MATH 60638 – Méthodes de prévision

Projet de semestre – Partie 2

Évaluation des méthodes de lissage et de régression

Présenté à

Pre. Debbie Dupuis

Par l'Équipe D

Christelle GEORGE – 11288106

Samy SENOUNE – 11290255

Le 19 mars 2021

À Montréal

Objectif

Dans la partie 1 du projet, il était question d'évaluer les méthodes naïves ainsi que d'analyser les variables explicatives possibles pour la prévision du pic de la demande horaire pour jour $t+1$ à Lincoln. Cette analyse a montré que la méthode du *naïve no change* ($\hat{y}_{t+1|t} = y_t$) obtenait la meilleure performance avec un MAPE de 8,1 % et un biais de -0,113 MWh ; méthode qui sert désormais de benchmark pour nos travaux. Dans cette partie 2, quelques ajustements ont été faits. La variable météorologique du vent a été ajoutée, en imputant les valeurs manquantes journalières par celles de la veille. Cette variable permet de calculer l'effet du refroidissement du vent (CP), qui sera utilisée dans la partie sur les méthodes de régression. La variable de tornade a été ajoutée également, puisque la ville de Lincoln est sujette à ce phénomène météorologique. Finalement, la température du référence du HDD a été ajustée à 10℃. Le but de la partie 2 est d'évaluer les méthodes de lissage et de régression quant à la performance de la prévision du pic de la demande horaire pour jour $t+1$ à Lincoln.

Méthodes de lissage

Méthodologie

La partie 1 a permis de mettre en évidence deux saisonnalités dans le pic de la demande horaire d'électricité : une saisonnalité hebdomadaire et une saisonnalité annuelle. La méthode de Taylor et le modèle TBATS permettent d'accommoder deux saisonnalités et seront évalués. Ceci dit, nous évaluerons également les méthodes du lissage simple, du lissage Holt, du lissage Holt Winters et les modèles à espace d'état. Le seul modèle écarté est celui de BATS, en raison du temps computationnel requis. L'ensemble des observations est divisé en trois groupes : l'échantillon d'entraînement (du 1er janvier 2011 au 31 décembre 2016), de validation (du 1er

janvier 2017 au 31 décembre 2018) et de test (du 1er janvier 2019 au 31 décembre 2020). D'autre part, étant donné la nature court-terme de la prévision demandée du pic de la demande (h=1), deux fenêtres de temps sont employées : une rolling et une expanding. La fenêtre rolling se déplace au fur et à mesure des prévisions sur la validation. Elle est de taille fixe, soit l'équivalent de 6 ans. La fenêtre expanding va s'agrandir avec le temps, incorporant progressivement les observations de l'échantillon de validation. Sa taille va varier, de l'équivalent de 6 ans à 8 ans. Les performances liées aux deux fenêtres seront comparées sur l'ensemble de validation: ceci permettra d'analyser qui du présent ou du passé est le plus représentatif du futur. Finalement, deux approches sont utilisées : dans le premier cas, le modèle est entraîné une unique fois sur l'ensemble d'entrainement et sera employé pour la totalité des prévisions. Dans le deuxième cas, le modèle sera re-entraîné pour chaque année de la validation, soit à deux reprises, afin de mettre à jour ses paramètres pour la prévision. Il s'agit alors de comparer les différentes performances des deux approches sur l'ensemble de validation afin de déterminer si l'impact du réentrainement est significatif dans la prévision du pic de la demande. Au final, chaque méthode et modèle sera évalué selon quatre configurations: avec une fenêtre rolling et expanding, avec et sans réentrainement. Pour comparer les méthodes entre elles, le test bilatéral de Diebold-Mariano est utilisé à un niveau 5% avec l'hypothèse nulle que l'erreur quadratique moyenne des prévisions des méthodes est la même. Ce test va donc indiquer si un ensemble de prévisions est significativement différent d'un autre afin de désigner le meilleur.

Analyse des résultats

Les méthodes de **lissage simple** et du **lissage double**, pour les quatre configurations énoncées plus haut, donnent les mêmes résultats avec un MAPE de 8,1%. Pour le lissage simple, le modèle choisi est celui avec erreurs additives et un alpha de 0,99. Pour le lissage double, c'est un modèle

avec erreurs et tendance additives, dont le paramètre du niveau, alpha, vaut 0,99 et le paramètre de tendance, beta, vaut 0,0001 : la pente n'est pas considérée. Sur nos données, le lissage double se comporte comme le lissage simple. D'autre part, la valeur de alpha des deux méthodes, quasi-égale à 1, implique que seules les observations récentes influent sur la prévision (la prévision $\hat{y}_{t+1}$ est quasiment égale à l'observation y_t). Ceci explique pourquoi la performance obtenue est similaire à celle de notre benchmark *naïve no change*. De plus, le réentrainement du modèle ne modifie que très légèrement ses paramètres (de l'ordre de la 3ème décimale). Ainsi, le type de fenêtre et le réentrainement du modèle n'a donc pas d'incidence sur la performance de la prévision. Les résultats obtenus (MAPE de 8,1%, similaires au benchmark) ne sont pas étonnants car les lissages simple et double n'accommodent ni tendance (en raison du paramètre beta quasi-nul du lissage double) ni saisonnalité. Ce ne sont donc pas des méthodes adaptées à la série du pic de la demande horaire. La méthode de **Holt-Winter** est ensuite considérée, puisqu'elle permet d'accommoder une tendance et une saisonnalité. Ici, seule la saisonnalité hebdomadaire peut être étudiée car la méthode ne tolère pas de grandes fréquences. Le modèle optimal obtenu est un modèle additif (erreurs, tendance et saisonnalité additives), qu'il soit réentraîné ou non. La meilleure performance est obtenue pour la fenêtre expanding avec un MAPE de 6,9 %, et pour laquelle 5,9% soit 43 observations sur les 730 de validation, se trouvent hors de l'intervalle de confiance à niveau 95%. La fenêtre rolling obtient une moins bonne performance, mais le réentrainement du modèle permet d'améliorer significativement la performance, passant d'un MAPE de 8,3 à 7,9 %. On voit aussi que les prévisions de la fenêtre rolling sont plus biaisées, surtout sans réentrainement du modèle, qui affiche un ME de -8,6 MWh. On conclue que la

Tableau 1 : Mesures de performance des méthodes du lissage simple et double sur validation

	LISSAGE SIMPLE				LISSAGE DOUBLE			
	SANS réentrainement		AVEC réentrainement		SANS réentrainement		AVEC réentrainement	
	RW	EW	RW	EW	RW	EW	RW	EW
ME	0.113	0.113	0.110	0.113	-0.097	-0.090	-0.037	-0.003
RMSE	54.025	54.025	53.990	54.025	53.999	53.999	53.998	54.033
MAE	39.015	39.015	39.033	39.015	39.033	39.034	39.035	39.018
MPE	-0.563	-0.563	-0.568	-0.563	-0.618	-0.616	-0.605	-0.593
MAPE	8.090	8.090	8.095	8.090	8.097	8.097	8.097	8.092

Tableau 2 : Mesures de performance de la méthode Holt-Winter et du modèle à espace d'état sur validation

	HOLT WINTER (Additif)				ETS			
	SANS réentrainement		AVEC réentrainement		SANS réentrainement		AVEC réentrainement	
	RW	EW	RW	EW	RW	EW	RW	EW
ME	-8.617	-0.388	-1.492	-0.345	-9.126	-2.697	-7.236	-21.759
RMSE	54.372	48.807	52.818	48.816	54.471	48.866	54.164	81.946
MAE	39.276	33.705	37.997	33.696	39.314	33.663	39.126	61.745
MPE	-2.503	-0.579	-0.997	-0.566	-2.606	-1.073	-2.215	-5.945
MAPE	8.256	6.877	7.883	6.876	8.269	6.890	8.207	13.318

méthode de Holt-Winter additive fournit de meilleures prévisions en tenant compte du passé plutôt que du présent. Pour le meilleur modèle (fenêtre expanding), son réentrainement n'influe pas sur la performance. À savoir, l'amortissement a été considéré également pour Holt-Winter. Les mesures de performance n'étant pas significativement différentes, ceci ne sera pas détaillé. Afin de trouver un modèle de lissage exponentiel optimal, on se place dans une structure de **modèles à espace d'état**. La saisonnalité hebdomadaire est considérée car encore une fois, de grandes fréquences ne sont pas permises. Dans la configuration sans réentrainement du modèle, le modèle optimal est ETS(M,A,A) soit celui avec des erreurs multiplicatives, et une tendance et saisonnalité additives. Ceci est le cas même lorsque les tendances multiplicatives sont permises. La meilleure performance est atteinte pour une fenêtre expanding, avec un MAPE de 6,9 %. Dans ce cas, on note que 6,7% des observations de validation se trouvent hors de l'intervalle de confiance à niveau 95%. Il est intéressant de noter que lorsqu'on réentraîne le modèle, le modèle reste le même ETS(M,A,A) pour la fenêtre expanding mais sa performance chute drastiquement. À contrario, le modèle s'adapte mieux aux données de la fenêtre rolling et devient ETS(M,N,A), ce qui lui accorde une amélioration de sa performance avec un MAPE de 8,2 %. On étudie à présent la **méthode de Taylor**, qui permet d'accommoder deux saisonnalités. Le cycle long (saisonnalité annuelle, m_2 = 364) doit être un multiple du cycle court (saisonnalité hebdomadaire, m_1 = 7).

Tableau 3 : Mesures de performance de la méthode de Taylor sur validation ainsi que valeur de ses paramètres pour son meilleur modèle

	DSHW				Paramètres du meilleur modèle DSHW		
	SANS réentrainement		AVEC réentrainement			Initial	Au bout de 1 an
	RW	EW	RW	EW	α	0,0027	0,1171
ME	11.841	6.912	-0.155	2.519	β	0	0
RMSE	142.578	50.624	51.869	50.247	ɣ	0,0461	0,0003
MAE	48.578	36.565	36.397	35.972	ω	0,2011	0,4124
MPE	0.929	0.631	-0.750	-0.223	ϕ	0,4452	0,5373
MAPE	9.453	7.377	7.457	7.300	λ	-	-

Avec cette méthode, le réentrainement du modèle est important : pour les deux fenêtres, il permet d'obtenir de meilleurs

performances, tant pour le MAPE que le biais. L'amélioration de la performance est la plus marquée sur la fenêtre rolling, qui passe d'un MAPE de 9,5 à 7,5%. Ceci dit, le meilleur modèle obtenu est le modèle réentraîné pour une fenêtre expanding, qui obtient alors un MAPE de 7,3%. On voit que le réentrainement du modèle accorde une plus grande pondération aux observations récentes et au comportement annuel de la série. Ceci dit, il n'est pas possible de calculer un intervalle de prévision pour cette méthode. La transformation de Box-Cox a également été considérée mais n'offre pas de meilleurs résultats. Elle ne sera pas détaillée. Le dernier modèle à l'étude est le **TBATS**. Il permet d'accommoder des saisonnalités plus complexes : dans notre cas, nous avons m_1 = 7 pour la saisonnalité hebdomadaire et m_2 = 365,25 pour la saisonnalité annuelle (le 29 février est inclus dans les données). Le modèle obtenu est le TBATS $(0, \{4{,}0\}, -, \{< 7{,}3 >, < 365.25{,}5 >\})$. Trois fonctions harmoniques sont nécessaires pour la saisonnalité hebdomadaire; cinq pour la saisonnalité annuelle. Le processus ARMA est un modèle autorégressif AR d'ordre 4. Le modèle TBATS performe le mieux pour la fenêtre expanding et obtient un MAPE de 6,5%. Aussi, 7,3% des observations de l'échantillon de validation se situent hors de l'intervalle de confiance à niveau 95%. L'évaluation des méthodes de lissage permet de conclure à ce stade :

Tableau 4 : Mesures de performance du modèle TBATS sur validation

TBATS	
RW	EW
-2.688	2.675
53.226	44.932
40.977	31.990
-2.121	-0.160
8.547	6.445

- La fenêtre expanding est souvent la meilleure pour la performance des prévisions, quelle que soit la méthode ou le modèle évalué. Le passé est donc important et devrait être considéré lors de la prévision du pic de la demande horaire d'électricité à Lincoln.

- Le meilleur modèle est celui du TBATS sur la fenêtre expanding, qui permet d'obtenir une performance de prévision significativement meilleure que celle de notre benchmark, avec un MAPE de 6,5%, ce qui représente une amélioration notable de la performance de 20%.

-Le modèle de Holt-Winter se place deuxième sur la fenêtre expanding, avec un MAPE de 6,9% : la saisonnalité hebdomadaire est donc importante dans la prévision du pic de la demande.

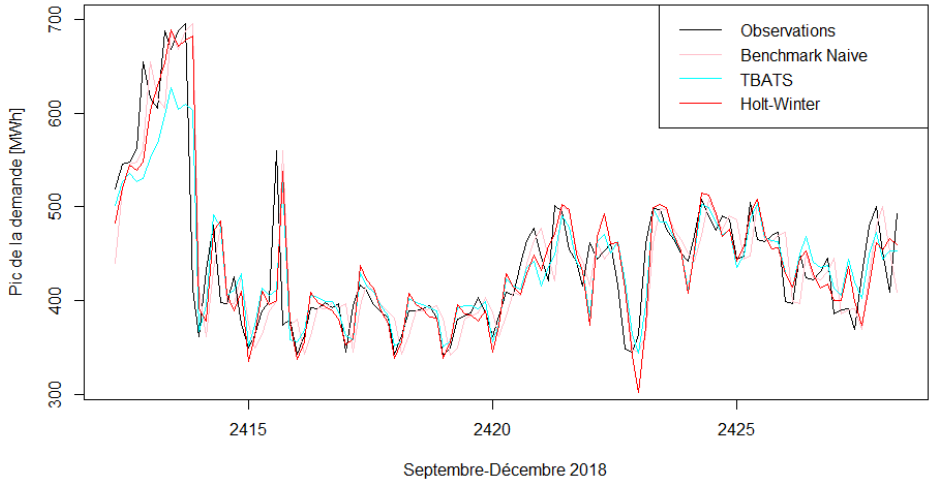

Figure 1 : Comparaison des prévisions du pic journalier de la demande sur l'échantillon de validation

Méthodes de régression

Méthodologie

Dans le cadre de l'évaluation des méthodes de régression linéaire multiple, trois approches différentes sont considérées. La première consiste à entraîner les paramètres du modèle une unique fois sur l'ensemble d'entraînement (2011-2016) et de les employer sur la validation (2017-2018). Ensuite, le modèle sera réentraîné pour chaque année de validation, soit à deux reprises, afin de mettre à jour ses paramètres pour la prévision : le deuxième modèle entraîné tiendra compte des observations de la première année de validation, ainsi que des 6 années d'entraînement (fenêtre expanding). Finalement, le modèle sera réentraîné biannuellement, soit à 4 reprises avec également une fenêtre expanding. Lors de la prévision pour le jour $t + 1$, il est important d'ajouter du bruit aux variables de températures moyennes et de vents moyens : en réalité, on a seulement accès aux prévisions et non aux observations de ces variables. Dans le cas présent, on pose l'hypothèse que les prévisions sont bonnes. Le bruit choisi suit une loi normale $N(0, 0.5)$ pour la température et $N(0, 1.5)$ pour le vent. D'autre part, les régressions linéaires aux

erreurs normales sont écartées : la structure résiduelle de ces modèles viole la condition essentielle de non corrélation des résidus. Ceci est dû à la nature chronologique de la série, qui induit une variance des résidus non constante et des erreurs ne suivant pas une loi normale. Dans ce rapport, on évaluera uniquement des modèles de régression linéaire multiple avec des erreurs ARMA, qui permettent une certaine corrélation entre les résidus. Finalement, des variables explicatives différentes seront étudiées dans plusieurs modèles ; leur p-value respective servira d'indicateur afin de déterminer si ces variables sont significatives. Dans ce rapport, seuls deux modèles sont correctement analysés : le meilleur modèle obtenu (modèle 1) et une de ses variante (modèle 2). Les autres modèles se trouvent dans le code R, et certains seront survolés, avec leurs performances sur l'ensemble de validation explicitées dans le tableau 6.

Analyse des résultats

Modèle 1 : Les variables explicatives sont le CDD, HDD, leurs lags 1 et 2, CP (refroidissement éolien avec $T_{ref} = 12°C$), des « dummy » pour les jours de la semaine (référence vendredi), les mois de l'année (référence décembre), les jours fériés, veille et lendemain du jour férié. Seuls les jours fériés dont l'effet est significatif sur le pic journalier de la demande sont considérés : le nouvel an, la fête nationale, la fête du travail, Thanksgiving et Noël. En moyenne, ces jours fériés engendrent une diminution de 61,8MWh du pic journalier, toutes choses étant égales par ailleurs. On note aussi que les coefficients estimés des mois de Janvier, Juin et Septembre ne sont pas significativement différents de 0 : les pics journaliers sont semblables au mois de référence Décembre. L'ajustement du modèle avec auto.arima indique une structure des résidus $c(1,0,0)$ qui s'exprime avec un seul coefficient $\varepsilon_t = 0{,}37\varepsilon_{t-1} + \omega_t$, pour les trois approches du réentrainement. Dans le premier cas ou le modèle est entrainé une unique fois initialement, l'étude du graphique ACF des résidus indique que le modèle de régression avec erreurs AR(1)

s'ajuste bien sur les observations. On note de minimes dépassements de l'intervalle de confiance aux lags 13, 14, 19, 20 et 23, ce qui est tout de même acceptable. Une analyse plus poussée sur la distribution des résidus standardisés entre 2011 et 2016 montre qu'ils semblent plus variables lors de la saison estivale, ce qui contredit la condition d'homoscédasticité de la variance. D'autre part, le graphique quantile-quantile semble bien ajusté au milieu mais a un comportement « heavy tailed » aux extrémités : ceci indique une grande proportion de grands résidus, positifs et négatifs. Ceci dit, ces comportements demeurent acceptables et ne sont pas disqualifiants vu le défi de taille du projet. De plus, le test Ljung-Box obtient de grandes p-values, ce qui valide l'hypothèse nulle selon laquelle les erreurs sont indépendantes. On obtient quand même de petites valeurs à partir du lag 14, ce qui est un comportement attendu vu la taille finie de l'échantillon. En terme de performance de la prévision, le modèle 1 fait très bonne figure, puisque c'est le meilleur modèle obtenu à ce stade : il obtient un MAPE de 4,1% et un biais de -1,61 MWh sur l'ensemble de validation. On cherche à comparer cette performance avec celle du cas idéal ou l'on possède les valeurs des températures et de vent au jour $t + 1$ pour la prévision du pic au jour $t + 1$. Ainsi, lorsqu'on fait abstraction du bruit, les résultats s'améliorent sans surprise avec un MAPE de 4,0%. Pour revenir à notre modèle 1 avec bruit, il est important de constater que 90% des observations sur la partie validation se trouvent dans l'intervalle de prédiction 95%. De plus, comme cela a été discuté dans l'ajustement du

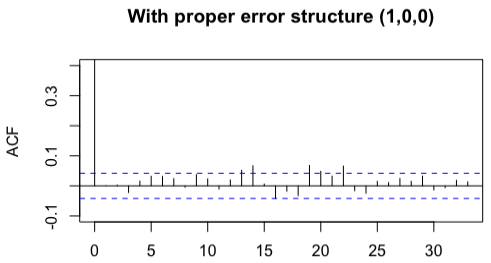

Figure 2 : Graphique ACF du modèle 1 entraîné une seule fois

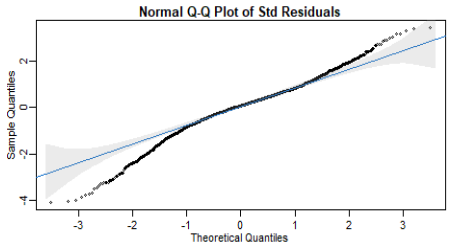

Figure 3 : Diagramme QQ des résidus du modèle 1

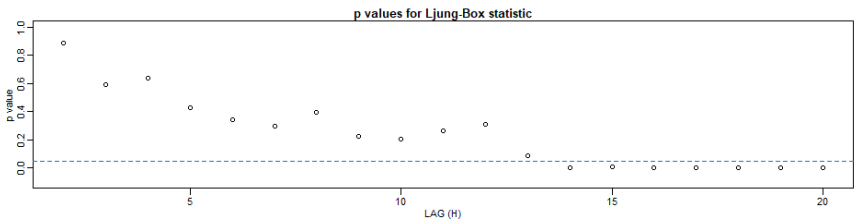

Figure 4 : Test Ljung-Box du modèle 1

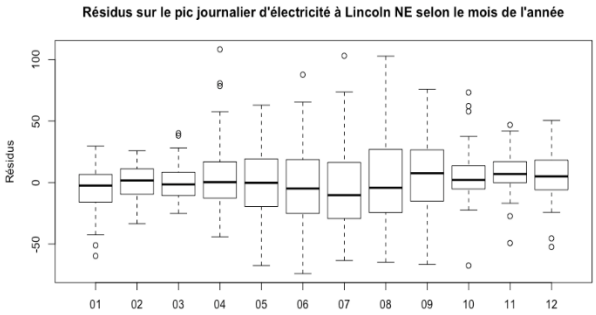

Figure 5 : Distribution des résidus standardisés sur l'ensemble de validation

modèle, on voit que les mois d'été sont associés à de plus grandes erreurs de prévision. Les résidus du mois d'Août sont caractérisés par un écart type particulièrement grand. À contrario, les mois de Février et Mars sont ceux dont l'erreur est la plus petite avec respectivement un MAE de 11,1 et 11,5 MWh, contre 31,2 en Août. On constate également qu'il n'y a pas de structure résiduelle selon le jour de la semaine ou selon l'année. À présent, on se penche vers les autres approches énoncées dans la méthodologie, soit celles ou le modèle est réentraîné à une certaine fréquence. Quelle que soit cette fréquence de réentrainement, la performance du modèle diminue progressivement. Pour un réentrainement annuel, il obtient un MAPE de 4,3% ; pour un réentrainement biannuel, il obtient un MAPE de 4,4%. Ainsi, les observations de validation n'ajoutent pas de pouvoir prédictif au modèle, que ce soit la première année entière ou par parties de 6 mois. Il est important de souligner que nous avons également considéré un ensemble d'entrainement ayant une fenêtre rolling de 6 ans et qu'avec cette approche, les résultats restent sensiblement les mêmes : ils ne seront pas détaillés dans le rapport.

Tableau 5 : Mesures de performance du modèle 1 de régression sur validation

Réentrainement	RMSE	MAPE	MAE	ME
Unique initial	27,09	4,09	19,83	1,61
Annuel	27,6	4,26	20,32	6,49
Biannuel	28,75	4,41	21,40	-0,83

Modèle 2 : Nous avions mentionné que certains mois n'étaient pas statistiquement significatifs dans le modèle précédent. Dans le modèle 2, on considère une autre approche pour tenter de saisir la saisonnalité annuelle : les variables « dummy » de saison sont introduites (référence printemps) à la place de celles des mois. L'ajustement du modèle sur l'échantillon d'entrainement permet d'obtenir des coefficients estimés tous significativement différents de zéro. Les analyses sur les résidus sont très similaires à celles effectuées sur le modèle 1 : les conditions de non-corrélation, homoscédasticité et normalité des résidus ne sont pas parfaitement respectées mais sont jugées acceptables. La structure résiduelle obtenue n'est certainement pas meilleure. La performance de ce modèle 2 se traduit par un MAPE de 4.5% : ce modèle est moins bon que le

premier. Les variables des mois sont donc plus pertinentes que celles des saisons pour la prévision. Aussi, les réentrainements de ce modèle ne permettent pas d'améliorer sa performance, quelle que soit leur fréquence. Au final, les mois d'été sont toujours les plus difficiles à prédire.

Autres modèles : Le modèle 3 comporte les mêmes variables que le modèle 1, auxquelles on a rajouté la variable « dummy » des tornades. Le modèle 4 représente notre modèle simple de référence. Il contient uniquement les variables de températures (CDD, HDD, lag 1 et 2), le jour de la semaine, les jours fériés et le vent (sans le CP). Bien que la ville de Lincoln soit soumise à des climats propices aux tornades, on voit que cette variable n'a pas de pouvoir dans la prévision du pic de la demande journalière d'électricité : le modèle 1 et 3 ont des performances identiques. Aussi, on peut remarquer que le modèle 4 est trop simple, puisqu'il obtient un MAPE de 5,3% et un biais de 8,2 MWh : plusieurs autres variables sont significatives pour la prévision. Il s'agit, entre autres, des mois de l'année et du refroidissement du vent. Plus d'une quinzaine de modèles différents ont été testés dans le code R mais ne sont pas discutés ici en raison de leur piètre performance. Pour conclure concernant les méthodes de régression :

Tableau 6 : Résumé des performances de l'ensemble des modèles de régression sur validation

Modèle	RMSE	MAPE	MAE	ME
1	27,09	4,09	19,83	1,61
2	29,21	4,50	21,42	6,22
3	27,09	4,09	19,83	1,61
4	34,62	5,25	26,17	-8,24

- Le réentrainement du modèle semble nuire à la qualité de la prédiction, dans notre cas. De plus, le type de fenêtres utilisé (rolling ou expanding) ne semble pas avoir d'incidence sur la performance des prévisions.

- Le modèle 1 de régression aux erreurs ARMA proposé est le meilleur modèle pour la prévision du pic de la demande journalière d'électricité à Lincoln. Son MAPE de 4,1% représente une amélioration notable de la performance de 50% par rapport à notre benchmark naïf. Il représente notre candidat pour les méthodes de régression.