test,aes(x=1:nrow(Y_test))) + geom_line(aes(y=y_vrai,colour="Valeur réelle")) + geom_line(ae

y_vrai_vent = as.matrix(y_vent[26497:26976])
y_pred_vent = as.matrix(y_predicted_vent[26497:26976])

Y_test_vent = data.frame(y_pred_vent,y_vrai_vent)
ggplot(data=Y_test,aes(x=1:nrow(Y_test_vent))) + geom_line(aes(y=y_vrai_vent,colour="Valeur réelle")) + ge

##############" MMAE
MMAE = 1/(24)*t(rep(1,24))%*%abs(y_vrai[1:24]-y_pred[1:24])
MMAE = 1/(120)*t(rep(1,120))%*%abs(y_vrai[1:120]-y_pred[1:120])
MMAE = 1/(480)*t(rep(1,480))%*%abs(y_vrai[1:480]-y_pred[1:480])
### 5.414306 l=0.442
```

```
###  5.432353 l =0.445
### 5.464035 l =0.45
```

MMAE