

Table des matières

Résumé.....	III
Abstract.....	V
Liste des figures.....	IX
Liste des tableaux.....	XI
Liste des abréviations.....	XIII
Dédicace.....	XV
Remerciements.....	XVII
Avant-propos.....	XIX
Introduction.....	1
1. Revue de la littérature	3
1.1 Revue de littérature méthodologique	3
1.1.1 Modèles dynamiques univariés	3
1.1.2 Modèle de lissage exponentiel Holt-Winters.....	6
1.1.3 Les modèles avec variables explicatives exogènes	7
1.1.4 L'approche basée sur l'intelligence artificielle.....	8
1.2 Revue des recherches et des résultats.....	8
2. Méthodologie	Erreur ! Signet non défini.
2.1 Le lissage exponentiel de Holt-Winters.....	12
2.2 Le modèle ARIMA saisonnier (SARIMA)	15
2.3 Le modèle ARIMA saisonnier avec variable exogènes.....	16
2.4 Procédure d'estimation des modèles SARIMA	17
2.4.1 Test de racine unitaire	17
2.4.2 Saisonnalité et identification des modèles ARIMA saisonnier	18
2.4.3 Prévisions et mesures de performance prédictives	19
3. Données.....	20
3.1 Statistiques descriptives agrégées	20
3.2 Profils chronologiques.....	23
4. Résultats.....	28
4.1 Modèles de lissage exponentiel.....	29
4.2 Modèles SARIMA.....	35
4.2.1 Tests de racine unitaire	35
4.2.2 Estimations SARIMA	36
4.3 Diagnostic des modèles.....	43
5. Conclusion	44
Références bibliographiques	47
ANNEXE I.....	51

Liste des figures

Figure 1 : Relation entre les variables climatiques et la consommation horaire d'électricité.....	21
Figure 2 : profil de consommation d'électricité au Québec de 2006 à 2010 (43824heures).....	23
Figure 3 : Profil de consommation d'une semaine type de chaque saison en 2006 et 2010.....	24
Figure 4 : Profil annuel de la consommation d'électricité à des heures spécifiques.....	25
Figure 5 : Profil annuel de consommation d'électricité en moyenne mobile 30 jours à des heures spécifiques	26
Figure 6 : Périodogramme de la consommation d'électricité au Québec en fréquence horaire	27
Figure 7 : ACF et PACF de la consommation horaire d'électricité	28
Figure 8 : Profil de prévision de la consommation d'électricité avec le modèle ETS(MAdA) pour 24h	31
Figure 9 : Profil de prévision de la consommation d'électricité avec le modèle ETS(MAdA) pour 72h.	31
Figure 10: Profil de prévision de la consommation d'électricité avec le modèle ETS(MAdA) pour 168h	32
Figure 11 : Profil de prévision de la consommation d'électricité avec le modèle ETS(AAdA) pour 24h.....	34
Figure 12 : Profil de prévision de la consommation d'électricité avec le modèle ETS(AAdA) pour 72h.....	34
Figure 13 : Profil de prévision de la consommation d'électricité avec le modèle ETS(AAdA) pour 168h.....	35
Figure 14 : Profil de prévision de la consommation d'électricité avec le modèle SARIMA pour 24h.....	38
Figure 15 : Profil de prévision de la consommation d'électricité avec le modèle SARIMA pour 72h.....	39
Figure 16 : Profil de prévision de la consommation d'électricité avec le modèle SARIMA pour 168h.....	39
Figure 17 : Profil de prévision de la consommation d'électricité avec le modèle SARIMAx pour 24h.....	41
Figure 18 : Profil de prévision de la consommation d'électricité avec le modèle SARIMAX pour 72h.	42
Figure 19 : Profil de prévision de la consommation d'électricité avec le modèle SARIMAX pour 168h.	42

Liste des tableaux

Tableau 1 : Statistiques récapitulatives de la consommation d'électricité sur cinq (5) années au Québec.....	22
Tableau 2 : Statistiques récapitulatives de la température sur cinq (5) années au Québec.....	22
Tableau 3 : Comparaison du critère d'information d'Akaike (AIC).	29
Tableau 4 : Comparaison des mesures de l'erreur intra échantillon	29
Tableau 5 : Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle MAdA.	30
Tableau 6 : Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle AAdA.....	33
Tableau 7: Résultat du test de racine unitaire avec un retard maximum $k=168$	35
Tableau 8: Coefficients et statistiques (SARIMA).....	37
Tableau 9: Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle SARIMA ..	38
Tableau 10: Coefficients et statistiques(SARIMAX)	40
Tableau 11: Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle SARIMAX	41

Liste des Abréviations

ACF	Autocorelation Fonction
AIC	Akaike information criteria
AR	Autoregressive
ARX	Autoregressive with exogenous variables
ARIMA	<i>Autoregressive integrated moving average</i>
ARIMAX	<i>Autoregressive integrated moving average with exogenous variables</i>
ARMA	Autoregressive moving average
ARMAX	Autoregressive moving average with exogenous variables
ETS	Exponential smoothing
NARX	Non-linear autoregressive with exogenous variables
MA	Moving Average
MAE	Mean Absolute Error (Erreur absolue moyenne)
MAPE	Mean Absolute Percent Error (Écart absolu moyen en pourcentage)
MASE	Mean Absolute Squared Error (Erreur quadratique absolue moyenne)
ME	Mean Error (Erreur moyenne)
MPE	Mean Percent Error (Écart moyen en pourcentage)
PACF	Partial Autocorelation Fonction
RMSE	Root Mean Squared Error (Racine carré de l'erreur moyenne)

*A Suzy Ken
et
Adeline*

Remerciements

Mes remerciements vont à toutes les personnes du département d'Économique de l'Université Laval qui ont contribué chacune à leur manière à la réalisation de ce travail. Je les adresse tout particulièrement :

Au Pr. Carlos Ordás Criado, qui a fait preuve de beaucoup d'engagement pour la réalisation de ce travail malgré sa lourde charge de travail;

Au Pr. Sylvain Dessy pour ses encouragements lors de mon engagement dans le programme de maîtrise;

À tous mes professeurs du département qui ont contribué à élargir mes connaissances dans différents domaines de l'économie;

À tous mes camarades du programme avec qui j'ai eu des échanges toujours constructifs;

À Hydro-Québec qui a gentiment mis notre disposition les données nécessaires à cette étude.

Ils vont aussi à ma fille pour sa patience et à ma très chère conjointe, Adeline, qui m'ont encouragé jour après jour lors de la réalisation de ce travail.

Avant-propos

Ce mémoire est un travail de recherche en économétrie appliquée aux séries chronologiques. L'économétrie des séries chronologiques est d'une importance cruciale dans la recherche empirique. Ses domaines d'applications vont de l'identification des propriétés statistiques fondamentales des séries économiques dans leur dimension temporelle, à la modélisation à des fins de simulation et de prévision. C'est dans cette logique que nous conduisons cette recherche. Je remercie le Pr. Carlos Ordás Criado du département d'Économie de l'Université Laval qui a su susciter en moi l'intérêt pour comprendre et utiliser ces méthodes. Ce mémoire m'a permis de compléter ma formation d'économiste en méthodes quantitatives.

L'objectif de ce mémoire est d'explorer les performances de trois modèles de base en séries temporelles pour analyser la consommation horaire d'électricité au Québec. La modélisation permettant d'effectuer des prévisions, nous nous sommes concentrés sur la prévision à court terme, en prenant en compte tant l'inertie de la consommation que l'effet de variables climatiques fondamentales. Cette recherche est un premier pas dans le développement d'outils rigoureux d'aide à la décision, qui pourraient être utiles aux opérateurs du secteur, principalement Hydro-Québec.

Toute l'analyse s'est faite à l'aide du logiciel statistique R. À cet égard, je remercie ici toute la communauté des chercheurs qui travaillent continuellement à améliorer cet outil de travail de qualité, accessible gratuitement et d'une grande puissance.

Étant conscient que la science est basée sur la critique, je suis toujours intéressé à recevoir toutes les remarques ou suggestions de correction ou d'amélioration de ce qui a été présenté dans ce mémoire.

Introduction

La production d'électricité revêt dans chaque pays des enjeux économiques, de sécurité et de bien-être social : l'électricité est un facteur de production fondamental dans l'activité économique moderne. La régularité de son offre pose des défis particuliers puisque ce flux continu est difficilement stockable. La prévision à court terme de la consommation d'électricité joue un rôle essentiel dans la gestion efficace des ressources allouées à la production d'électricité. Les erreurs de prévisions peuvent, en effet, occasionner des coûts opérationnels importants (Haida et Muto, 1994). D'après Hobbs et al. (1999), une réduction de l'erreur moyenne de prévision de 1% peut épargner des milliers, voire même des millions de dollars dans une unité de production d'électricité. Soares et Medeiros (2008) citent l'estimation effectuée en 1984 dans une installation électrique au Royaume-Uni, où un accroissement de 1% de l'erreur de prévision avait causé un accroissement annuel des coûts opérationnels de 10 millions de livres sterling. Lorsque le secteur de l'électricité est fortement régulé¹, les opérateurs en situation de monopole utilisent la prévision à court terme pour assurer la fiabilité de l'offre (Weron, 2006). Ainsi selon cet auteur, du point de vue de l'opérateur, prévoir la consommation d'électricité est nécessaire pour la constitution d'outils d'aide à la décision et pour minimiser les coûts de sur/sous-production qui ne sont pas facilement transférables dans les prix.

Du point de vue de la théorie économique, les modèles de prévision à court et moyen terme se concentrent sur une modélisation des points d'équilibre, où l'offre et la demande s'égalisent à un instant t . À l'équilibre, l'opérateur offre la quantité demandée à laquelle s'ajoute les pertes de charge. À court terme, la demande d'électricité est très dépendante des variables liées aux conditions climatiques, à l'organisation sociale et au cycle économique (Weron et Misiorek, 2008). Elle est donc sujette à d'importantes fluctuations. Du point de vue de l'offre, Taylor (1975) précise que la production d'électricité dépend de la capacité productive des installations électriques. Elle est fixe à court terme et elle se mesure en général en watt.

Dans ce travail, nous nous concentrons sur la consommation domestique (résidentielle et agricole) d'électricité de la province du Québec de la période du 1^{er} Janvier 2006 au 31 décembre 2010. Ce marché est fortement régulé et Hydro-Québec y est l'opérateur unique, verticalement intégré (production, transport, distribution d'électricité). Cette entreprise exploite essentiellement l'énergie

¹ Ce qui est le cas au Québec.

d'origine l'hydraulique (95%), le reste étant composé d'énergie d'origine nucléaire, gazière et éolienne, généralement achetée auprès de producteurs privés. Selon le Ministère des Ressources naturelles et de la Faune du Québec et Statistique Canada, en termes de consommation finale par secteur d'activité au cours de ces dix dernières années², le secteur industriel arrive en tête pour 44.8% approximativement, suivi du secteur résidentiel, 34.4%. Le secteur du transport est celui dont la consommation est la plus faible (0.2%) tandis que le secteur commercial consomme 20.06%. La consommation totale était en baisse en 2009 de 2.14% par rapport à 2008.

Dans ce travail, nous nous intéressons à la consommation électrique domestique de l'ensemble de la province de Québec. Durant la période d'analyse, cette consommation a été légèrement croissante de 2006 à 2009 et en léger repli en 2010. Par conséquent, la série ne présente pas de tendance haussière prononcée. Le pic de consommation a été observé le 16 janvier 2009 à 7h du matin. Les 18120 Mwatt³ consommés ce jour-là représentent un écart de 40% par rapport à la moyenne de ce même mois de janvier qui était de 12 725 Mwatt. Un tel écart est emblématique de la pression qui peut être imposé au réseau électrique à très court terme, d'où l'importance d'estimer des modèles capables de prédire la consommation horaire avec fiabilité.

Dans notre recherche, nous nous intéressons à la modélisation de la consommation d'électricité à court terme avec des techniques de séries chronologiques qui prennent en compte plusieurs types de saisonnalités liées aux conditions climatiques et à l'organisation de l'activité économique. La température, l'humidité, les points de rosé, la nébulosité sont les principales variables climatiques exogène qui influencent directement la demande horaire d'électricité. Les événements particuliers liés aux jours fériés affectent l'organisation économique et agissent aussi directement sur la consommation d'électricité. Notons que très peu d'études récentes ont été consacrées à la modélisation de la consommation de l'électricité pour la région du Québec, en particulier sur des données à haute fréquence. À notre connaissance, celles qui se sont intéressées à la consommation d'électricité se sont limitées aux approches microéconométriques et microéconomiques⁴. Plusieurs méthodologies sont à disposition pour modéliser la consommation d'électricité à court terme. On peut les classer en deux grandes catégories : *i) les modèles basés sur les techniques économétriques traditionnelles (lissage exponentiel simple, méthode Holt-Winters, régression linéaire ou robuste,*

²Ministère des Ressources naturelles et de la Faune du Québec. « Consommation d'électricité [archive] ». <http://www.mrn.gouv.qc.ca/>. Site internet consulté le 2012-08-14.

³ D'après notre base de données sur la consommation d'électricité au Québec.

⁴Voir Bernard, J. T., D. Bolduc, et al. (1996) ; Bernard, J. T., D. Bolduc, et al. (2011).

modèles autorégressifs et de moyennes mobiles) et *ii) les techniques basées sur l'intelligence artificielle (le réseau de neurones, système de réseau flou, système expert et machine à vecteur support)*. Nous nous concentrons ici sur trois méthodes économétriques de séries temporelles.

Ce travail est divisé en cinq sections. La première propose une revue de littérature sur nos modèles d'intérêt et les travaux les plus récents sur le sujet. La seconde section décrit avec plus de détails la méthodologie empirique retenue. La description des données est faite en section trois. Nos résultats figurent en section quatre et nous exposons nos conclusions dans la section 5.

1. Revue de la littérature

Au fil des années, différents techniques ont été développés pour modéliser la charge d'électricité tant avec les outils classiques de l'économétrie des séries temporelles⁵, qu'avec les méthodes de l'intelligence artificielle⁶. Certaines études se concentrent sur la comparaison de modèles spécifiques à l'une de ses approches, alors que d'autres comparent ces deux classes de modèles. Dans la section suivante, nous tâchons dans un premier temps de donner une vision globale de l'ensemble des modèles à disposition pour étudier la consommation d'électricité dans un but de prévision. En section 1.2, nous proposons une revue de littérature de travaux récents dans le domaine.

1.1 Revue de littérature méthodologique

Ces approches se subdivisent en deux catégories : l'une se concentre exclusivement sur les propriétés temporelles de la variable étudiée (méthodes univariées) tandis que l'autre considère la relation entre la variable étudiée, son passé et d'autres variables explicatives exogènes. La première approche utilise exclusivement l'inertie de la série pour prédire ses réalisations futures, la seconde approche introduit dans la dynamique de la chronique l'influence d'autres variables explicatives.

1.1.1 Modèles dynamiques univariés

⁵ Voir Bunn et Farmer, 1985a

⁶ Voir Hippert et al. (2001); Mohandes (2002); Chen et Chang (2004); Metaxiotis et al. (2003); Weron (2006, 75-78);

La modélisation univariée d'une série temporelle se fait en général à l'aide d'une régression linéaire qui inclut deux composants : un terme autorégressif (AR) qui établit un lien (généralement linéaire) entre les réalisations présentes et passées de la variable d'intérêt et une composante de moyenne mobile (MA), qui établit un lien (généralement linéaire) entre les déviations aléatoires présentes et passées entre les réalisations de la série en tous temps et les valeurs intra-échantillon prédites par le modèle. Dans notre contexte, lorsqu'un modèle purement $AR(p)$ est utilisé pour modéliser le profil temporel de consommation d'électricité, la réalisation présente de cette variable est une combinaison linéaire des consommations observées aux p périodes précédentes auquel on ajoute une erreur de mesure aléatoire pour la période présente. L'avantage du modèle autorégressif est sa simplicité. Son inconvénient est que les méthodes autorégressives présupposent la stationnarité des séries. Une série est dite *faiblement*⁷ stationnaire si sa moyenne, $E(z_t) = \mu_z$ ne dépend pas l'indice de temps t et si la covariance entre z_t et z_{t-k} dépend seulement du décalage k . Si une série est stationnaire et normalement distribuée alors, la moyenne (μ_z) et la variance (σ_z^2) suffisent pour caractériser la série. Cependant, la covariance est importante pour l'identification des fonctions d'autocorrélation (Aragon, 2011 ; 57).

L'économétrie distingue en général deux types de non stationnarité : déterministe et stochastique. Un processus non stationnaire *déterministe* à la forme générale $y_t = f(t) + \varepsilon_t$, où $f(t)$ est une fonction déterministe du temps (t) et ε_t est un terme d'erreur aléatoire de moyenne nulle, de variance finie et en général indépendant et identiquement distribué. Un processus non stationnaire stochastique est un processus explosif vis-à-vis de ses réalisations passées. Le plus connu d'entre eux est la marche aléatoire (ou processus avec une racine unitaire sans dérive), qui a la forme fonctionnelle $y_t = y_{t-1} + \varepsilon_t$. Il est facile de montrer que la variance de ce processus croît avec t . Une simple transformation des données permet en général de stabiliser les séries non stationnaires. Dans le cas de la tendance déterministe, il faudra identifier la forme paramétrique de $f(t)$ et la soustraire à y_t pour obtenir une transformation stationnaire de la série. Dans le cas de la tendance stochastique, la différenciation des données permet de rendre le processus stable. Notons qu'un processus peut posséder une tendance à la fois déterministe et stochastique.

⁷ Dans l'étude de la stationnarité, on distingue la *stationnarité stricte* et la *stationnarité faible*. La stationnarité stricte implique que $f(x_1, \dots, x_t)$ et $f(x_{1+k}, \dots, x_{t+k})$ sont de même loi.

Les séries chronologiques de consommation d'électricité sont généralement non stationnaires car, soit elles croissent dans le temps (la moyenne croît avec t), soit la variance croît dans le temps, soit les deux. Cette non stationnarité en moyenne et/ou variance n'est pas problématique car elle se traite comme indiqué ci-dessus.

Une autre approche de modélisation des séries chronologiques est la technique des moyennes mobiles (*MA*) sur les erreurs. Cette technique considère qu'une chronique peut s'exprimer comme une combinaison linéaire de *bruits blancs*⁸ à l'horizon $t-q$. Lorsqu'un processus $MA(q)$ est *inversible*⁹, il peut être alors combiné à un processus autorégressif $AR(p)$ pour construire un modèle à composants $AR(p)$ et $MA(q)$, appelé $ARMA(p,q)$ (Aragon, 2011;70-71). Selon Weron (2006;83), la méthode des moyennes mobiles n'est pas efficace pour la modélisation de la consommation d'électricité et n'est généralement utile que pour le *lissage* des séries chronologiques. Pour Aragon (2011;67), l'utilisation des deux composantes permet de mieux capturer la dynamique sous-jacente, qui est le fruit de l'autocorrélation, des variations saisonnières et d'aspects aléatoires. La combinaison des deux modèles pose parfois des problèmes d'*inversibilité* du modèle $MA(q)$. Il convient alors de vérifier que les racines du polynôme caractéristique¹⁰ du processus $MA(q)$ sont strictement supérieures à 1. Lorsque cette condition est vérifiée, on dit que le modèle $MA(q)$ est *inversible*.

Il existe une version saisonnière du modèle $ARMA(p,q)$, qui permet aux composants *AR* et *MA* de s'adapter aux cycles observés (semaine, mois, saison). On dénote ses modèles par $SARMA(p,q)(P,Q)_s$ ou encore $ARMA(p + sP, q + sQ)$, où s est la cyclicité du processus (ex : $s=12$ pour un cycle mensuel), p et q sont les paramètres habituels de notre ARMA non saisonnier et P , Q et s sont des paramètres spécifiques à la composante saisonnière. En effet, un processus ARMA qui présente des cycles ne pourra pas être caractérisé par une structure d'erreurs aléatoires sans autocorrélation et les termes aléatoires du modèle ne respecteront pas l'hypothèse de « *bruits blancs* ». Il est alors indiqué de résoudre le problème en modélisant le résidu lui-même par un ARMA

⁸ En l'occurrence, il s'agit d'un bruit blanc, c'est-à-dire une variable aléatoire ε_t qui a les caractéristiques suivantes : $E(\varepsilon_t) = \mu$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t \varepsilon_s) = 0 \quad \forall t \neq s$.

⁹ Un processus $MA(q)$ est *inversible* s'il peut être représenté comme un processus autorégressif infini, voir Aragon (2011 :67).

¹⁰ $1 + \theta_1 z + \dots + \theta_q z^q = 0$



ayant pour unité de temps la période de la saisonnalité. On obtient alors un modèle ARMA saisonnier sous l'hypothèse de la stationnarité (Aragon, 2011;83).

Des modèles dynamiques équivalents à ceux présentés ci-dessus existent pour les séries non stationnaires, pour lesquelles la non stationnarité est d'origine stochastique. Il s'agit des modèles $ARIMA(p,d,q)$ et $SARIMA(p,d,q)(P,D,Q)_s$, qui sont des modèles adaptés aux séries qui deviennent stationnaires par différenciation d'ordre d pour la partie non saisonnière et d'ordre D pour la partie saisonnière.

1.1.2 Modèle de lissage exponentiel Holt-Winters

Le lissage exponentiel englobe une série de méthodes intuitives de lissage et de prévision apparues dans les années 50. Au fil des années, ces méthodes ont laissé place à des spécifications rigoureuses. Ces techniques permettent de mettre à jour les prédictions en $t+1$ sur la base de moyennes pondérées des valeurs passées. Dans sa version la plus simple, le lissage exponentiel s'exprime par l'équation : $\mu_t = c_0 y_t + c_1 y_{t-1} + c_2 y_{t-2} + \dots$, où les poids $c_t \geq 0$ peuvent être définis de nombreuses manières, voir Aragon (2011;121) pour une présentation concise. Si μ_t constitue notre prévision de y en $t+1$, il est intuitif d'attribuer plus de poids aux valeurs y_t récentes. Le modèle de base sous cette hypothèse attribue des poids qui décroissent exponentiellement, selon la formule $c_i = \alpha(1 - \alpha)^i, i = 0, 1, \dots$ et $0 \leq \alpha \leq 1$, d'où l'appellation de lissage exponentiel. Plus le paramètre α est proche de 1, plus le passé immédiat influence la prévision μ_t . Ce schéma de pondération conduit à l'expression équivalente $\mu_t = \alpha y_t + (1 - \alpha)\mu_{t-1}$, dans laquelle la mise à jour de la prévision en $t+1$ s'effectue facilement dès qu'une information en t est connue, en séparant la contribution du passé lointain et du présent immédiat. Dans ce travail, nous nous concentrons sur des méthodes de lissage exponentiel de type Holt-Winters, qui permettent d'ajouter à la composante autorégressive du modèle, une tendance et une saisonnalité. Notons également que le lissage exponentiel peut s'exprimer sous la forme de modèles ARIMA spécifiques, voir à ce sujet Hyndman et al (2011, Ch.11).

Jusqu'à présent, nous avons laissé de côté les éléments de nature exogène qui influencent l'évolution d'une chronique. Il est donc important de s'intéresser à l'inclusion de variables exogènes dans le processus de génération de données temporelles.

1.1.3 Les modèles avec variables explicatives exogènes

Certaines méthodes statistiques de prévision de la consommation d'électricité basées sur les séries chronologiques utilisent la consommation passée et les valeurs courantes ou passées des variables explicatives exogènes pour prévoir la consommation courante d'électricité. La régression multiple basée sur l'estimateur des moindres carrées est la plus utilisée. Il s'agit ici de rechercher le meilleur ajustement de la variable expliquée aux variables explicatives telle que la somme des carrés de résidus soit minimisée (Weron, *op. cit.*;81). Ainsi, lorsqu'un processus $AR(p)$, $ARMA(p,q)$ ou $ARIMA(p,d,q)$ est identifié pour modéliser une série chronologique, on peut lui ajouter des variables explicatives exogènes dont on soupçonne l'influence directe. On obtient alors des modèles appelés ARX , $ARMAX$ ou $ARIMAX$. La composante X du modèle indique que le processus dépend de variables explicatives exogènes. Les méthodes de modélisation de la consommation d'électricité utilisent par exemple les variables liées aux conditions climatiques telles que la température, l'humidité, etc. Lorsque l'effet de ces variables explicatives sur la consommation est non linéaire¹¹, certains auteurs font appel à une *fonction (non-linéaire) de transfert*, qui spécifie l'incidence des processus, tels que le chauffage et la climatisation, sur la relation entre les variations des variables climatiques et de consommation d'électricité. Bisgaard et Kulahci (2011;203-261) ont présenté une méthodologie d'identification des *fonctions de transfert dynamique*, basée sur la théorie développée par Box et Jenkins (1969), qui permet de mieux capter les mécanismes de la relation non linéaire (« avec retard ») entre les variables explicatives et la variable expliquée. Selon ces deux auteurs, elle est définie par opposition à la *régression linéaire simple* qui spécifie un effet immédiat d'une variable explicative sur une variable dépendante. Plusieurs autres techniques de régression permettent de tenir compte de la sélection des variables, de la corrélation entre les variables et de l'existence des valeurs extrêmes dans les variables et les résidus : ce sont les techniques de *régression robuste*¹².

L'analyse en composante principale (ACP) est adaptée à la situation où il existe un grand nombre de variables explicatives dans un modèle de régression, dont l'effet individuel est difficilement identifiable. L'ACP a pour but de réduire la dimension d'un ensemble de variables explicatives

¹¹ Les ménages utilisent les dispositifs de chauffage ou de climatisation seulement à partir de certains niveaux de température extérieure.

¹² Voir Tibshirani (1996); Wang et Jiang (2007); Wang et Tsai (2007).

corrélées. Il s'agit de décrire les variations d'un ensemble de variable corrélée par un sous ensemble de variables non corrélées dérivé par combinaison linéaire des variables initiales. Étant donné le grand nombre de variables qui peuvent influencer la consommation d'électricité à court terme, certains auteurs ont recours à ces méthodes (voir-ci-dessous).

1.1.4 L'approche basée sur l'intelligence artificielle

Hahn, Meyer-Nieberg et Pickl (2009) ont recensé plusieurs études sur la prévision de la consommation de l'électricité basée sur l'intelligence machine : *Les réseaux neuronaux artificiels, système de réseau flou, système expert et machine à vecteur support*. La particularité de ces méthodes est qu'elles utilisent les techniques de l'intelligence artificielle (i.e. la programmation sur ordinateur) et tiennent compte de relations complexes et non linéaires entre les variables endogènes et exogènes et la consommation courante d'électricité que les méthodes de régression peinent à appréhender. Selon Weron (*op.cit.*;75-78), ces méthodes sont une « *boîte noire* » et leur performance prédictive n'est pas entièrement convaincante. Nous ne faisons pas une présentation détaillée de cette littérature car notre but est d'utiliser les méthodes de séries chronologiques.

1.2 Revue des recherches et des résultats

Taylor, Menezes et McSharry (2006) comparent la précision prédictive à court terme (jusqu'à un jour) de six méthodes univariées d'estimation de la demande d'électricité. Les approches analysées incluent le modèle ARIMA à double saisonnalité multiple, le lissage exponentiel pour double saisonnalité et une nouvelle méthodologie basée sur l'analyse en composantes principales. Les méthodes sont appliquées à la demande d'électricité horaire de Rio de Janeiro entre le 5 mai 1996 et le 30 novembre 1996 et à la demande pour chaque demi-heure en Angleterre et au Pays de Galles couvrant la période du 27 mars 2000 au 22 octobre 2000. La méthode de lissage exponentielle par double saisonnalité performe bien avec les deux séries puisqu'elle fournit le pourcentage de l'erreur absolue moyenne de prévision le plus faible. Il ressort également de l'étude que les approches les plus simples et robustes, qui exigent peu de connaissances spécifiques à l'industrie électrique, peuvent surpasser des modèles plus complexes.

En appliquant une méthodologie similaire à celle de Taylor et al. (2006), Taylor et McSharry (2007) ont utilisé les données intra journalière de consommation d'électricité de dix pays européens pour effectuer de la prévision à court terme. Ils définissent un modèle ARIMA, un modèle AR périodique, une extension pour double saisonnalité par lissage exponentiel de *Holt-Winters* et une méthode basée sur l'analyse en composantes principales. Leurs modèles qui tiennent compte de la présence d'un cycle journalier et hebdomadaire dans les données sont appliqués aux données journalières de consommation d'électricité des dix pays sur une période de trente semaines (3 avril 2005 au 29 octobre 2005). La comparaison de la performance prédictive de ces méthodes aboutit à peu de différences entre modèles. Les méthodes ARIMA et de l'analyse en composantes principales se montrent les plus performantes en terme d'erreurs absolues.

Amjadi (2001) propose une méthodologie basée sur seize modèles ARIMA modifiés pour modéliser la consommation d'électricité en Iran. Ces modèles sont issus du modèle ARIMA régulier (Box-Jenkins, 1976). Pour modifier les modèles, une variable qui est la prévision initiale de la variable dépendante effectuée par l'opérateur du système et une variable de température calculée sont inclus dans le modèle. La variable de température est calculée à partir de trois sections isothermiques définies sur la base des différents types de climat observé en Iran. La température horaire utilisée dans chaque section est finalement obtenue en pondérant la température observée par la proportion de la consommation d'électricité de chaque section. Les modèles ARIMA modifiés sont subdivisés en deux catégories comprenant chacune huit modèles : la première catégorie modélise la consommation horaire des jours de semaine et de weekend et la deuxième catégorie est utilisée pour modéliser les pics journaliers de consommation. Une variable supplémentaire associée à des intervalles horaires pendant lesquelles les pics sont observés durant chaque saison est ajoutée aux modèles de la deuxième catégorie. Les modèles sont estimés en régression linéaire multiple et testés avec les données horaires de consommation d'électricité, des prévisions de l'opérateur et de températures observées en Iran dans la période de 1991 à 1997. Deux critères d'appréciation sont retenus pour évaluer l'efficacité des modèles : l'erreur absolue moyenne en pourcentage et l'erreur absolue maximum en pourcentage. En comparant les modèles estimés au modèle ARIMA régulier, il ressort que les modèles ARIMA modifiés produisent une meilleure prévision que le modèle ARIMA régulier à cause de la variable, « prévision de l'opérateur », incluse dans les modèles modifiés qui est fortement corrélée avec la variable dépendante estimée.

Kagan, Goktas et Hepsag (2009) ont comparé trois modèles : le modèle ARIMA ajusté de la saisonnalité, le modèle SARIMA et un modèle de régression avec variables binaires séquentielles pour la prévision de la consommation horaires d'électricité en Turquie. Pour définir le modèle ARIMA, ils ont évalué les effets saisonniers en faisant une régression de la demande d'électricité sur des variables binaires associées à la périodicité des effets saisonniers. Ils ont ensuite extrait la saisonnalité par la méthode des moyennes mobiles multiplicatives et la série obtenue a été rendue stationnaire en appliquant une différenciation régulière d'ordre un. Les fonctions d'autocorrélation et d'autocorrélation partielle ont servi à identifier respectivement l'ordre des processus MA et AR. Le modèle SARIMA est utilisé pour tenir compte de la saisonnalité qu'ils ont identifiée par régression linéaire. Le modèle de régression avec variables binaires séquentielles est proposé comme une méthode alternative à la méthode Box-Jenkins. Les chocs structurels sont identifiés dans ce modèle par les tests CUSUM et CUSUM-SQ sur les résidus de la régression et définis par les variables binaires dans le modèle de régression. Les trois modèles sont testés sur les données horaires de la Turquie sur la période de janvier 1997 à décembre 2005. Les résultats de cette étude montrent que la méthode de régression avec variables saisonnières latentes (binaires) est une méthode alternative à la méthode Box-Jenkins (ARIMA ou SARIMA) en cas de forte saisonnalité et de rupture structurelle dans la variable étudiée. Elle fournit une meilleure prédiction avec l'erreur absolue moyenne en pourcentage la plus faible.

Soares et Medeiros (2008) étendent les travaux de Cottet et Smith (2003) et Soares et Souza (2006) et s'intéressent à la prévision de la consommation d'électricité des régions du Sud du Brésil durant la période allant du 1^{er} Janvier 1990 au 31 décembre 2000. Leur modèle autorégressif, appelé «*modèle autorégressif saisonnier à deux étapes*», décompose les séries en (i) une partie déterministe qui capture les effets de long terme, les saisonnalités annuelles et les événements particuliers à l'aide de tendances (*linéaires et non linéaires*) et de variables muettes et (ii) une composante *stochastique* dans laquelle le terme d'erreur suit un processus autorégressif. En appliquant leur modèle sur les données par tranche horaire (la charge horaire, pour chaque tranche horaire étant traitée comme une série chronologique séparée), ils comparent sa performance prédictive à celle d'un modèle ARIMA saisonnier. Leurs résultats indiquent que la subdivision des données en plages horaires améliore significativement la performance prédictive.

Yang, Huang et Huang (1995) utilisent comme seule variable explicative exogène, la température, pour estimer un modèle ARMAX. Leur méthodologie est basée sur ce qu'ils ont qualifié d'approche de «*programmation évolutionnaire*¹³ de prévision de charge à court terme». Dans cette approche ils définissent des ensembles «*parents*» et «*fil*s » dont les éléments sont différents ordres du modèle ARMA ou les paramètres estimés de ce modèle. Ils ont sélectionné les éléments de chaque ensemble par un processus de compétition aléatoire en fonction des valeurs qui procurent un meilleur modèle. Ils ont testé leur modèle avec les données horaires d'électricité et de température à Taiwan, regroupées en deux catégories : données de jour de semaine et données de Weekend. Le critère d'erreur de prédiction finale d'Akaike est adopté pour identifier la valeur appropriée de l'ordre du modèle et une fonction de perte est utilisée pour identifier les paramètres appropriés. La capacité prédictive de leur modèle ARMAX s'est avérée supérieure à celle proposée par le logiciel commercial SAS pour la prévision d'un jour à une semaine, justifiant la possibilité d'améliorer les capacités des modèles ARMAX par l'algorithme de «*programmation évolutionnaire*». La particularité de leur étude est que leur algorithme est principalement basé sur un jeu d'optimisation permettant de sélectionner le meilleur ordre du modèle ARMA et les meilleurs coefficients des variables du modèle ARMAX.

Espinoza et al. (2007) construisent un modèle pour identifier la relation non linéaire entre la charge d'électricité et les variables exogènes susceptibles en Belgique. Leur modèle non linéaire autorégressif avec variables exogènes (NARX) est constitué de trois ensembles de variables explicatives : le premier ensemble est la partie autorégressive contenant les données de consommation horaire d'électricité de deux jours précédents; le second est constitué des variables liées à la température mesurant les effets de l'utilisation des équipements de chauffage et de refroidissement et le troisième ensemble regroupe les variables d'information calendaire sous forme de variables binaires sur le mois de l'année, le jour de la semaine et l'heure de la journée. Ces variables exogènes influencent la variable dépendante à travers une fonction inconnue qu'ils définissent à l'aide d'une technique de moindres carrés¹⁴ sous l'hypothèse de non linéarité. Le modèle est testé sur les données horaires de consommation d'électricité et de température. L'erreur absolue moyenne, s'est située en dessus de 3% dans les différents échantillons. Ce qui a permis de

¹³ Pour une meilleure compréhension de cette méthodologie, voir, Fogel D.B (1994).

¹⁴ Voir Espinoza, M., J. A. K. Suykens, et al. (2006).

conclure que cette méthodologie permet d'obtenir de meilleurs résultats de prévision en comparaison à la méthode des moindres carrés ordinaires.

2. Méthodologie

Dans cette partie, nous présentons de manière plus détaillée les modèles retenus dans notre étude ainsi que la démarche adoptée pour les estimer. Le premier modèle est un modèle de lissage exponentiel avec erreurs multiplicative/additive, tendance additive et saisonnalité additive, ETS (M,A,A)/ETS(A,A,A)¹⁵. Le deuxième modèle est un modèle ARIMA saisonnier (SARIMA) sans variables exogènes. Le dernier modèle retenu est un ARIMA saisonnier avec variables exogènes (climatiques). Ces modèles sont les plus couramment utilisés dans la littérature sur le sujet.

2.1 Le lissage exponentiel de Holt-Winters

Le lissage exponentiel de Holt-Winters, aussi appelé lissage exponentiel triple, décompose la variable d'intérêt en deux composants, une tendance (T) et une saisonnalité (S), auxquels on peut ajouter un terme d'erreur (E). Il existe deux grandes versions de ce modèle : une version additive qui est indiquée lorsque la variance est stable dans le temps, et une version multiplicative plus adéquate lorsque la variance croît/décroît dans le temps. Dans ce travail, nous privilégierons la version additive, car nos données corroborent plutôt l'hypothèse d'une variance de la consommation horaire d'électricité relativement stable dans le temps (voir la section 3). Néanmoins, nous étudions également différentes versions de ces modèles de lissage, selon la méthodologie proposée par Hyndman, Koehler, Ord et Snyder (2008,11 :27)¹⁶. Nous comparons ces modèles en termes de critère d'information et de performance prédictive.

Hyndman, Koehler, Ord et Snyder (2008) proposent une méthodologie d'estimation des modèles de lissage exponentiel basé sur l'approche *espace-état*, qu'ils mettent en pratique dans la fonction **ets()** de la librairie *forecast* du logiciel R. Les paramètres sont estimés par maximum de vraisemblance. Ces auteurs précisent que, dans le cas des modèles additifs de type Holt-Winters avec une tendance à deux composantes (le niveau (*level*) et la pente (*growth*)), l'estimation pour h périodes futures est donnée par l'équation suivante :

¹⁵ ETS(MAA) signifie "Exponential Smoothing with Multiplicative errors, Additive trend and Additive seasonality"

¹⁶ Forecasting with Exponential Smoothing: The State Space Approach

$$T_h = l_t + bh \quad [1]$$

Où l est le *niveau* et b est la *pente*.

Cette tendance additive est dans certain cas *amortie* (notée *Ad* pour « Additive damped trend ». Gardner et McKenzie (1985)¹⁷ ont proposé une modification de la méthode linéaire de Holt pour permettre de prendre en compte cette version amortie. Les équations de la tendance deviennent alors :

$$\text{Niveau : } l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + \phi b_{t-1}) \quad [2.a]$$

$$\text{Pente : } b_t = \beta^* (l_t - l_{t-1}) + (1 - \beta^*) \phi b_{t-1} \quad [2.b]$$

$$\text{Prévision : } \hat{y}_{t+h|t} = l_t + ((\phi + \phi^2 + \dots + \phi^h) b_t) \quad [2.c]$$

$0 < \phi \leq 1$ montre bien l'aspect amorti de l'estimation : plus la prévision est éloignée, plus l'effet de la pente décroît. Par exemple, la pente pour la prévision en $t+1$, $t+2$, etc., sont b_t , $(\phi + \phi^2)b_t, \dots$, etc. Une valeur de ϕ proche de 1 donne plus de poids au présent immédiat. D'après Hyndman et al. (op.cit), la saisonnalité peut être incluse dans l'expression antérieure et s'exprimer en termes de *niveau* et de *pente* de manière additive :

$$\text{Niveau : } l_t = \alpha (z_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad [3.a]$$

$$\text{Pente : } b_t = \beta (l_t - l_{t-1}) + (1 - \beta)b_{t-1} \quad [3.b]$$

$$\text{Saisonnalité : } s_t = \gamma (z_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m} \quad [3.c]$$

$$\text{La prévision : } \hat{z}_{t+h|t} = l_t + b_t h + s_{t-m+h_m} \quad [3.d]$$

$$\text{avec } h_m = [(h-1) \bmod m] + 1$$

Les équations [3.a] à [3.d] peuvent être réécrites en fonction de l'erreur de prévision, selon que cette erreur, notée ε_t , est multiplicative ou additive. On obtient alors les formulations suivantes:

1. Modèle avec erreurs additive, tendance additive et saisonnalité additive (AAA) :

¹⁷ Gardner, E. S., et McKenzie, E. D. (1985). Forecasting trends in time series. Management Science, 31(10), 1237-1246.

$$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$$

[4.a]

$$l_t = l_{t-1} + b_{t-1} + \alpha \varepsilon_t \quad [4.b]$$

$$b_t = b_{t-1} + \beta \varepsilon_t \quad [4.c]$$

$$s_t = s_{t-m} + \gamma \varepsilon_t \quad [4.d]$$

2. Modèle avec erreurs additive, tendance *additive amortie* et saisonnalité additive (AA_dA) :

$$\mu_t = l_{t-1} + \phi b_{t-1} + s_{t-m} \quad [5.a]$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t \quad [5.b]$$

$$b_t = \phi b_{t-1} + \beta \varepsilon_t \quad [5.c]$$

$$s_t = s_{t-m} + \gamma \varepsilon_t \quad [5.d]$$

3. Modèle avec erreurs multiplicative, tendance additive et saisonnalité additive (MAA) :

$$\mu_t = l_{t-1} + b_{t-1} + s_{t-m} \quad [6.a]$$

$$l_t = l_{t-1} + b_{t-1} + \alpha(l_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t \quad [6.b]$$

$$b_t = b_{t-1} + \beta(l_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t \quad [6.c]$$

$$s_t = s_{t-m} + \gamma(l_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t \quad [6.d]$$

4. Modèle avec erreurs multiplicative, tendance additive amortie et saisonnalité additive (MA_dA) :

$$\mu_t = l_{t-1} + \phi b_{t-1} + s_{t-m} \quad [7.a]$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha(l_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t \quad [7.b]$$

$$b_t = \phi b_{t-1} + \beta(l_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t \quad [7.c]$$

$$s_t = s_{t-m} + \gamma(l_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t \quad [7.d]$$

Le terme ε_t est supposé indépendant et identiquement distribué et suit une distribution gaussienne de moyenne 0 et de variance constante.

2.2 Le modèle ARIMA saisonnier (SARIMA)

Le deuxième modèle estimé dans ce travail est le modèle ARIMA saisonnier. Rappelons les hypothèses implicites à ce modèle, énoncées en section 1.1.1 : (i) existence d'une corrélation linéaire entre les observations présentes et passées de la variable d'intérêt; (ii) existence d'un effet linéaire aléatoire présent et passé sur la variable d'intérêt et (iii) absence d'effet significatif d'autres variables exogènes. Dans sa forme la plus standard, voir Aragon (2011;103), ce modèle s'écrit :

$$Z_t = \beta + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + \varepsilon_t, \quad \phi_p \neq 0, \quad \varepsilon_t \sim BB(0, \sigma_v^2) \quad [8]$$

Z_t est la consommation d'électricité à l'instant t

ϕ_i est un paramètre d'autorégression correspondant au retard i tel que $i = 1, \dots, p$

ε_t est un choc aléatoire de la série chronologique et est un Bruit blanc de moyenne nulle et de variance constante.

En intégrant l'opérateur de retard l'équation [8] devient :

$$\Theta(B)Z_t = \beta + v_t \quad [9]$$

tel que $\Theta(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \quad [10]$

B^i est l'opérateur de retard pour le retard i . Z_t est un processus $AR(p)$ stationnaire ; les racines du polynôme $\Theta(B) = 0$ n'appartiennent pas à l'intervalle $[-1, 1]$.

Sous l'hypothèse (iii), nous considérons que une fraction de la consommation d'électricité suit une marche aléatoire et qu'une observation peut aussi être définie par :

$$Z_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad [11]$$

θ_i est un paramètre de la composante en moyenne mobile du retard i tel que $i = 1, \dots, q$,

En introduisant l'opérateur de moyenne mobile, nous obtenons :

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q, \quad [12]$$

$$\text{Tel que l'équation, } Z_t = \mu + \Theta(B)\varepsilon_t \quad [13]$$

n'est pas un bruit blanc (voir la note de bas de page, numéro 8). Les racines de l'équation [12] sont strictement supérieures à 1 en valeur absolue. En incorporant les effets saisonniers aux équations [8] et [13] à travers les paramètres associés à la saisonnalité (P, D, Q), nous obtenons le modèle SARIMA défini par :

$$\Theta_P(B^s)\phi_P(B)(1-B^s)^D(1-B)^d Z_t = \Phi_Q(B^s)\theta_Q(B)\varepsilon_t \quad [14]$$

Sous l'hypothèse (iii), nous ignorons l'effet de facteurs exogènes sur la consommation d'électricité. Le modèle est ainsi basé sur les hypothèses de non stationnarité, d'autocorrélation des valeurs de la série, de l'existence d'une composante aléatoire et de la saisonnalité. Les valeurs de p, d, q, P, D, Q et s que nous obtiendrons du modèle estimé nous permettront de rejeter ou de ne pas rejeter ces hypothèses pour la série chronologique de consommation d'électricité dans la région de Québec. Ce modèle est identifié selon la méthodologie proposée par Hyndman et Khandakar (2008). Cette procédure est implémentée dans la fonction **auto.arima()** du package forecast du logiciel R.

2.3 Le modèle ARIMA saisonnier avec variable exogènes

Le modèle ARIMA saisonnier précédent ignorait l'impact des variables exogènes, comme les variables climatiques. Or, l'influence de ces variables peut permettre de capturer les spécificités climatiques ignorées par le passé de la consommation d'électricité. L'estimation de ce modèle suit une procédure de minimisation du critère AIC pour déterminer les ordres p, q et P, Q. Nous imposons d=D=0 et une forme fonctionnelle linéaire pour $f(X_t)$:

$$\Theta_P(B^s)\phi_P(B)(1-B^s)^D(1-B)^d Z_t = \Phi_Q(B^s)\theta_Q(B)\varepsilon_t + f(X_t) \quad [15]$$

X_t est la matrice des variables climatiques (température, humidité relative, point de rosée). La fonction f est estimée par maximum de vraisemblance dans la fonction **auto.arima()**.

2.4 Procédure d'estimation des modèles SARIMA

L'estimation des modèles SARIMA requiert dans un premier temps de tester la stationnarité de la série et de procéder au besoin à la transformation des séries pour obtenir une série transformée stationnaire. On adopte ensuite une méthodologie d'identification des ordres autorégressifs des processus AR et MA. Le modèle estimé est ensuite testé pour vérifier qu'il obéit à toutes les hypothèses statistiques fondamentales de la modélisation dynamique présentée dans la revue méthodologique 1.1 et en section 2.2. Finalement, la qualité des prévisions hors échantillon est analysée. Cette procédure globale est souvent appelée « procédure Box-Jenkins (1976) ».

2.4.1 Test de racine unitaire

Les tests de stationnarité des séries sont des classiques de l'économétrie des séries temporelles. Dans les années 90, plusieurs tests de stationnarité saisonnière ont été introduits¹⁸. Ici, nous nous contentons d'appliquer le test statistique le plus populaire pour tester la stationnarité de la série, sans nous préoccuper des cycles dans un premier temps. Le test de stationnarité de la série de consommation d'électricité que nous adoptons est la version « augmentée » de Dickey et Fuller (1981), dont la stratégie d'implémentation est décrite de manière très pratique par Pfaff (2008 :61). Cette stratégie se base sur les équations [16] à [18], selon que l'on considère la présence ou l'absence d'une tendance déterministe (appelée « trend » en anglais) et d'un terme de dérive (la constante aussi appelée « drift » en anglais) :

$$\Delta Z_t = \pi Z_{t-1} + \sum_{j=1}^k \gamma_j \Delta Z_{t-j} + \mu_t \quad [16]$$

$$\Delta Z_t = \beta_1 + \pi Z_{t-1} + \sum_{j=1}^k \gamma_j \Delta Z_{t-j} + \mu_t \quad [17]$$

$$\Delta Z_t = \beta_1 + \beta_2 t + \pi Z_{t-1} + \sum_{j=1}^k \gamma_j \Delta Z_{t-j} + \mu_t \quad [18]$$

¹⁸ Voir notamment Osborn-Chui-Smith-Birchenhall (1988), Franses (1990), Hylleberg et al. (1990) et Canova-Hansen (1995).

Nous avons effectué les tests de stationnarité sur la base des équations [16], [17] et [18], appelées respectivement « None », « Drift » et « Trend » dans notre tableau de résultats, pour un nombre maximum de retards de $k = 168$ (1 semaine). Le paramètre k est choisi par minimisation du AIC. Notons que le terme qui englobe la somme à droite des équations [16] à [18] n'intervient pas directement dans l'interprétation des tests de racine unitaire. Il n'est là que pour « blanchir » l'erreur, en déterminant k par minimisation du critère d'information AIC par exemple. L'estimation et le calcul des différentes statistiques de test sont disponibles dans la fonction **ur.df ()** du *package urca* du logiciel R. L'équation [16] permet de tester la présence d'une racine unitaire sans dérive sous l'hypothèse nulle (statistique τ_1). L'équation [17] permet de tester la présence d'une racine unitaire sous l'hypothèse nulle (statistique τ_2), et l'hypothèse jointe de présence de racine unitaire et d'une constante nulle (statistique ϕ_1). L'équation [18] teste l'hypothèse nulle de présence de la racine unitaire (statistique τ_3), l'hypothèse jointe de « tendance déterministe nulle et de présence de racine unitaire » (statistique ϕ_3) ainsi que l'hypothèse jointe de « constante nulle, tendance déterministe nulle et de présence de racine unitaire » (statistique ϕ_2). Les valeurs critiques sont fournies automatiquement par le logiciel, conformément aux tables statistiques de Fuller (1996). Finalement, sans entrer dans trop de détails, la stationnarité de la composante saisonnière est également vérifiée avec le test de Canova et Hansen (1995). La procédure est implémentée dans la fonction **nsdiffs()** de la librairie *forecast*. Si ni la série, ni la saisonnalité de la série ne possèdent de racines unitaires alors les paramètres $d = D = 0$. En cas de présence de racines unitaires, la série et/ou sa composante saisonnière sont différenciées le nombre de fois requis.

2.4.2 Saisonnalité et identification des modèles ARIMA saisonnier

Une fois la stationnarité des séries testée et les ordres d et D déterminés, la construction des modèles SARIMA et SARIMAX nécessite l'identification des ordres (p,q) et (P,Q) . Nous adoptons l'approche d'optimisation des coefficients de la régression par *maximum de vraisemblance*, et cherchons à minimiser le critère d'information d'Akaike (1973), AIC, pour déterminer les ordres optimaux. Ce critère est défini par :

$$AIC = T \ln(\sigma_\varepsilon^2) + 2k \quad [19]$$

où $\hat{\sigma}_\varepsilon^2 = \frac{1}{T} \sum \hat{\varepsilon}_t^2$ est l'estimation maximum de vraisemblance (MV) de σ_ε^2 ; k est le nombre de paramètres estimés et T est le nombre d'observations (Aragon, 2011 :44). Cette procédure est implémentée dans la fonction **auto.arima()** du logiciel R. Notons que dans notre cas, le grand nombre d'observations à disposition rend le modèle particulièrement lourd à estimer, et cela d'autant plus que la fréquence horaire d'observation nous incite à autoriser un nombre de *lag* maximum p, q et P, Q d'une semaine, soit 168 heures. C'est la raison pour laquelle une procédure d'estimation rapide est appliquée (voir les arguments *stepwise* et *approximation* de la fonction **auto.arima()**). Ceci requiert néanmoins une demi-journée de calcul sur un ordinateur de bureau actuel.

Une fois estimés les paramètres des modèles, nous testons l'hypothèse que les résidus sont des bruits blancs (voir la note de bas de page, numéro 8) issus d'une même distribution gaussienne. Nous utilisons le test de Jarque et Bera¹⁹ (1980) pour tester la normalité des résidus et la statistique de *Ljung-Box* (1978) définie par l'équation suivante :

$$Q^*(h) = n(n+2) \sum_{k=1}^h \frac{\rho_k^2}{n-k}. \quad [20]$$

La statistique [20] suit une distribution de probabilité Khi-deux (χ^2) dont le nombre de degré de liberté est égale à h, où h est le nombre de coefficient d'autocorrélation testé.

2.4.3 Prévisions et mesures de performance prédictives

Finalement, pour mesurer et comparer les erreurs de prévision de chaque modèle, nous avons retenu le critère de l'erreur moyen absolue (MAPE), exprimé en pourcentage :

$$MAPE = 100 * \frac{1}{N} * \sum_{t=1}^N \left| \frac{Z_t - \hat{Z}_t}{Z_t} \right| \quad [21]$$

où Z_t est l'observation actuelle et $\hat{Z}_t$, l'observation de la prévision. Nous retenons le modèle pour lequel ce critère est minimum. Mais d'autres indicateurs usuels sont également fournis par le logiciel.

¹⁹ Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), 255-259.

3. Données

Les données utilisées pour cette étude sont des observations horaires sur la consommation d'électricité domestiques et agricoles, la température, le point de rosé, la nébulosité et l'humidité relative des régions de Québec et Montréal (province du Québec). Toutes ces données sont celles enregistrées à chaque heure de la journée sur la période de 2006 à 2010 par Hydro Québec, qui est la plus grande compagnie de production et de distribution d'électricité au Québec. La consommation d'électricité est mesurée en mégawatts (MW). À titre indicatif, on retiendra que 1 MW délivre une énergie équivalente à 0.086 tonnes de pétrole (86 kg). La température et le point de rosée sont définis en degré Celsius (°C). Le point de rosée étant la température à laquelle l'humidité de l'air se condense pour former les gouttelettes d'eau. L'humidité relative est donnée en pourcentage (%) et peut être déduite de la température ambiante et de la température de point de rosée. La nébulosité est exprimée en pourcentage et correspond à l'intensité de la rosée ou au niveau de couverture du ciel par les nuages. Nous avons retenu les variables climatiques mesurées à Montréal et à Québec.

3.1 Statistiques descriptives agrégées

La figure 1 montre la relation entre la consommation d'électricité et la température : la consommation d'électricité décroît avec la hausse de la température jusqu'à atteindre un point de retournement situé vers 15°C. Cette allure se justifie par l'utilisation des systèmes de chauffage en périodes froides et des systèmes de refroidissement en périodes chaudes. La corrélation positive entre la température et le point de rosé est très forte. De ce fait, la relation entre le point de rosé, la température et la consommation d'électricité est similaire. Il existe un point d'humidité relative maximum auquel la consommation d'électricité cesse de croître et diminue progressivement avec son accroissement. L'effet de la variation de la nébulosité sur la demande d'électricité est presque indéfini ou indirect. L'impact des conditions climatiques paraît ainsi important non seulement dans le cycle journalier mais aussi sur la quantité totale annuelle d'électricité consommée. Le test de corrélation linéaire de Pearson montre une forte corrélation entre la consommation d'électricité et la température (-85%) et le point de rosée (-84.7%). Le point de rosée étant fortement corrélé avec la température. La corrélation de la consommation est plus faible avec l'humidité relative (-4.5%) et la nébulosité (9.2%) et c'est la raison pour laquelle nous privilégions dans cette analyse le lien entre la consommation d'électricité et la température. Les corrélations avec les autres variables sont analysées en

profondeur dans les différents modèles que nous utilisons dans cette étude. La figure 1 permet aussi d'identifier la relation non linéaire entre les variables climatiques (température et point de rosé). On constate que la relation entre la consommation d'électricité et la nébulosité est diffuse.

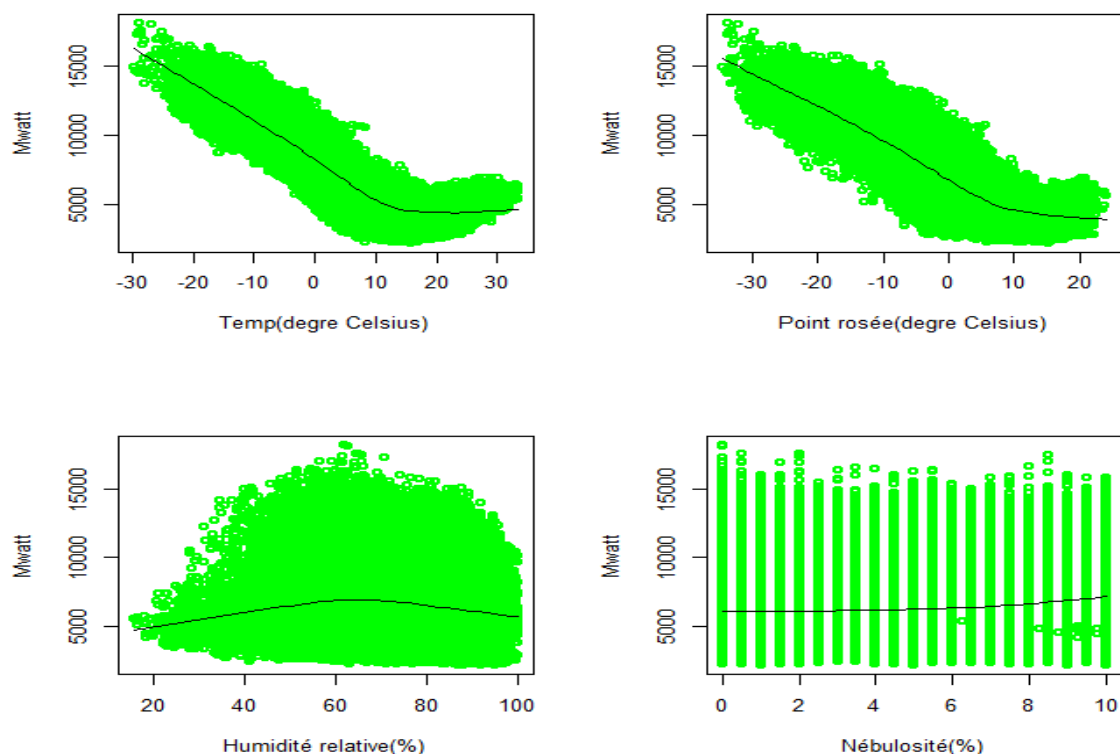


Figure 1 : Relation entre les variables climatiques et la consommation horaire d'électricité.

Les tableaux 1 et 2 nous montrent quelques statistiques de l'évolution de la consommation d'électricité et de la température. La consommation moyenne annuelle domestique et agricole d'électricité s'est accrue de 11.24% entre 2006 et 2009 et a chuté de 4.7% entre 2009 et 2010. Ce qui représente une croissance effective de 6.54% de 2006 à 2010. La consommation moyenne annuelle s'est accrue de 2006 à 2009, passant de 6367 MW à 7083 MW. La température moyenne a évolué en sens inverse sur la même période passant de 7.46°C à 6°C. L'année 2010 a été un peu plus chaude que les quatre années précédentes : la température moyenne est remontée de 5.96°C à 7.86°C et la consommation moyenne d'électricité est descendue à 6752 MW. La dispersion autour de la moyenne, pour les deux variables, s'est aussi accrue durant cette période et les consommations minimales et maximales annuelles ont aussi beaucoup variées à la hausse. On

remarque une forte variabilité de la température et de la demande d'électricité entre les périodes froides (décembre à mars) et les périodes chaudes (Juin à Aout). La stabilité de la température autour des valeurs extrêmes (maximales et minimales) durant ces périodes explique aussi les niveaux de consommation minimum et maximum annuels. On remarque une valeur extrême en 2009 de 18120 MW (Tableau 1) qui représente un pic de consommation identifié le 16 janvier 2009 à 7heures du matin pour une température moyenne de -28.65°C très proche du minimum observé de cette même année (-29.75°C).

Tableau 1 : Statistiques récapitulatives de la consommation d'électricité sur cinq (5) années au Québec.

Année	Minimum	1^{er} Quartile	Médiane	Moyenne	3^{ème} Quartile	Maximum	Écart-type
2006	2157	4285	5574	6367	8358	14910	2716.36
2007	2287	4304	5736	6784	9230	16070	3127.006
2008	2208	4455	5990	6882	9356	15770	3040.708
2009	2403	4533	6016	7083	9428	18120	3283.68
2010	2425	4537	5855	6752	8812	16330	2911.777

Tableau 2 : Statistiques récapitulatives de la température sur cinq (5) années au Québec.

Année	Minimum	1^{er} Quartile	Médiane	Moyenne	3^{ème} Quartile	Maximum	Écart-type
2006	-20.100	-0.200	7.500	7.459	16.150	32.000	10.86551
2007	-24.850	-3.100	6.800	6.115	16.600	32.100	12.49541
2008	-24.350	-3.350	7.250	6.174	16.450	31.150	11.76655
2009	-29.750	-2.050	6.950	5.961	15.700	31.450	11.86332
2010	-21.100	-0.500	7.600	7.863	16.900	33.400	11.03362

3.2 Profils chronologiques

La figure 2 illustre le profil de **consommation horaire d'électricité au Québec pour l'ensemble de la période d'analyse**. On ne remarque pas de tendance croissante. La variance ne semble pas toujours constante autour de la moyenne saisonnière à court terme mais elle reste relativement stable sur des périodes plus longues. Le cycle annuel est aussi clairement répétitif. Ce graphique confirme la forte hausse de la consommation en automne et hiver et la baisse pour le printemps et l'été.

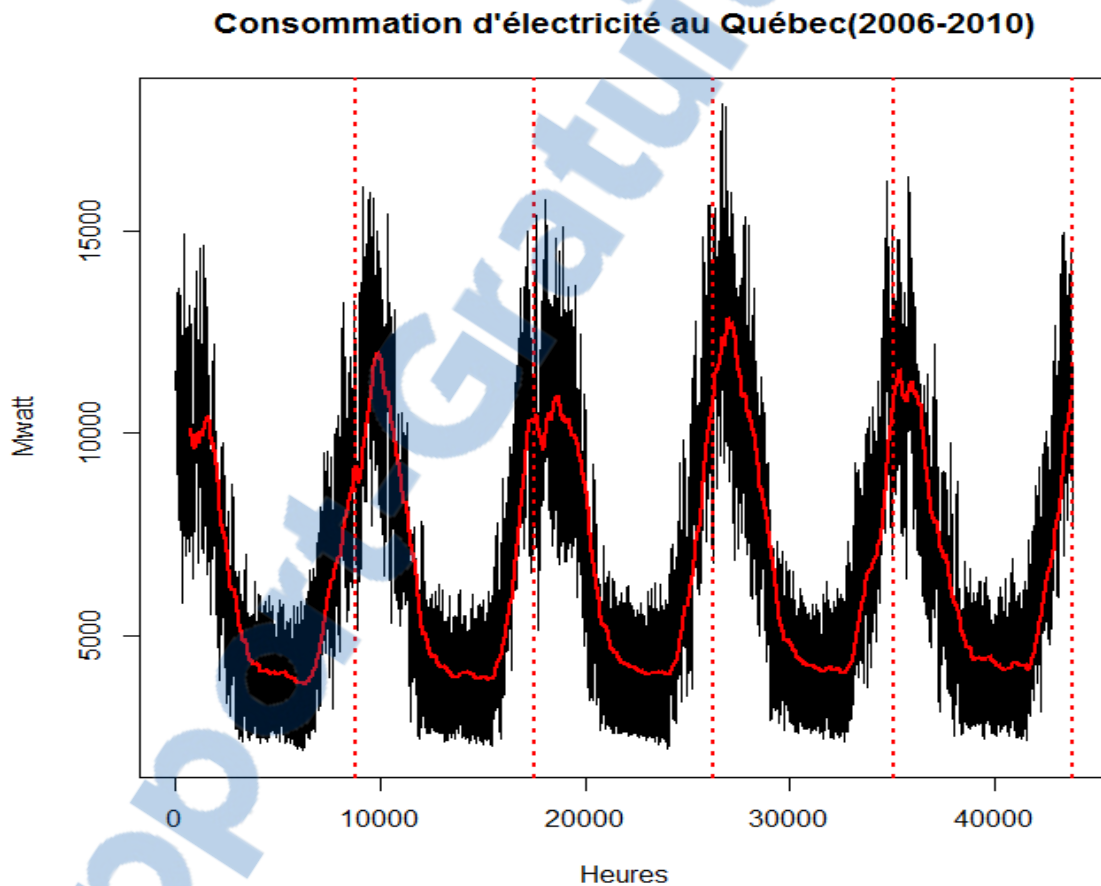


Figure 2 : Profil de consommation d'électricité au Québec de 2006 à 2010 (43824heures). Les lignes rouges verticales en pointillée délimitent le début des différentes années.

La figure 3 présente des profils de consommation hebdomadaire de la première semaine de chaque saison pour les années 2006 et 2010 par rapport à leur moyenne annuelle.

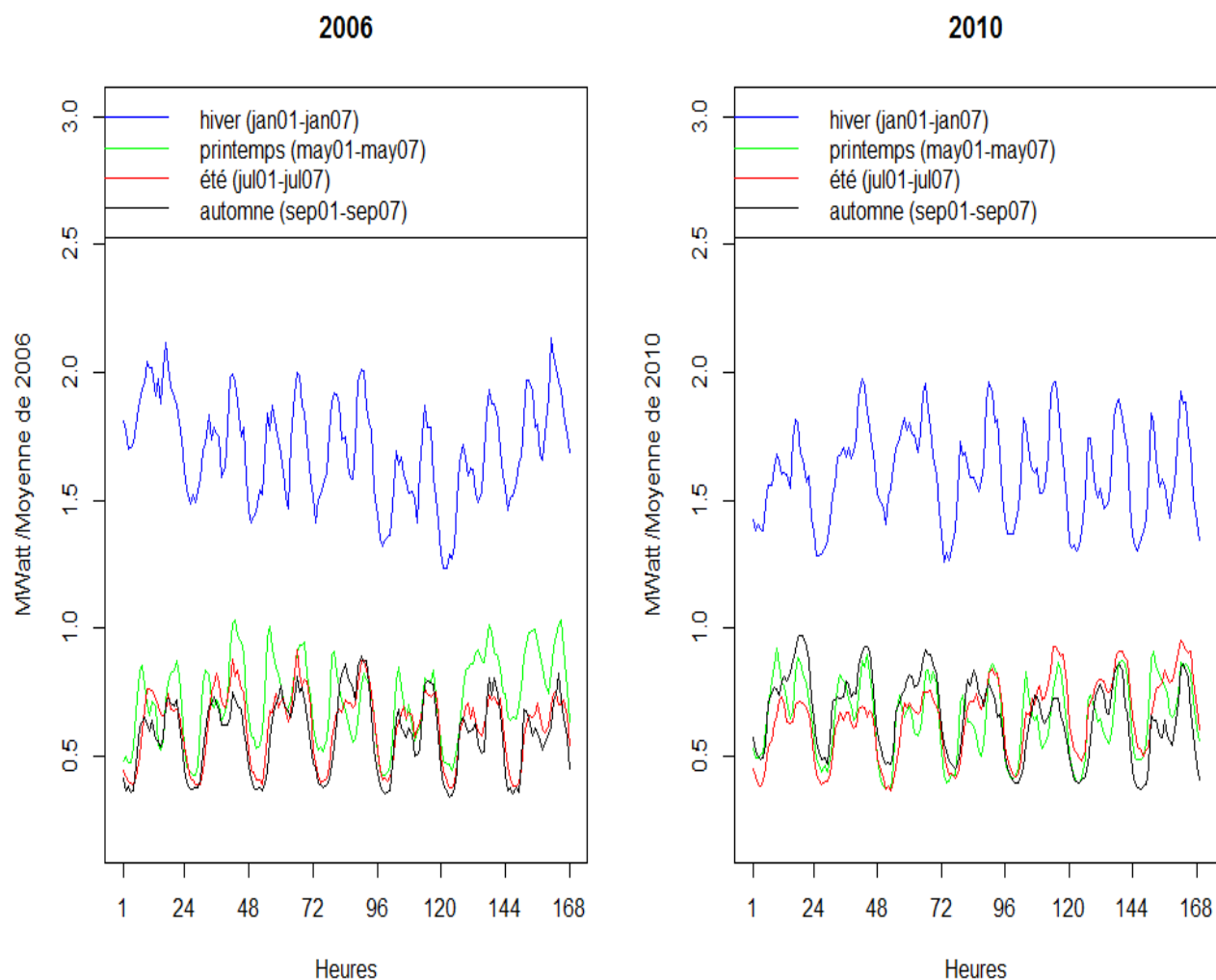


Figure 3 : Profil de consommation d'une semaine type de chaque saison en 2006 et 2010.

Les figures 4 et 5 se concentrent sur les profils de consommation aux différents moments de la journée. La figure 4 se concentre sur les cycles à des heures spécifiques de la journée (6am, 12pm, 6pm et 12am) pour l'ensemble des 5 années (~ 7 jours x 4 semaines x 12 mois x 5 années = 1680 heures). La ligne rouge est la moyenne mobile sur 30 jours. On constate que les profils de consommation aux différentes heures de la journée se ressemblent, avec des pics en hiver et une consommation plus basse en été. On observe cependant des différences importantes dans les niveaux de consommation selon les heures : la consommation moyenne est nettement inférieure à minuit et particulièrement élevée à 6pm, ce qui est conforme à ce que l'on attend.

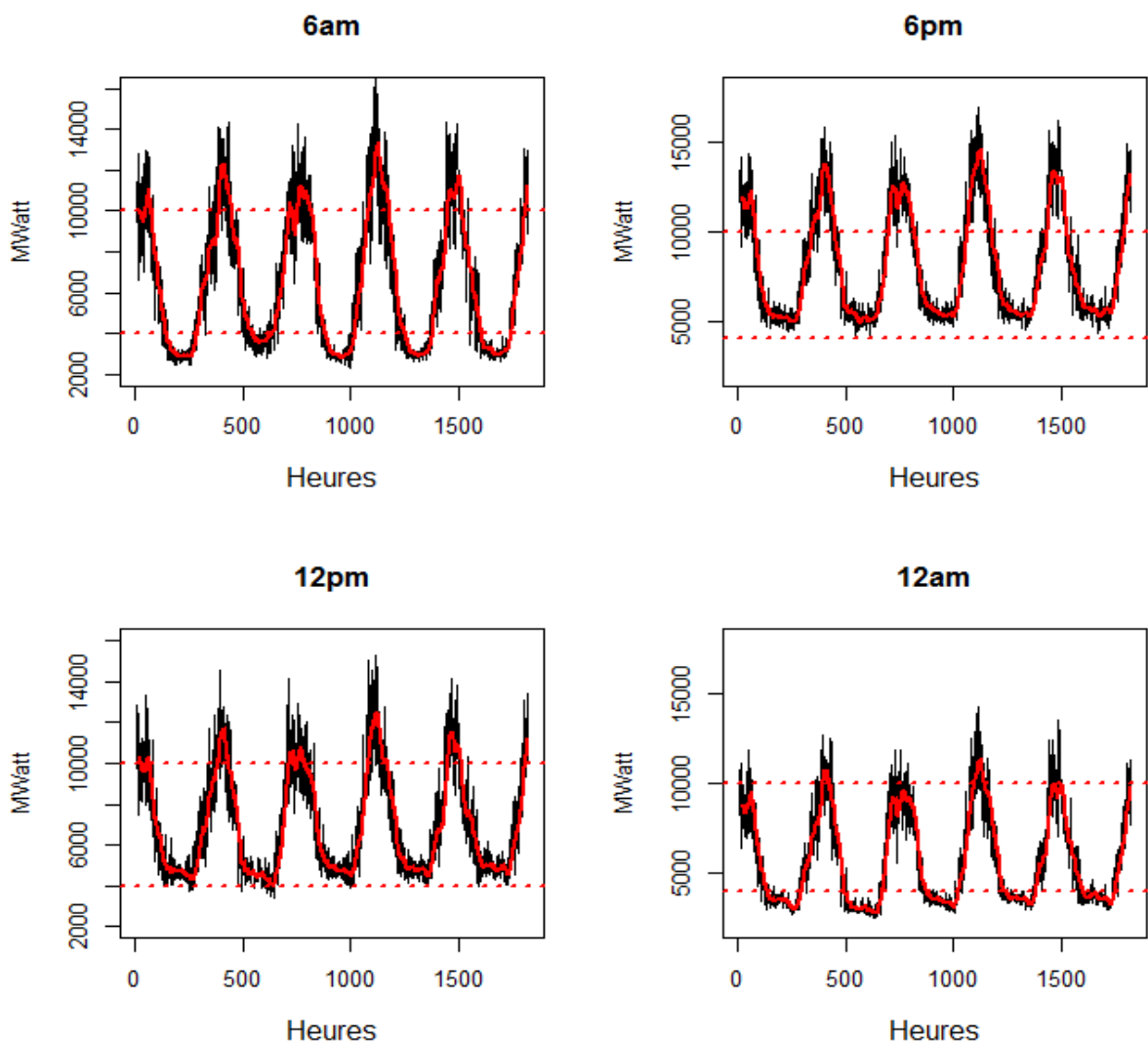


Figure 4 : Profil annuel de la consommation d'électricité à des heures spécifiques. Les lignes rouges horizontales en pointillée permettent une meilleure comparaison des niveaux de consommation d'électricité entre différentes plages horaires.

La figure 5 montre que les niveaux de consommation aux différentes heures de la journée sont restés relativement stables aux cours des années analysés.



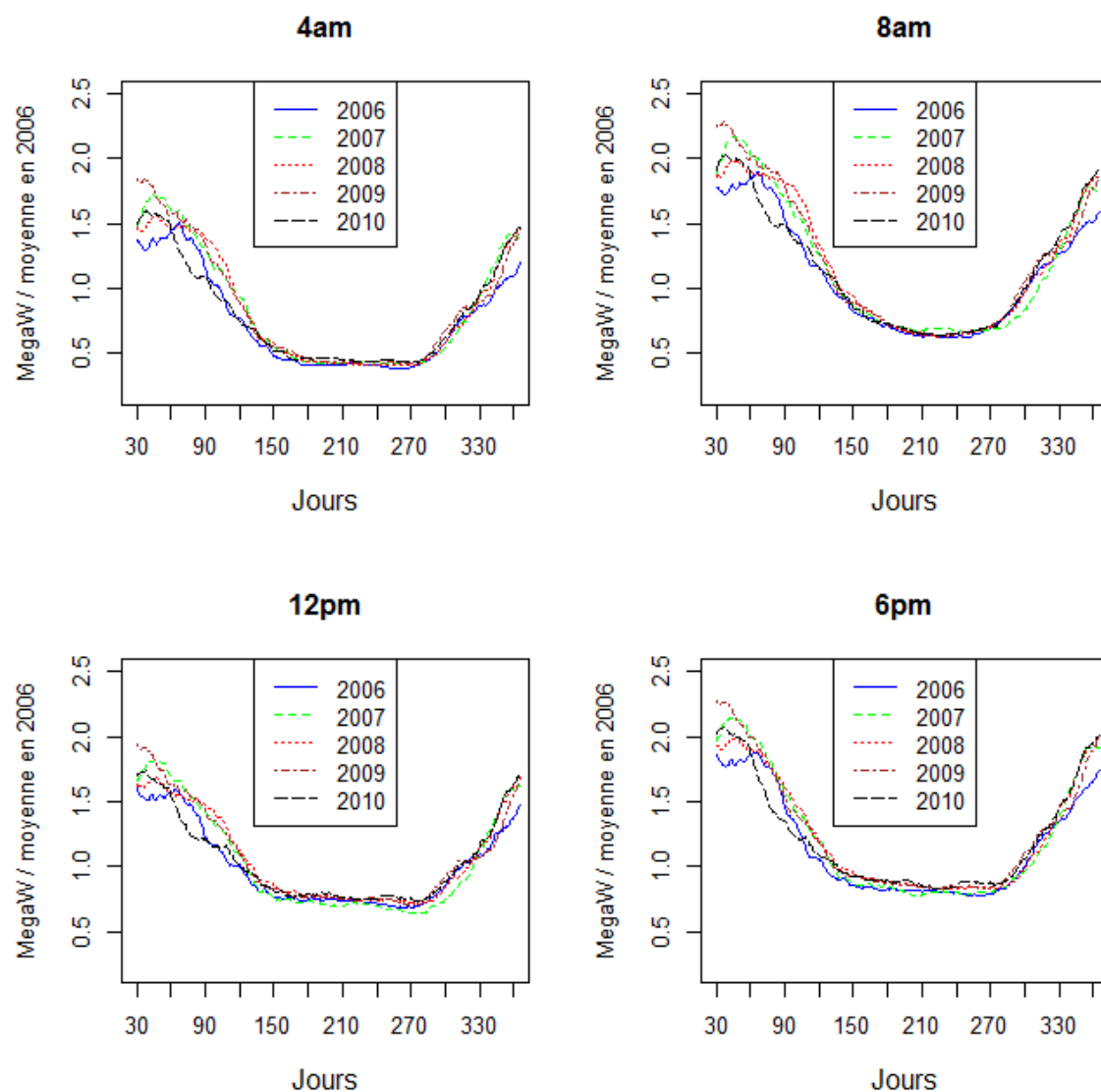


Figure 5 : Profil annuel de consommation d'électricité en moyenne mobile 30 jours à des heures spécifiques.

Une technique intéressante qui permet de détecter les cyclicités/fréquences les plus pertinentes d'une chronique est l'analyse discrète de Fourier ou analyse spectrale à travers le périodogramme (Weron, 2006;40-41). Le but de cette approche est de décomposer la série en vagues sur la base de fonctions sinusoïdales. Les vagues sont exprimées en termes de fréquences qui représentent le nombre de cycles par unité de temps. Selon la formulation de Weron (2006;39) le périodogramme (ou *densité spectrale*) est définie par :

$$I_n(w_k) = \frac{1}{n} \left| \sum_{t=1}^n x_t \exp^{-i(t-1)w_t} \right|^2 \quad [29]$$

$w_k = 2\pi(k/n)$, sont des fréquences exprimées en termes de radians par unité de temps.

Soit w , la fréquence et T la période de temps requise pour un cycle complet, alors T est l'inverse de la fréquence ($T=1/w$). La figure 6 permet d'identifier la périodicité de la saisonnalité de la consommation d'électricité au Québec. On remarque dans l'image du bas que la densité spectrale est plus forte autour de zéro. Cela provient du cycle hebdomadaire de 168 heures ($T=1/168=0.005952380$). On remarque deux autres pics à 24 heures ($T=1/24=0.0416$) et à 12heures ($T=1/12=0.0832$). Les cycles mensuels et saisonniers sont peu perceptibles sur nos données à haute fréquence.

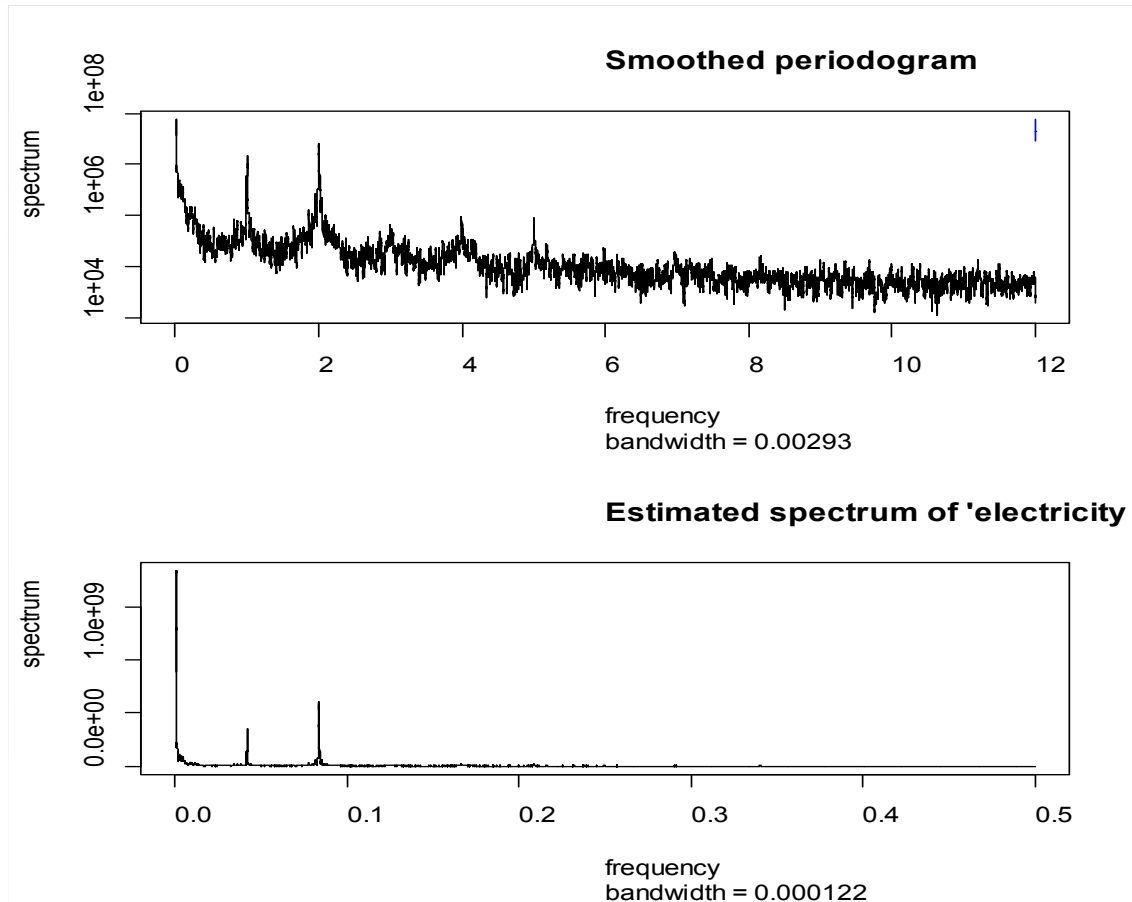


Figure 6 : Périodogramme de la consommation d'électricité au Québec en fréquence horaire.

Finalement, il est aussi intéressant de jeter un coup d'œil aux fonctions d'autocorrélation et d'autocorrélation partielle de la série de consommation d'électricité car elles nous indiquent la persistance des liens entre le présent de la série et son passé. La figure 7 présente des résultats. La consommation horaire d'électricité est fortement autocorrélée, comme le montre très clairement l'autocorrélogramme (ACF) en bas à gauche, avec une forte persistance sur les 70 heures précédentes, et une décroissance sinusoïdale très lente avec des pics aux multiples de 24h. L'autocorrélogramme partiel (PACF) montre des pics aux retards 24h et 48h.

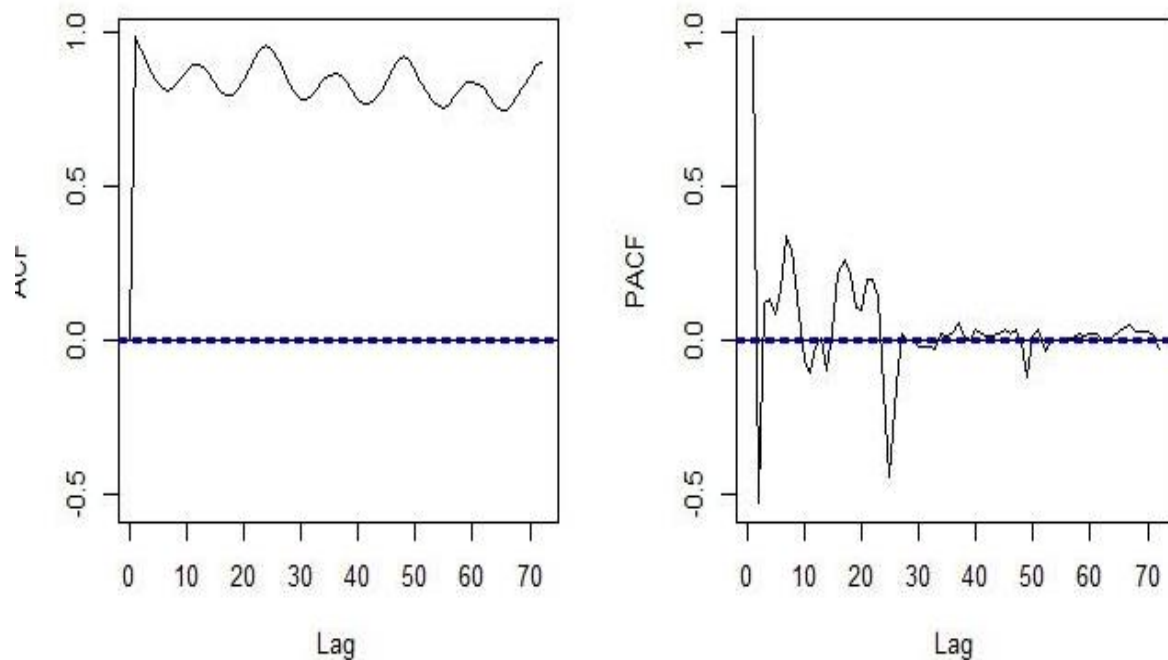


Figure 7 : ACF et PACF de la consommation horaire d'électricité.

4. Résultats

Nous avons estimé trois modèles pour ajuster et prédire la consommation horaire d'électricité au Québec de 2006 à 2010. Pour rappel, les trois méthodes retenues sont : le lissage exponentiel Holt-Winters, les modèles SARIMA et SARIMAX. Ces modèles sont présentés ci-dessous dans l'ordre énoncé.

4.1 Modèles de lissage exponentiel

Nous avons retenu deux versions du modèle de lissage exponentiel. Sous les hypothèses d'erreur multiplicative (M), de tendance additive (A) et de saisonnalité additive « *amortie* » (Ad), nous avons estimé le modèle MAdA. Sous les hypothèses d'erreur additive (A), de tendance additive amortie (Ad) et de saisonnalité additive (A), nous avons estimé le modèle AAdA. Selon Hyndman et al. (2008 :18), le meilleur critère pour choisir entre les modèles à erreur additive ou multiplicative reste la minimisation du *critère d'information d'Akaike* (AIC). En matière de performance prédictive intra et hors échantillon, nous présentons tous les indicateurs les plus classiques, mais nous concentrons sur l'erreur absolue moyenne en pourcentage (MAPE) pour établir la meilleure performance prédictive hors-échantillon. C'est l'indicateur le plus usuel dans la littérature sur le sujet. Notre échantillon d'estimation des paramètres du modèle de lissage exponentiel va du 1^{er} janvier 2006 au 24 décembre 2010. La prévision hors échantillon porte sur les 24, 72 et 168 heures qui suivent la dernière heure du 24 décembre 2010.

Tableau 3 : Comparaison du critère d'information d'Akaike (AIC).

	AIC
ETS(MAdA)	970224.3
ETS(AAdA)	974766.1

Tableau 4 : Comparaison des mesures de l'erreur²⁰ intra échantillon.

	ME	RMSE	MAE	MPE	MAPE	MASE
ETS(MAdA)	0.010	338.106	250.776	-0.14	3.953	0.581
ETS(AAdA)	0.042	337.463	250.828	-0.135	3.965	0.581

Les résultats montrent que le modèle de lissage exponentiel avec erreurs multiplicative MAdA domine le modèle AAdA en termes de critère d'information. Il se révèle légèrement plus performant en termes de prévision intra-échantillon. Sa MAPE est légèrement supérieure, mais la différence est très petite.

²⁰ Voir Annexe I pour les définitions mathématiques de ces différentes mesures d'erreurs.

ME = l'erreur moyen (Mean Error); RMSE = racine carré de l'erreur moyenne (Root Mean Squared Error); MAE = l'erreur absolue moyen (Mean Absolute Error); MPE = l'écart moyen en pourcentage (Mean Percent Error); MAPE = l'écart absolu moyen en pourcentage (Mean Absolute Percent Error); MASE = l'erreur quadratique absolue moyenne (Mean Absolute Squared Error).

1. Résultats du modèle ETS(M,Ad,A) sur l'échantillon du 1^{er} Janvier 2006 au 24 décembre 2010.

Paramètres de lissage : $\alpha = .0752$; $\beta = 1e - 04$; $\gamma = 0.2$; $\phi = 0.93$

Valeurs initiales d'état (*state*) et de croissance (*growth*):

$$l_0 = 12473.4932$$

$$b_0 = -41.05$$

$$s_{-23} = -1021.915, s_{-22} = -333.26, s_{-21} = 413.63, s_{-20} = 872.74, s_{-19} = 1151.47, s_{-18} = 1319.66, \\ s_{-17} = 1333.88, s_{-16} = 799.955, s_{-15} = -73.262, s_{-14} = -469.882, s_{-13} = -435.265, s_{-12} = -123.31 \\ s_{-11} = 192.7307, s_{-10} = 276.461, s_{-9} = 467.82, s_{-8} = 680.134, s_{-7} = 925.152, s_{-6} = 738.105, \\ s_{-5} = -299.190, s_{-4} = -1024.378, s_{-3} = -1277.51, s_{-2} = -1378.354, s_{-1} = -1402.03, s_0 = -1333.388$$

sigma: 0.0524

La prévision de consommation de 24 heures, 72 heures et 168 heures suivantes est comparée avec la consommation réelle sur les figures 8 à 10 et les différentes mesures de l'erreur de prévision hors échantillon sont présentées dans le tableau 5.

Tableau 5 : Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle MAdA.

24h				
ME	RMSE	MAE	MPE	MAPE
-356.841	734.155	595.864	-2.725	4.932

72h				
ME	RMSE	MAE	MPE	MAPE
105.362	659.35	544.260	0.979	4.452

168h				
ME	RMSE	MAE	MPE	MAPE
-793.33	1527.55	1167.150	-7.99	11.15

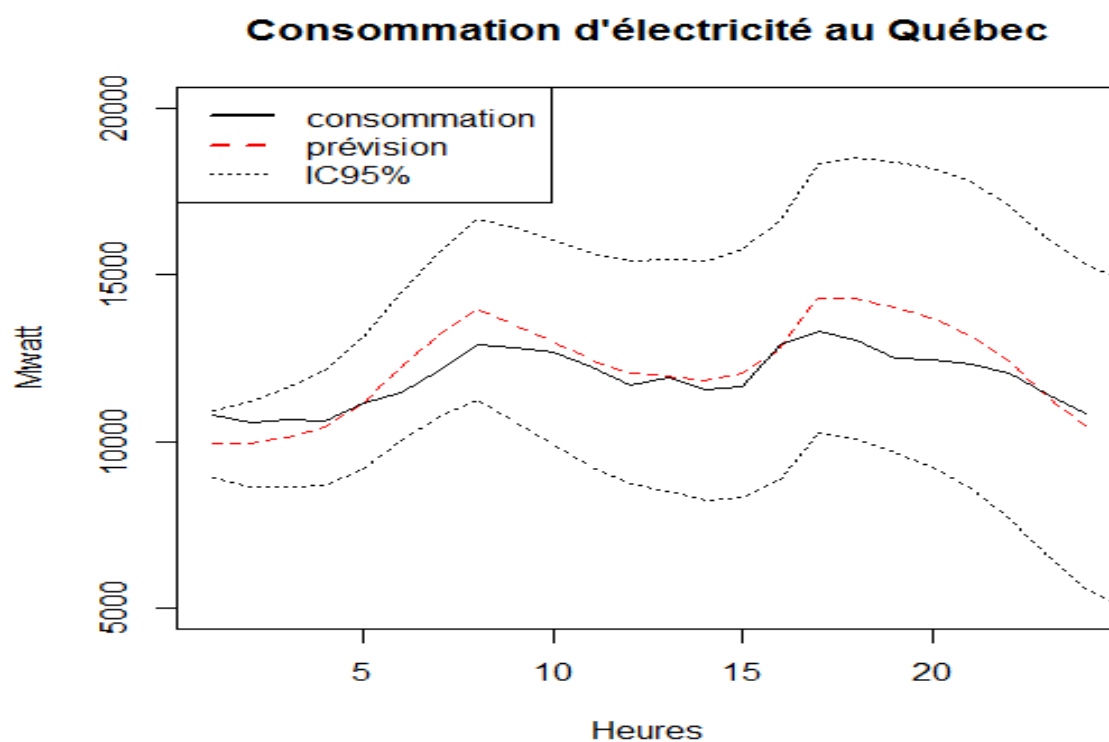


Figure 8 : Profil de prévision de la consommation d'électricité avec le modèle ETS(MAdA) pour 24h.

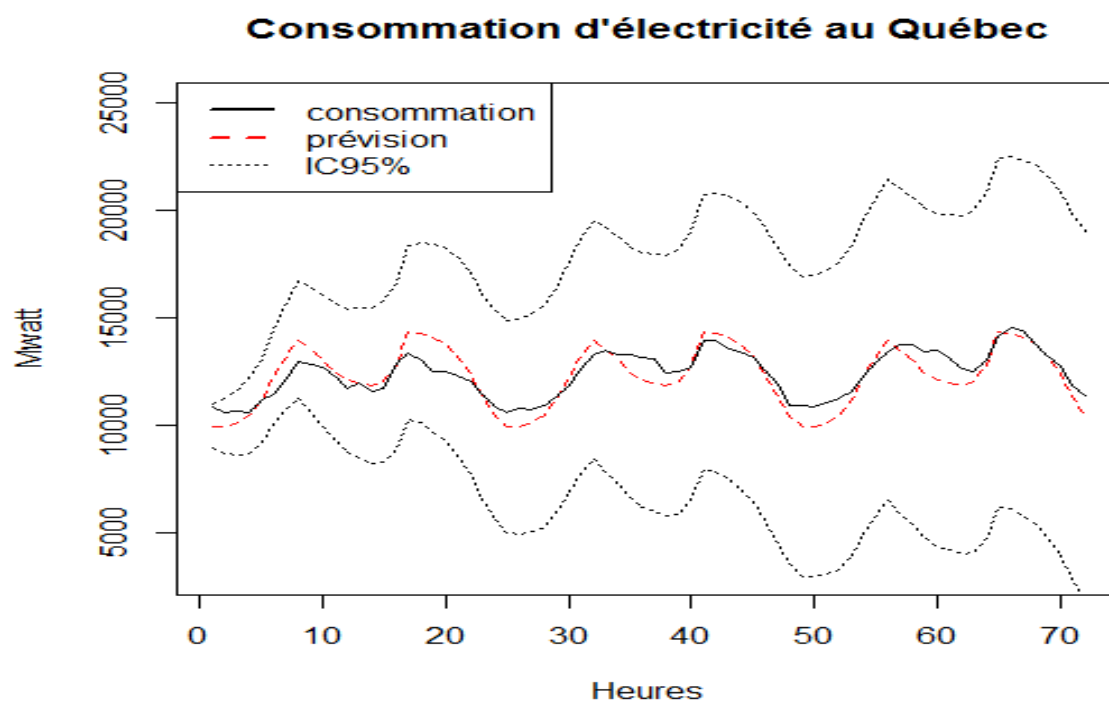


Figure 9 : Profil de prévision de la consommation d'électricité avec le modèle ETS(MAdA) pour 72h.

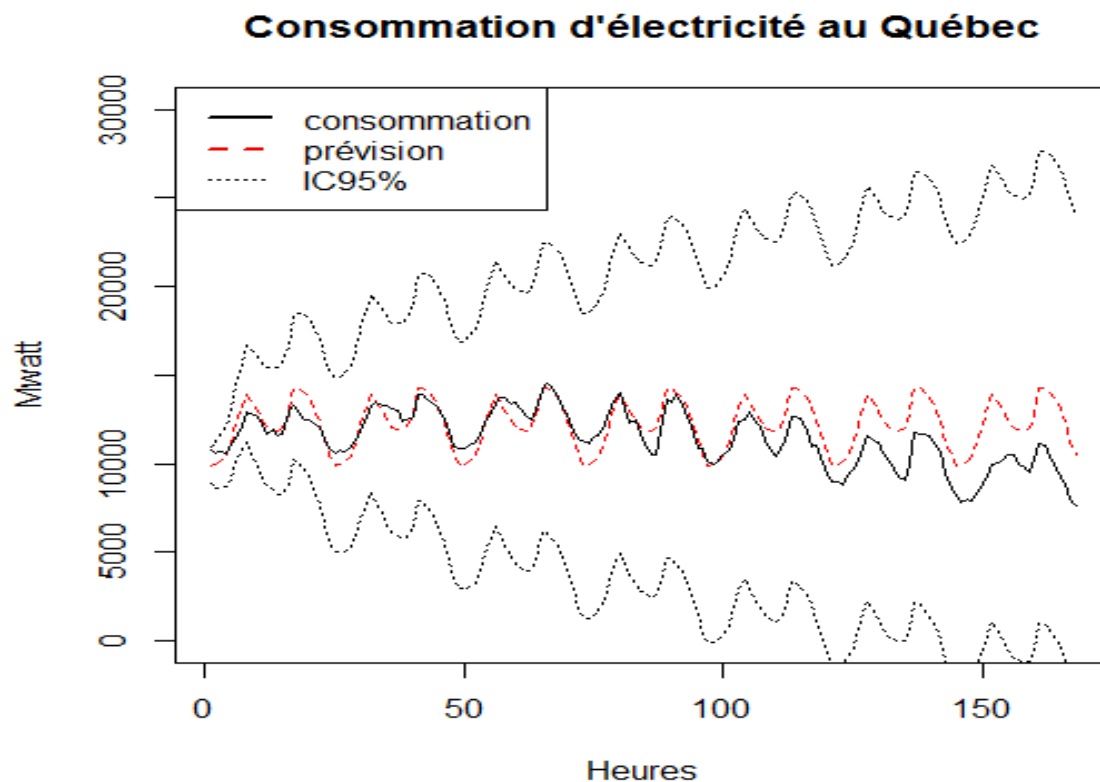


Figure 10: Profil de prévision de la consommation d'électricité avec le modèle ETS(MAdA) pour 168h.

L'intervalle de confiance à 95% de la prévision de long terme croît à un niveau d'incertitude peu utile dans la pratique. Raison pour laquelle, il semble plus adéquat de se concentrer sur la prévision hors-échantillon à 24 ou 72 heures maximum.

2. Résultats du modèle ETS(A,Ad,A) sur l'échantillon du 1^{er} Janvier 2006 au 24 décembre 2010

Paramètres de lissage:

$$\alpha = 0.796; \beta = 1e-04; \gamma = 0.204; \phi = 0.98$$

Valeurs initiales d'état (*state*) et de croissance (*growth*):

$$l_0 = 12706.33$$

$$b_0 = -41.077$$

$$s_{-23} = -1029.478, s_{-22} = -333.2131, s_{-21} = 413.662, s_{-20} = 872.7408, s_{-19} = 1151.502, \\ s_{-18} = 1319.56, s_{-17} = 1335.661, s_{-16} = 802.7571, s_{-15} = -73.3635, s_{-14} = -469.9532, \\ s_{-13} = -435.3548, s_{-12} = -123.3631, s_{-11} = 192.7279, s_{-10} = 276.3995, s_{-9} = 467.8013, s_{-8} = 680.1058,$$

$s_{-7}=925.0983$, $s_{-6}=738.0614$, $s_{-5}=-299.1451$, $s_{-4}=-1024.284$, $s_{-3}=-1277.396$, $s_{-2}=-1378.387$,
 $s_{-1}=-1401.949$, $s_0=-1330.197$

sigma: 337.4638

Tableau 6 : Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle AAdA.

24h				
ME	RMSE	MAE	MPE	MAPE
-423.303	772.245	636.072	-3.280	5.255

72h				
ME	RMSE	MAE	MPE	MAPE
37.015	650.996	538.759	0.430	4.403

168h				
ME	RMSE	MAE	MPE	MAPE
-863.630	1566.544	1198.038	-8.617	11.444

On remarque que la performance prédictive hors-échantillon du modèle multiplicatif est un moins meilleure que celle du modèle additif. La prévision de consommation à l'horizon 24 heures, 72 heures et 168 heures est comparée à la consommation réelle sur les figures 11 à 13 ci-dessous, avec un intervalle de confiance de 95%. Il est intéressant de constater que même si la qualité prévisionnelle ponctuelle du modèle additif est un peu inférieure, les prévisions hors-échantillon de long terme (168h) sont moins incertaines que celles du modèle multiplicatif, car la taille de l'intervalle de confiance à 95% est plus faible.

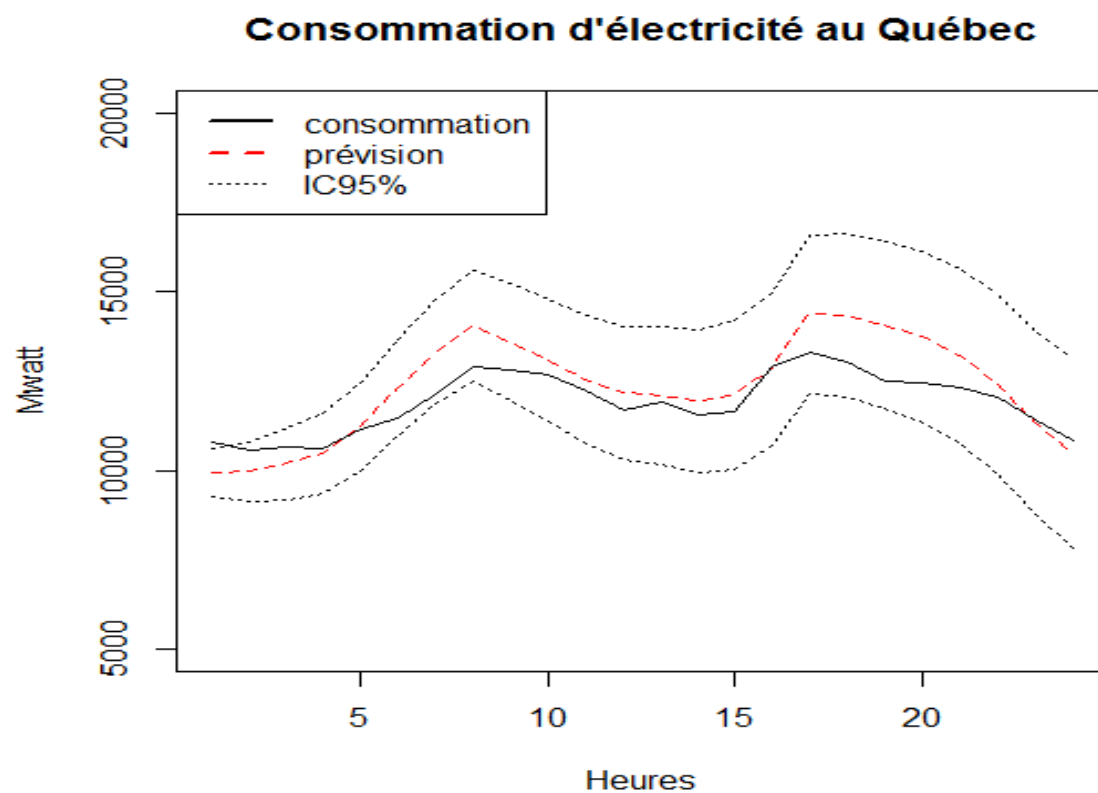


Figure 11 : Profil de prévision de la consommation d'électricité avec le modèle ETS(AAdA) pour 24h.

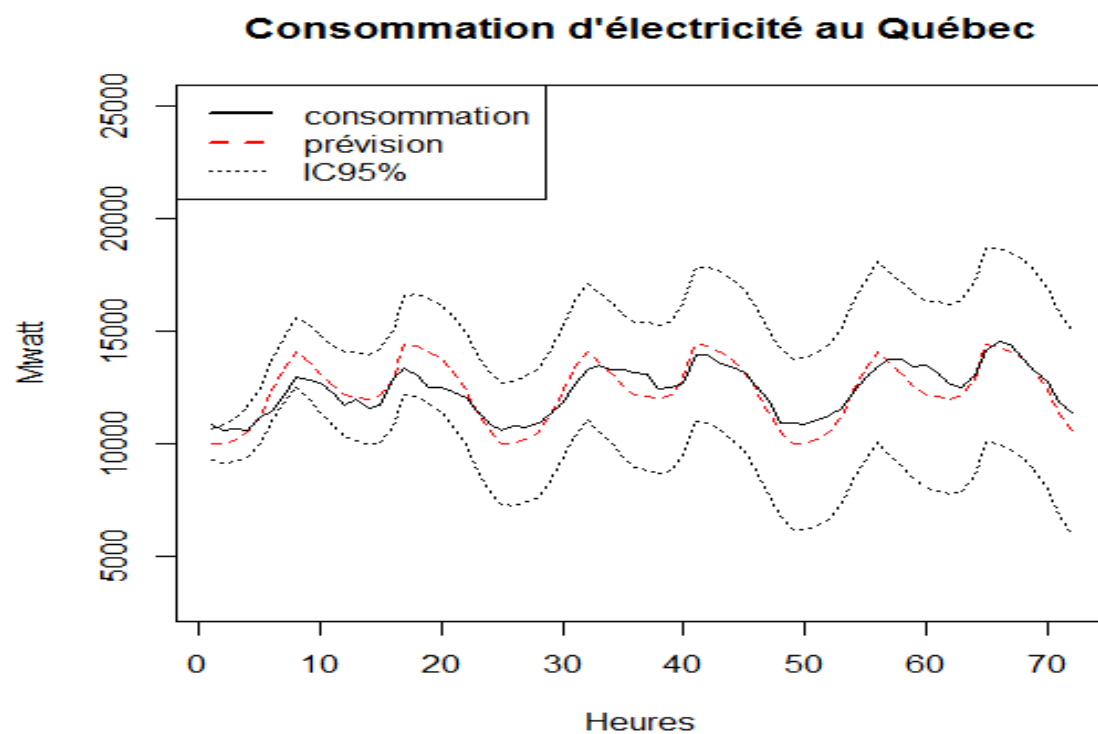


Figure 12 : Profil de prévision de la consommation d'électricité avec le modèle ETS(AAdA) pour 72h.

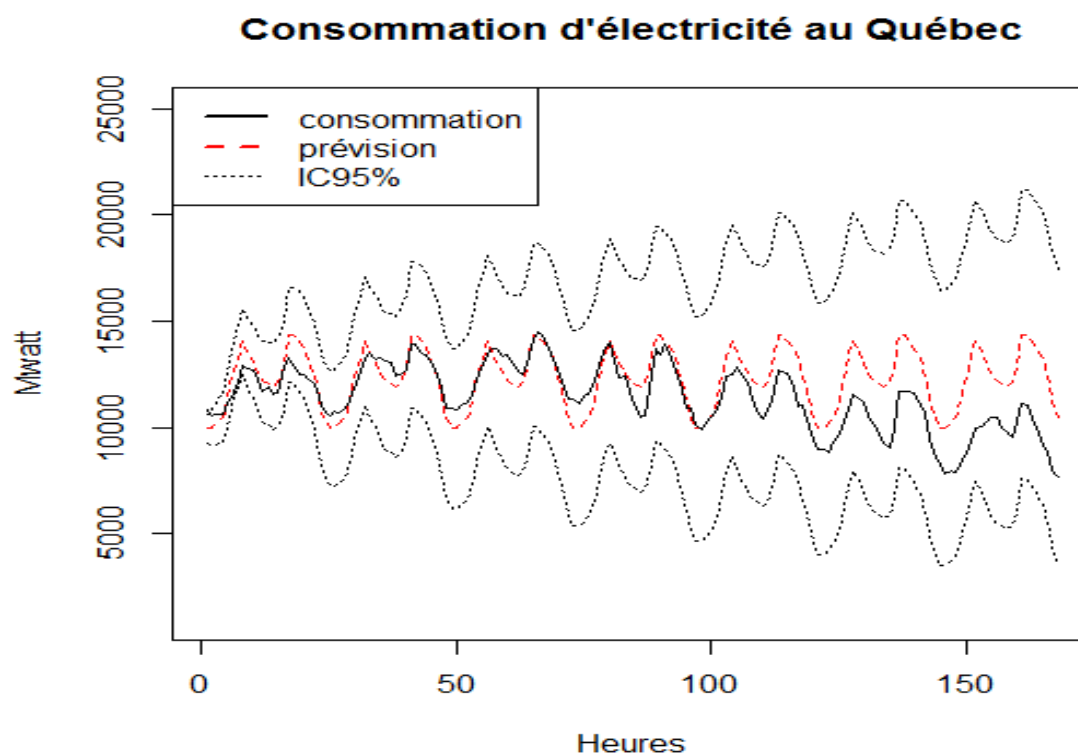


Figure 13 : Profil de prévision de la consommation d'électricité avec le modèle ETS(AAdA) pour 168h.

4.2 Modèles SARIMA

4.2.1 Tests de racine unitaire

Comme énoncé dans les sections 2.4 et 3.2, la stationnarité des séries est une hypothèse fondamentale dans la modélisation de type ARIMA/SARIMA. Le tableau 9 donne les résultats des tests ADF de stationnarité sur la chronique.

Tableau 7: Résultat du test de racine unitaire avec un retard maximum $k=168$.

Type de test	Valeurs du test de Dickey-Fuller augmenté, retard max $k=168$ h											
	empirique						Théorique (1%)					
	τ_1	τ_2	τ_3	ϕ_1	ϕ_2	ϕ_3	τ_1	τ_2	τ_3	ϕ_1	ϕ_2	ϕ_3
None	-1.18	--	--	--	--	--	-2.58	--	--	--	--	--
Drift	--	-3.01	--	4.54	--	--	--	-3.43	--	6.43	--	--
Trend	--	--	-3.05	--	3.25	4.87	--	--	-3.12	--	4.03	5.34

On remarque que nous acceptons au seuil de 1% la présence de racine unitaire dans toutes les configurations du test ADF, et qu'il n'y a pas de tendance déterministe linéaire (voir la statistique ϕ_3) dans les données. Nous concluons qu'il existe une évidence statistique solide pour considérer notre série de consommation d'électricité comme étant non stationnaire. Par conséquent, nous avons effectué une différenciation d'ordre $d=1$. Pour ce qui est de l'ordre de différenciation de la composante saisonnière, le terme D du SARIMA, notons que le test de Canova-Hansen conclut à la nécessité de différencier une fois la composante saisonnière. Par conséquent, $D=1$.

4.2.2 Estimations SARIMA

Le deuxième type de modèle que nous avons estimé est le modèle ARIMA saisonnier. Le premier de ces modèles est un SARIMA sans variables exogènes. À nouveau, l'échantillon d'estimation des paramètres couvre la période du 1er Janvier 2006 au 24 décembre 2010 et l'échantillon de prévision hors-échantillon va du 25 au 31 décembre 2010. Nous avons imposé une différenciation des séries à l'ordre 1, donc $D=d=1$. La procédure de minimisation du critère AIC conduit aux ordres (p,q) et (P,Q) présentés dans le tableau 8 :

Modèle SARIMA

ARIMA(2,0,9)(9,0,9)[24] with non-zero mean

Tableau 8: Coefficients et statistiques (SARIMA).

Coefficients	estimation	Erreur standard	t.stat	p.val
constante	6755.851	2230.716	3.028557	0.002457
ar1	0.888	0.1782	4.982028	0.000001
ar2	0.093	0.1749	0.530751	0.595591
ma1	-0.045	0.1781	-0.251749	0.801235
ma2	0.025	0.0254	0.991951	0.321221
ma3	0.0007	0.0083	0.090040	0.928256
ma4	-0.017	0.0050	-3.358811	0.000783
ma5	-0.013	0.0056	-2.314586	0.020636
ma6	-0.050	0.0052	-9.718054	0.000000
ma7	-0.066	0.0098	-6.717319	0.000000
ma8	-0.052	0.0118	-4.397182	0.000011
ma9	-0.026	0.0095	-2.700742	0.006918
sar1	0.527	0.0465	11.330212	0.000000
sar2	0.008	0.0509	0.162367	0.871017
sar3	0.222	0.0451	4.924021	0.000001
sar4	-0.078	0.0400	-1.943678	0.051934
sar5	-0.086	0.0351	-2.445276	0.014474
sar6	0.237	0.0347	6.838904	0.000000
sar7	0.648	0.0389	16.667059	0.000000
sar8	-0.149	0.0538	-2.773392	0.005548
sar9	-0.335	0.0374	-8.959914	0.000000
sma1	-0.338	0.0465	-7.268741	0.000000
sma2	-0.0006	0.0454	-0.012424	0.990087
sma3	-0.170	0.0405	-4.203676	0.000026
sma4	0.096	0.0346	2.786976	0.005320
sma5	0.116	0.0317	3.647425	0.000265
sma6	-0.166	0.0315	-5.271097	0.000000
sma7	-0.561	0.0341	-16.466177	0.000000
sma8	0.025	0.0455	0.549203	0.582866
sma9	0.295	0.0294	10.017634	0.000000
σ^2 estimée= 66916; log likelihood = -304480.3; aic = 609026				

Mesures de l'erreur intra-échantillon du modèle SARIMA

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
0.112	258.035	193.466	-0.1436	3.040	0.323	-0.00015

Ce modèle se révèle meilleur que les méthodes de lissage au niveau de la fiabilité prédictive intra-échantillon, pour la plupart des indicateurs, et en particulier en termes de MAPE.

Les figures 14 à 16 comparent respectivement la consommation d'électricité observée à la prévision effectué pour 24h, 72h, 168h avec le modèle SARIMA dans un intervalle de confiance de 95%.

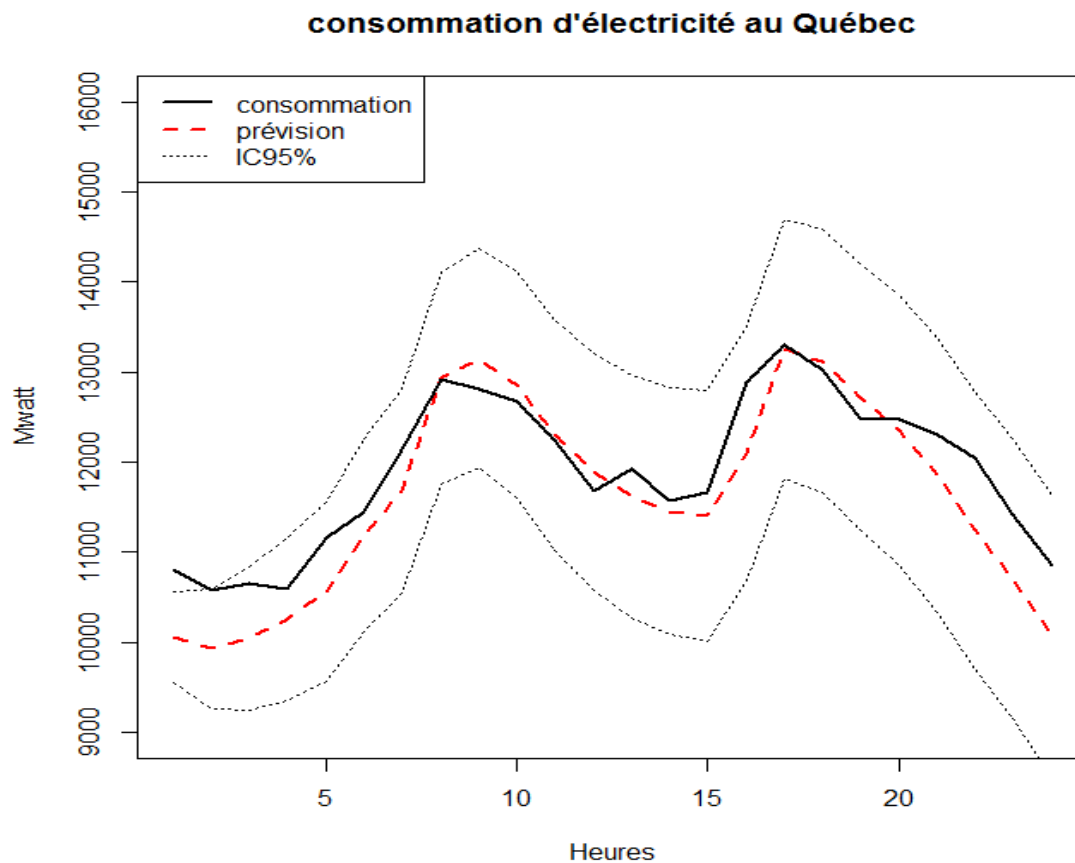


Figure 14 : Profil de prévision de la consommation d'électricité avec le modèle SARIMA pour 24h.

Tableau 9 : Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle SARIMA.

24h				
ME	RMSE	MAE	MPE	MAPE
286.14	462.74	382.164	2.54	3.308

72h				
ME	RMSE	MAE	MPE	MAPE
1036.181	1247.605	1068.19	8.4	8.65

168h				
ME	RMSE	MAE	MPE	MAPE
391.4	1287.703	1084.394	2.439	9.72

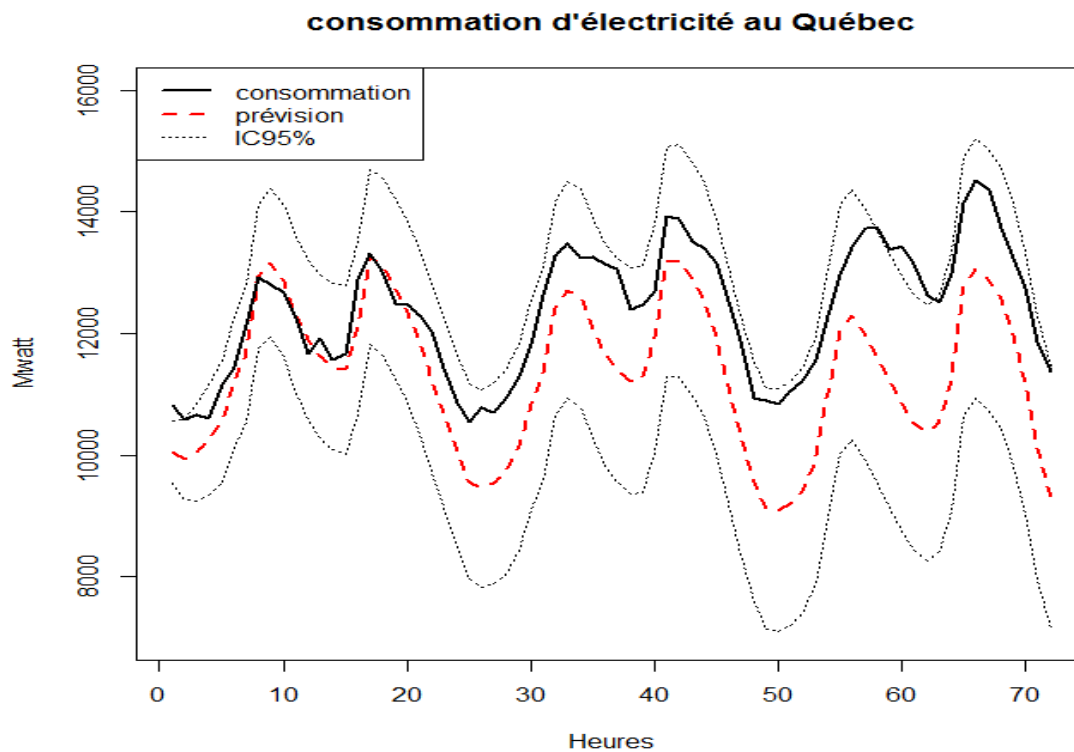


Figure 15 : Profil de prévision de la consommation d'électricité avec le modèle SARIMA pour 72h.

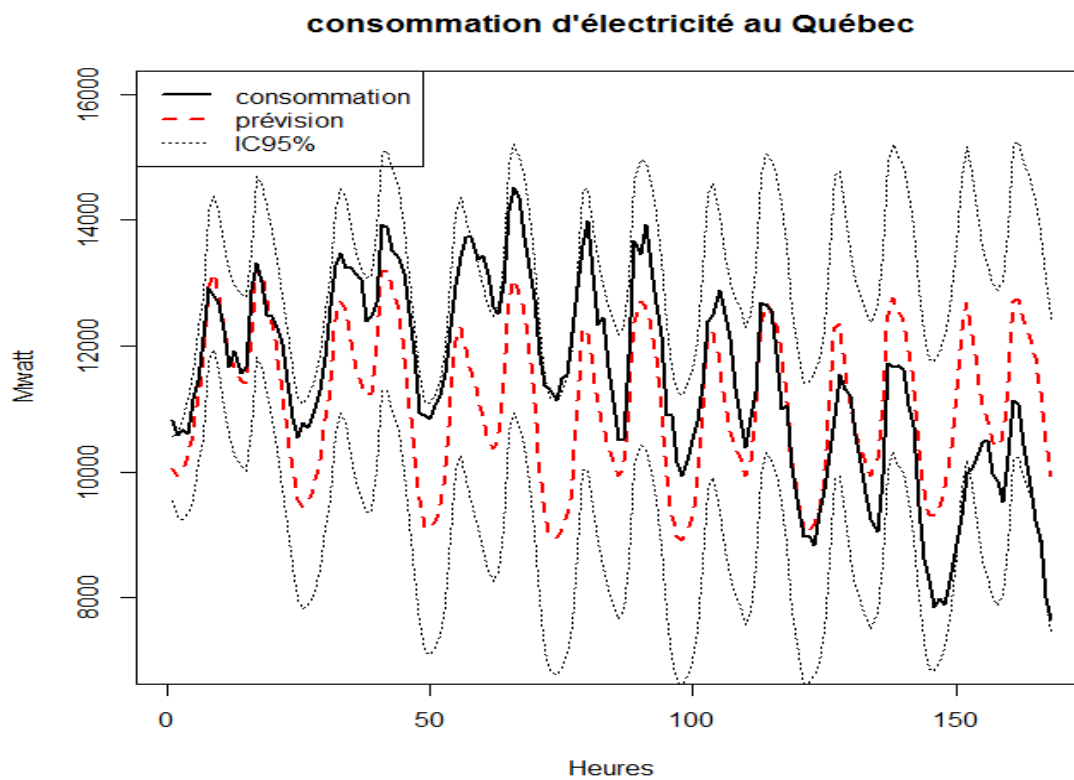


Figure 16 : Profil de prévision de la consommation d'électricité avec le modèle SARIMA pour 168h.

La performance hors-échantillon du SARIMA est meilleure que le lissage exponentiel pour les échantillons de 24h et de 168h mais pas sur 72h. L'incertitude de la prévision sur 160h est notablement inférieure avec le SARIMA qu'avec le lissage exponentiel.

Le second modèle est un SARIMAX avec variables exogènes climatiques. La même procédure d'estimation est suivie que pour le SARIMA. Les résultats sont les suivants :

Modèle SARIMAX

ARIMA(1,0,0)(4,0,3)[24] with non-zero mean

Tableau 10: Coefficients et statistiques(SARIMAX).

Coefficients	estimation	Erreur standard	t.stat	p.val
constante	6873.112	820.6521	8.375	0
ar1	0.927	0.0020	466.040	0
sar1	2.135	0.0057	377.507	0
sar2	-1.998	0.0124	-160.863	0
sar3	0.772	0.0121	63.764	0
sar4	0.088	0.0053	16.523	0
sma1	-1.983	0.0046	-433.941	0
sma2	1.830	0.0070	261.561	0
sma3	-0.724	0.0044	-165.901	0
hr	5.066	0.6865	7.379	0
pr	-50.941	2.8498	-17.875	0
Tem	-64.124	2.8470	-22.523	0
N	5.494	0.7733	7.104	0
σ^2 estimée = 75486: log likelihood=-307110.8 AIC=614251.7 AICc=614251.8 BIC=614373.3				

Mesures de l'erreur intra-échantillon du modèle SARIMAX

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-0.017	274.442	207.148	-0.195	3.305	0.346	-0.117

On constate d'abord que la performance prédictive intra-échantillon du SARIMAX est inférieure à celle du SARIMA. Par conséquent, le nouveau modèle obtenu lorsque l'on prend en compte les variables climatiques n'améliore pas l'ajustement des données intra-échantillon. Ce modèle est tout de même supérieur aux deux modèles de lissage exponentiel.

Les figures 17 à 19 comparent respectivement la consommation d'électricité observée à la prévision effectué pour 24h, 72h, 168h avec le modèle SARIMAX dans un intervalle de confiance de 95%.

Tableau 11: Mesure de l'erreur de prévision hors échantillon sur 24h, 72h et 168h avec le modèle SARIMAX.

24h				
ME	RMSE	MAE	MPE	MAPE
234.108	462.672	388.78	2.099	3.332

72h				
ME	RMSE	MAE	MPE	MAPE
973.55	1177.96	1025.109	7.906	8.32

168h				
ME	RMSE	MAE	MPE	MAPE
563.41	1032.63	857.63	4.370	7.445

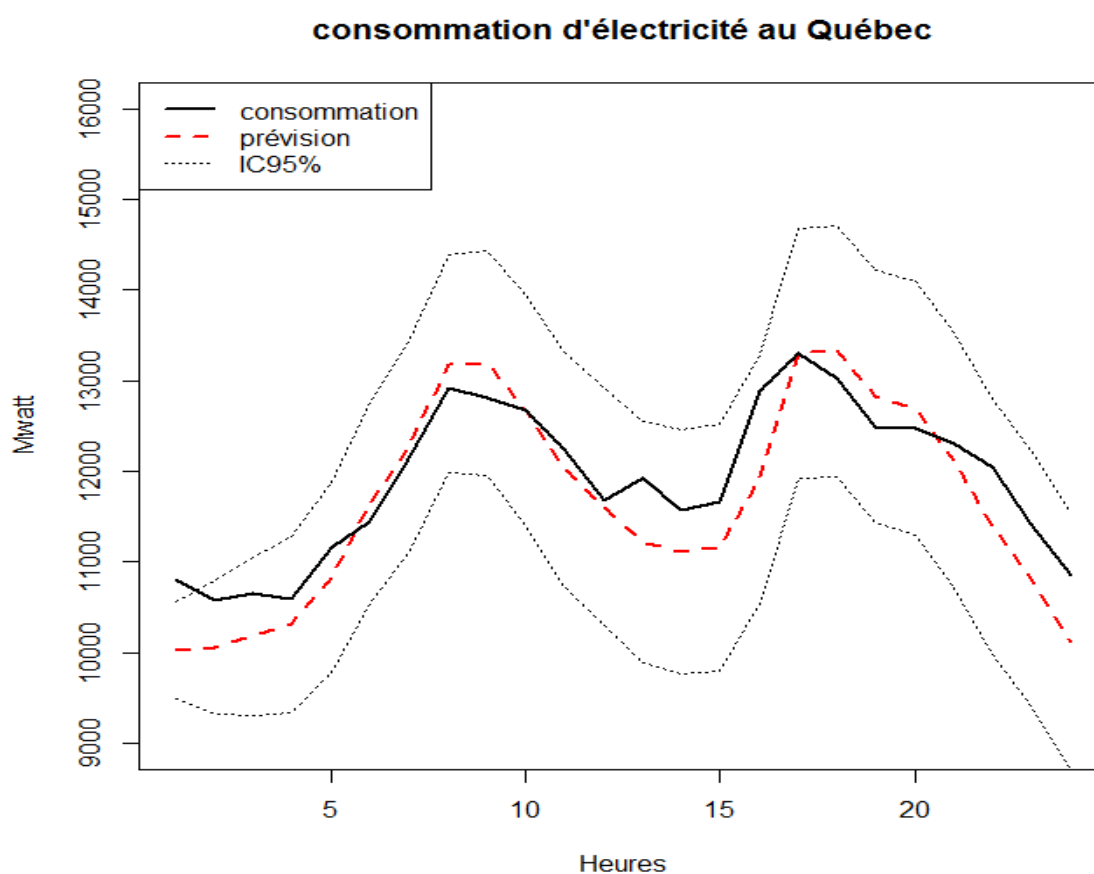


Figure 17 : Profil de prévision de la consommation d'électricité avec le modèle SARIMAX pour 24h.

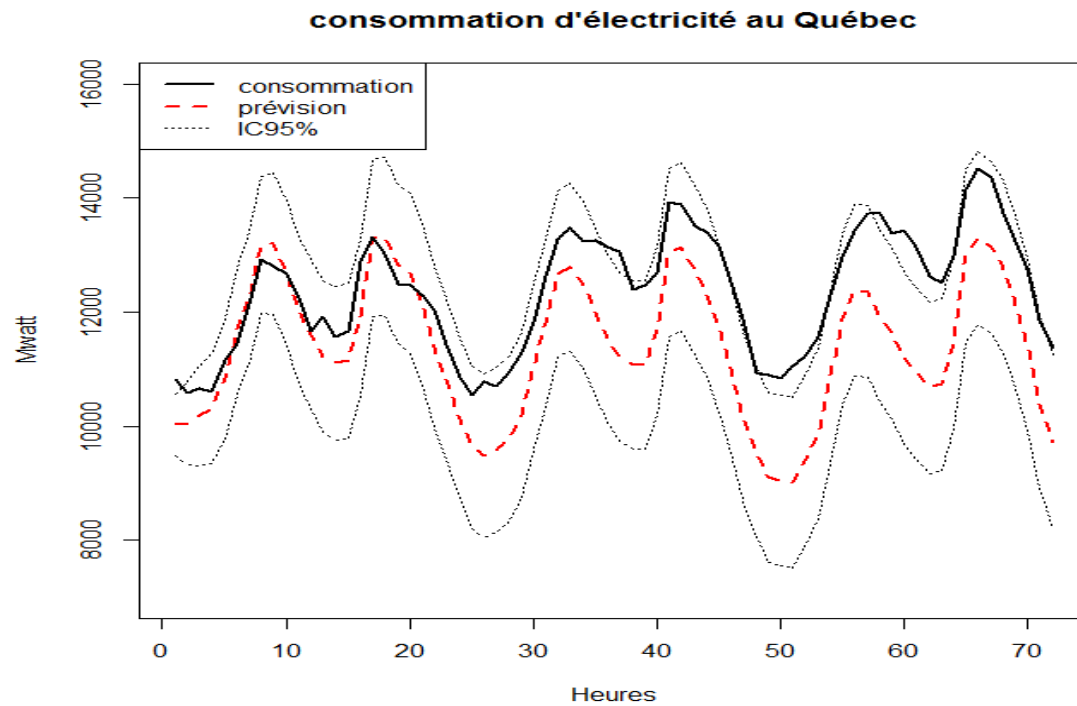


Figure 18 : Profil de prévision de la consommation d'électricité avec le modèle SARIMAX pour 72h.

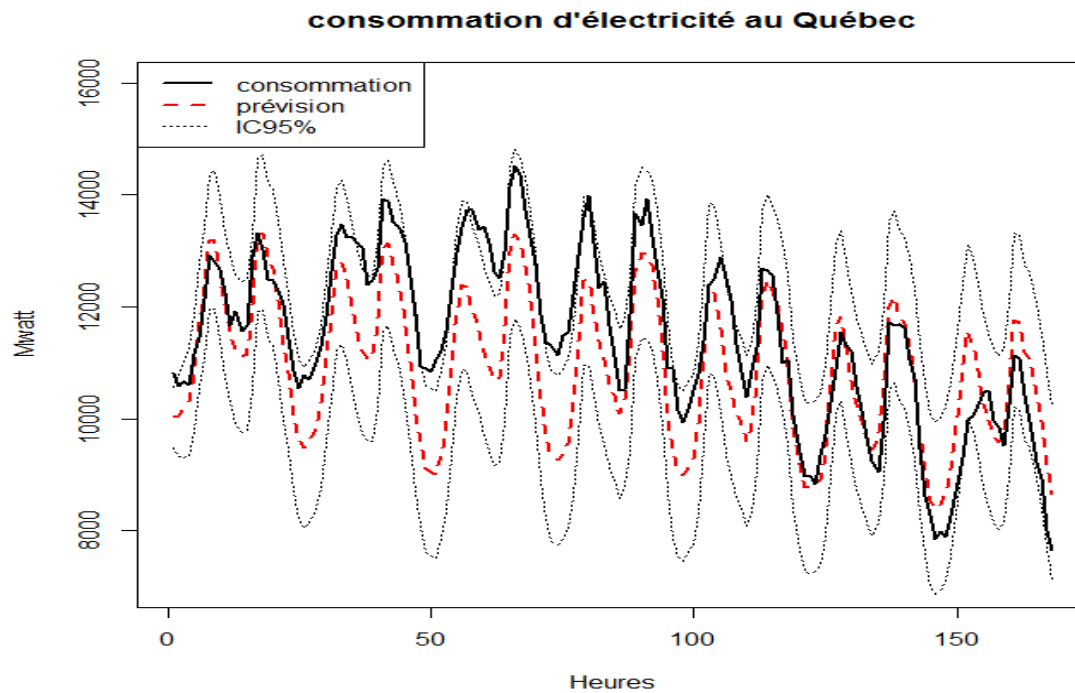


Figure 19 : Profil de prévision de la consommation d'électricité avec le modèle SARIMAX pour 168h.

Hors échantillon, il s'avère que le SARIMAX est le meilleur modèle pour prédire ponctuellement à l'horizon 72h et 168h. En effet, la MAPE pour ces horizons temporels est la plus faible de tous les modèles. Il ressort aussi clairement des graphiques que le SARIMAX est le moins incertain à long terme.

4.3 Diagnostic des modèles

Nous développons dans cette section quelques tests de spécification des modèles SARIMA et SARIMAX.

4.3.1 Tests pour le modèle SARIMA

La normalité des erreurs est importante pour la fiabilité de nos intervalles de confiance asymptotiques. Pour tester la normalité des résidus du modèle SARIMA, nous avons effectué ci-dessous le test de Jarque-Bera.

Résultat du test effectué sur le logiciel *R* avec la fonction **jarqueberaTest()** :

```
Jarque - Bera Normality Test
Test Results:
STATISTIC:
X-squared: 8948.5565
P VALUE: Asymptotic p Value: < 2.2e-16
```

Nous concluons que les résidus de ce modèle ne sont pas normalement distribués.

La statistique de *Ljung-Box* permet de vérifier l'hypothèse nulle d'absence d'autocorrélation des erreurs. Nous rejetons aussi cette hypothèse sur les résidus du modèle SARIMA aux retards 24,72 et 168. Les résidus de ce modèle sont par conséquent fortement corrélés.

Résultat du test effectué sur le logiciel *R* avec la fonction **Box.test ()** :

```
Lag=24, Box-Ljung test
X-squared = 270.1232, df = 24, p-value < 2.2e-16
```

```
Lag=72, Box-Ljung test
X-squared = 705.9115, df = 72, p-value < 2.2e-16
```

```
Lag=168, Box-Ljung test
X-squared = 1026.484, df = 168, p-value < 2.2e-16
```

4.3.2 Tests pour le modèle SARIMAX

Les mêmes tests que ci-dessus sont appliqués au modèle SARIMAX.

Résultat du test de normalité effectué sur R par la fonction **jarqueberaTest()** :

Jarque - Bera Normality Test
Test Results:
STATISTIC:
X-squared: 8948.5565
P VALUE: Asymptotic p Value: < 2.2e-16

Résultat du test d'autocorrélation effectué sur R par la fonction **Box.test ()** :

Lag=24, Box-Ljung test
X-squared = 1521.21, df = 24, p-value < 2.2e-16

Lag=72, Box-Ljung test
X-squared = 2582.233, df = 72, p-value < 2.2e-16

Lag=168, Box-Ljung test
X-squared = 4016.581, df = 168, p-value < 2.2e-16

Nous rejetons à nouveau les hypothèses nulles de normalités et d'absence d'autocorrélation des résidus. Notons finalement que ces mêmes hypothèses sont rejetées pour les estimations Holt-Winters. Par conséquent, les intervalles de confiance basés sur la normalité des erreurs ne sont pas fiables. Cependant, l'estimation ponctuelle reste non biaisée.

5. Conclusion

Dans cette étude, nous avons utilisé les données journalières de consommation d'électricité et de variables climatiques (température, nébulosité, point de rosée, etc) du Québec pour tester des modèles de prévision de cette consommation à court terme. L'exploration de ces données nous a révélé des caractéristiques tels que la non stationarité stochastique de la série de consommation d'électricité et sa très forte inertie (autocorrélation forte et longue). Un seul processus de différenciation est nécessaire pour rendre la série stationnaire. Il en va de même pour la saisonnalité stochastique journalière, qui requiert une différenciation. Nous avons effectué des prévisions tant intra que hors échantillon avec trois types de modèles : la méthode de lissage exponentiel Holt-Winters, un SARIMA et un SARIMAX avec variables exogènes climatiques. Nous avons proposé des

prédictions sur trois horizons temporels : 24h (1 jour), 72h (3 jours) et 168h (1 semaine). Les modèles SARIMA et SARIMAX se montrent généralement plus performants que les lissages Holt-Winters à erreurs additive et multiplicative. La qualité des prévisions est tout de même globalement satisfaisante, avec l'erreur absolue moyenne en pourcentage en dessous de 10% dans la plupart des modèles. Par ailleurs, le modèle SARIMA s'avère le plus performant pour prédire la consommation d'électricité sur 24 heures, alors que le SARIMAX se révèle le plus performant sur 72h et 168h. Nous concluons donc que les variables climatiques sont particulièrement intéressante pour prédire à moyen (3 jours) et long terme (1 semaine) la consommation d'électricité avec des données horaires.

Les modèles estimés doivent cependant être améliorés car certaines hypothèses statistiques fondamentales n'ont pu être validées. Une piste évidente à poursuivre est de prendre en compte les cycles mensuelles et saisonniers (printemps, été, automne, hiver) comme variables explicatives exogènes dans le modèle SARIMAX. Nos tentatives dans ce sens se sont heurtées à des problèmes de convergence de l'estimateur et aux limites computationnelles de notre ordinateur. Il serait également intéressant d'explorer des modèles à mémoire longue, de type ARFIMA, pour capturer les très fortes inerties présentes dans nos données. Des recherches supplémentaires permettraient sans doute de fournir à Hydro-Québec des modèles statistiques plus fiables pour effectuer des prévisions performantes de très court terme dans le cadre d'analyse que nous avons exploré.

Références bibliographiques

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second international symposium on information theory (pp. 267-281). Akademiai Kiado.
- Alfares, H. K. and M. Nazeeruddin (2002). Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science* (1): 23-34.
- Amaral, L. F., R. C. Souza, et al. (2008). A smooth transition periodic autoregressive (STPAR) model for short-term load forecasting. *International Journal of Forecasting* (4): 603-615.
- Amjady, Nima (2001). Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *Power Systems, IEEE Transactions on* 16.3 (2001): 498-505.
- Aragon, Y. (2011). *Séries temporelles avec R : Méthodes cas*. Springer
- Bernard, J. T., D. Bolduc, et al. (1996). Quebec residential electricity demand: a microeconomic approach. *Canadian Journal of Economics*: 92-113.
- Bernard, J. T., D. Bolduc, et al. (2011). A pseudo-panel data model of household electricity demand. *Resource and Energy Economics* (1): 315-325.
- Bisgaard, S. and M. Kulachi (2011). *Times series analysis and forecasting by example*. John Wiley & sons
- Box, G. E., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-D. iv, San Francisco.
- Box, G. E. P. G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Time Series and Digital Processing
- Bunn, D., & Farmer, E. D. (1985). *Comparative models for electrical load forecasting*. John Wiley & Son
- Canova, F., & Hansen, B. E. (1995). Are seasonal patterns constant over time? A test for seasonal stability. *Journal of Business & Economic Statistics*, 13(3), 237-252.
- Campbell P. R. J. et Adamson K. (2006). *Methodologies for Load Forecasting*. Intelligent systems
- Chen, B. J., & Chang, M. W. (2004). Load forecasting using support vector machines: A study on EUNITE competition 2001. *Power Systems, IEEE Transactions on*, 19(4), 1821-1830.
- Cho, M. Y., Hwang, J. C., et Chen, C. S. (1995). Customer short term load forecasting by using ARIMA transfer function model. In *Energy Management and Power Delivery*
- Cottet, R., & Smith, M. (2003). Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464).
- Cowpertwait, P. S., et Metcalfe, A. V. (2009). *Introductory time series with R*. Springer.
- Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: Journal of the Econometric Society*, 1057-1072.
- Espinoza, M., J. A. K. Suykens, et al. (2007). Electric load forecasting. *Control Systems, IEEE* 27(5): 43-

- Espinoza, M., J. A. K. Suykens, et al. (2006). LS-SVM regression with autocorrelated errors. Proc. of the 14th IFAC Symposium on System Identification (SYSID).
- Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer
- Fan J.Y. and McDonald J.D. (1994). A Real-Time Implementation of Short-Term Load Forecasting for Distribution Power Systems. IEEE Transactions on Power Systems, 9:988-994,
- Fogel D.B (1994). An Introduction to Simulated Evolutionary Optimization. IEEE Transactions on Neural Networks.
- Gep, B. O. X., Gailim M. Jenkins, and Gregory C. Reinsel. (1976). "Time series analysis: forecasting and control." Oakland CA: Holden-Day.
- Hahn, H., S. Meyer-Nieberg, et al. (2009). Electric load forecasting methods: Tools for decision making. European Journal of Operational Research 199(3): 902-907.
- Haida, T. and S. Muto (1994). Regression based peak load forecasting using a transformation technique. Power Systems, IEEE Transactions on 9(4): 1788-1794.
- Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. Power Systems, IEEE Transactions on, 16(1), 44-55.
- Hobbs, B. F., Jitprapaikulsaarn, S., Konda, S., Chankong, V., Loparo, K. A., & Maratukulam, D. J. (1999). Analysis of the value for unit commitment of improved load forecasts. Power Systems, IEEE Transactions on, 14(4), 1342-1348.
- Hyndman, R., Koehler, A., Ord, K., & Snyder, R. (2008). Forecasting with Exponential Smoothing: The State Space Approach. *Springer Series in Statistics*
- Hyndman, R.J. and Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, 26(3).
- Sumer, K. K., Goktas, O., & Hepsag, A. (2009). The application of seasonal latent variable in forecasting electricity demand as an alternative method. Energy policy, 37(4), 1317-1322.
- Kulahci, M. (2011). Time Series Analysis and Forecasting by Example, Wiley.
- Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2), 297-303.
- Metaxiotis, K., Kagiannas, A., Askounis, D., & Psarras, J. (2003). Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. Energy conversion and Management, 44(9), 1525-1534.
- Mohandes, M. (2002). Support vector machines for short-term electrical load forecasting. International Journal of Energy Research, 26(4), 335-345.
- Nowicka-Zagrajek, J. and R. Weron (2002). Modeling electricity loads in California: ARMA models with hyperbolic noise. Signal Processing 82(12): 1903-1915.
- Pardo A., Meuno V. et Valdor E. (2002). Temperature and seasonality influences on Spanish electricity load. Energy economics.

- Rahman S. (1990). Formulation and Analysis of a Rule-Based Short-Term Load Forecasting Algorithm. *Proceedings of the IEEE*, 78:805-816
- Soares, L. J. and M. C. Medeiros (2008). Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. *International Journal of Forecasting* (4): 630-644.
- Soares, L. J., & Souza, L. R. (2006). Forecasting electricity demand using generalized long memory. *International Journal of Forecasting*, 22(1), 17-28.
- Taylor, L. D. (1975). The demand for electricity: a survey. *The Bell Journal of Economics*: 74-110.
- Taylor, James W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society* 54.8 : 799-805.
- Taylor, J. W., L. M. De Menezes, et al. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting* (1): 1-16.
- Taylor, J. W. and P. E. McSharry (2007). Short-term load forecasting methods: An evaluation based on european data. *Power Systems, IEEE Transactions* (4): 2213-2219.
- Taylor, James W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research* 204.1 (2010): 139-152.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
- Wang, H., Li, G., & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3), 347-355.
- Wang, H., Li, G., & Tsai, C. L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 63-78.
- Weiss, A. A., & Andersen, A. P. (1984). Estimating time series models using the relevant forecast evaluation criterion. *Journal of the Royal Statistical Society. Series A (General)*, 484-487.
- Weron, R. (1998). Modeling and forecasting electricity loads and prices: A statistical approach. *HSC Boo*
- Weron, R. (2006). Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach (The Wiley Finance Series).
- Weron, R. and A. Misiorek (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting* (4): 744-763.
- Yang, H. T., C. M. Huang, et al. (1995). Identification of ARMAX model for short term load forecasting: an evolutionary programming approach. *Power Industry Computer Application Conference, 1995. Conference Proceedings., 1995 IEEE, IEEE*.

Rapport-Gratuit.com

ANNEXE I

Formules mathématiques des différentes mesures des résidus de la modélisation.

ME = l'erreur moyen (Mean Error)

$$ME = \frac{1}{n} \sum_t e_t$$

RMSE = racine de l'erreur quadratique moyen (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n} \sum_t e_t^2}$$

MAE = l'erreur absolue moyen (Mean Absolute Error)

$$MAE = \frac{1}{n} \sum_t |e_t|$$

MPE = l'écart moyen en pourcentage (Mean Percent Error)

$$MPE = 100 \frac{1}{n} \sum_t \frac{e_t}{Y_t}$$

MAPE = l'écart absolu moyen en pourcentage (Mean Absolute Percent Error)

$$MAPE = 100 \frac{1}{n} \sum_t \left| \frac{e_t}{Y_t} \right|$$

MASE = l'erreur quadratique absolu moyen (Mean Absolute Squared Error)

$$MASE = \frac{1}{n} \sum_t |e_t|^2$$