MATH 60638 – Méthodes de prévision

Projet de semestre – Partie 1

Présenté à

Pre. Debbie Dupuis

Par l'Équipe G

Abderezzak AMIMER – 11241872

Maryvonne ANGELO – 11286237

Franck BENICHOU - 11249092

Théo LEFEBVRE – 11314432

Joe YOUNES – 11285603

Le 10 février 2022

À Montréal

Ce rapport est la première partie du projet de **prévision de la somme de la demande horaire d'électricité pour jour t +1**, dans la zone du Texas composée des régions **contigües** Nord, Centrale Nord, et Est, desservies par le *Electric Reliability Council of Texas* (ERCOT). Nous passerons par une brève introduction d'ERCOT et des régions en analyse, une analyse exploratoire des données, une description des variables explicatives et leur source et l'évaluation des méthodes naïves selon des métriques d'erreur de prévision.

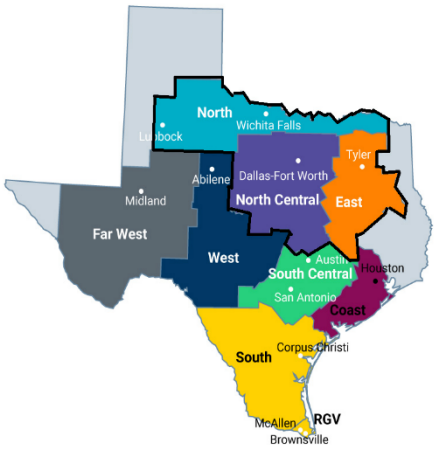


Figure 1 ERCOT Weather Zone Map

Introduction

ERCOT est un operateur électrique à but non lucratif, qui gère le flux électrique de 26 millions de clients au Texas, représentant 90 % des besoins en électricité. Son système couvre 75% du territoire texan et ses membres incluent des particuliers, des coopératives, des municipalités et des producteurs d'électricité[1]. En 1970, ERCOT se forme pour répondre aux exigences fédérales. Cependant, la grille électrique texane n'est pas soumise à la régulation fédérale et se trouve isolée des interconnections Est et Ouest américaines (bloquant toute fourniture extérieure majeure en électricité vers le Texas)[2]. ERCOT a été frappé de coupures d'électricité en février 2011 et du 13 au 17 février 2021, dans les deux cas, à la suite d'une vague de froid extrême[3]. Un rapport de la NERC de 2019 alertait déjà que le réseau ERCOT avait l'une des réserves de marges anticipées

[1] Company Profile (ercot.com)
[2] https://en.wikipedia.org/wiki/Electric_Reliability_Council_of_Texas
[3] https://en.wikipedia.org/wiki/Electric_Reliability_Council_of_Texas

les plus basses aux USA, l'empêchant de répondre aux pics de demande d'électricité pendant les grands froids et l'été, notamment en 2011 et 2021[4]. Le cœur économique de nos trois régions se trouve dans la région *North Central* représentant près de 90% de la demande d'électricité, en particulier dans le centre urbain de Dallas-Fort Worth. *North Central* est plus résidentielle et urbaine que les deux autres régions[5]. Selon le *Comptroller General* du Texas, la population de la métropole de Dallas a cru de 19% sur la période 2010-2019[6]. Selon le *US Bureau of Economic Analysis* (BEA), le PIB du comté de Dallas était de 239 milliards de dollars en 2020[7], avec une population de 2,6 millions d'habitants[8]. Le secteur tertiaire y est aussi important que les secteurs manufacturiers, du pétrole et du gaz, et les industries de l'aviation et de l'aérospatiale [9]. La région *North* a un pourcentage d'activités agricoles et industrielles plus important que les deux autres régions. Les populations et PIB des deux grandes villes de *North*, soit Lubbock et Wichita Falls, sont respectivement de 310,000 habitants[10] et 12 milliards de dollars[11], pour l'un, et de 129,500 habitants et de 5 milliards de dollars pour l'autre (2020). Ces 2 métropoles ont cru de 11% et 0.5% respectivement de 2010 à 2019. La région Est est spécialisée dans le secteur pétrolier et gaz de schiste. La ville principale y est Tyler avec 230,000 habitants et un PIB de 6 milliards de dollars (2020) a vu sa population croître de 11% de 2010 à 2019. Le Texas accueille 14 bases militaires qui emploient directement 226,000 personnes[12]. Nos régions d'étude en accueillent 2 à JRB Fort Worth et à Wichita Falls[13]. Selon le *US Energy Information Administration* (EIA), la majeure

[4] https://en.wikipedia.org/wiki/Electric_Reliability_Council_of_Texas
[5] https://worldpopulationreview.com/us-counties/states/tx
[6] https://comptroller.texas.gov/economy/economic-data/regions/2020/texas.php
[7] https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=5#reqid=70&step=1&isuri=1&acrdn=5
[8] https://www.texas-demographics.com/counties_by_population
[9] https://realestate.usnews.com/places/texas/dallas-fort-worth/jobs#:~:text=In%20the%20Dallas%20area%2C%20the,slightly%20below%20the%20national%20average
[10] https://www.texas-demographics.com/counties_by_population
[11] https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=5#reqid=70&step=1&isuri=1&acrdn=5
[12] https://comptroller.texas.gov/economy/economic-data/regions/2020/texas.php
[13] https://militarybases.com/texas/

partie de la demande d'électricité de l'État vient du secteur résidentiel, où 3 ménages texans sur 5 utilisent l'électricité pour le chauffage et l'aération[14]. Le chauffage en hiver et l'aération en été sont deux des raisons principales de consommation d'électricité au Texas. L'EIA note que les pics de demande d'électricité journaliers sont atteints en été à cause de l'augmentation de l'utilisation de l'aération[15]. La région *East* possède un climat subtropical avec quelques vagues de froid du Nord, des précipitations de 890 à 1500 mm/an, des plaines côtières et des collines dans les terres[16]. Les régions *North* et *North Central* ont un climat subtropical / continental avec des précipitations entre 700 et 1200 mm/an. Ces deux régions sont des plaines et subissent de nombreuses tornades au printemps car elles sont dans la *tornado alley*. *North* et *North Central* reçoivent des vents froids du Nord et chauds du Golfe du Mexique. Les hivers y sont tempérés mais parfois ponctués de vague de froid extrême en janvier ou février. Les trois régions connaissent des hivers doux avec des étés chauds voir tropicaux, menant à des vagues de chaleur[17].

Analyse exploratoire des données

L'analyse porte sur l'évolution de la demande journalière d'électricité des régions étudiées de 2012 à 2021. Nous observons que la région *North Central* présente la plus grande demande d'électricité **(Figure 2)** car celle-ci est la plus peuplée et industrielle des 3 régions. De 2012 à 2021, la moyenne de la demande journalière a

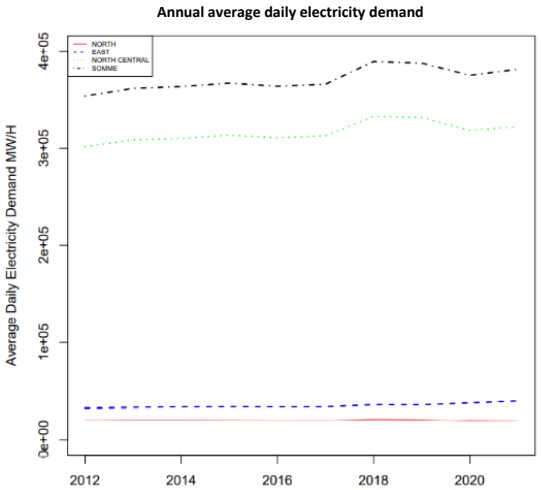


Figure 2 Graphique de la moyenne annuelle de la demande d'électricité journalière 2012-2021

[14] https://www.eia.gov/state/analysis.php?sid=TX
[15] https://www.eia.gov/state/analysis.php?sid=TX
[16] https://en.wikipedia.org/wiki/East_Texas
[17] https://en.wikipedia.org/wiki/North_Texas

augmenté de 7.8 % dans l'agrégat des 3 régions (*SOMME*), de 6.8% dans la région *North Central* (dû à la croissance économique et démographique de Dallas), de 22% dans la région *East,* et diminué de 1% dans la région *North*. La stagnation dans *North* peut être due à une désindustrialisation relative et une croissance moins importante de la population. *East* observe une augmentation sur la décennie probablement à cause du développement du commerce de gaz de schiste et l'apparition de nouvelles raffineries. En **Figure 2,** de 2012 à 2017, la croissance de la moyenne annuelle de la demande journalière d'électricité oscille entre 0 à 3%. L'année 2018 voit une croissance de la demande annuelle d'électricité de 6% grâce au boom économique et à la production pétrolière record. S'ensuit un déclin à partir de 2019, accentué par la Covid-19 pour toutes les régions à l'exception de *East*.

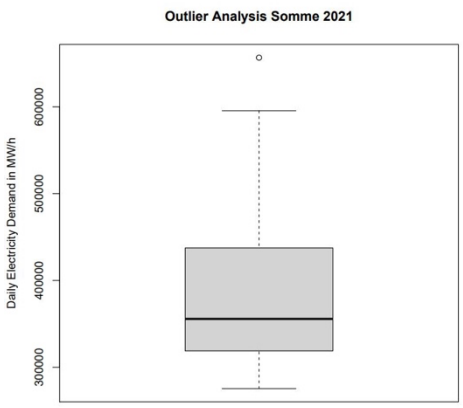


Figure 3 Boxplot de la demande journalière d'électricité en 2021

À travers la composante de saisonnalité de la décomposition *stl,* nous observons une saisonnalité annuelle des pics de de la demande journalière en hiver (décembre à mars le pic le moins grand) et été (juin-août le pic le plus grand des deux) dans les 3 régions. Cette saisonnalité hiver/été est plus prononcée en *North Central*, région plus sensible aux changements de température du fait de sa population plus importante, qu'en *North* et *East* où la part des secteurs primaires et secondaires

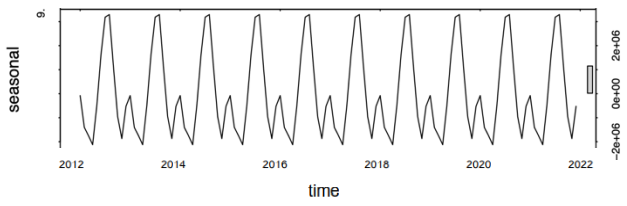


Figure 4 Graphique de la décomposition stl de la demande journalière d'électricité 2012-2021

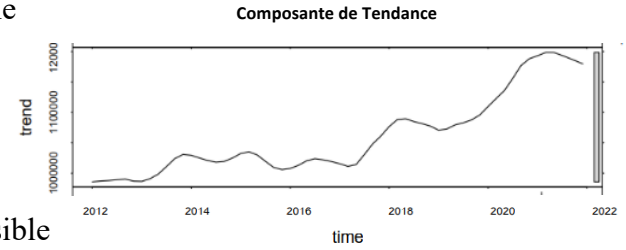
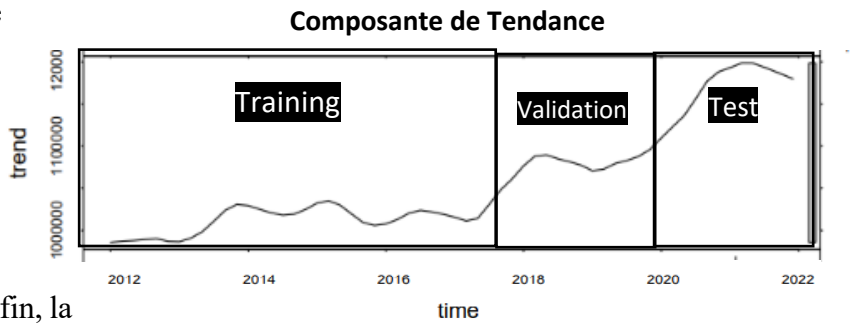


Figure 5 Graphique de la tendance de la demande journalière d'électricité 2012-2021

(moins sensible aux changement de température) est plus importante. Globalement, sur les 3 régions, la tendance sur la période entière est à la hausse. Post crise financière, la tendance est à croissance modérée avant de croître plus rapidement à partir de 2017. Cette tendance à la hausse est soutenue par *North Central et East.* Manifestement, les données sont non-stationnaires. L'analyse des *IQR* avec les boxplots nous montre via la figure 3 une seule valeur aberrante dans l'aggrégat des 3 régions en février 2021, lors de la tempête hivernale extrême. La région North présente 2 valeurs aberrantes en 2017 et 2021 respectivement. North Central en présente 1 en février 2021 (tempête hivernale) de même que la région East.

En dessaisonalisant la série chronologique, nous aurions pu améliorer cette analyse. Aucun traitement n'est fait, pour le moment. Enfin, la



separation des ensemble d'entraînement, de validation, et de test se fait selon le critère de cohérence dans la dynamique de marché de la période. Suivant la tendance de la décomposition *stl,* nous choisissons une période d'entraînement allant de 2012 à 2017. Les périodes de validation et de test vont respectivement de 2018 à fin 2019, et de 2019 à fin 2021.

Variables explicatives

Afin d'améliorer la qualité des prédictions, il sera primordial d'explorer des modèles utilisant des variables explicatives Pour les choisir, la linéarité de ces variables avec nos données de demande journalière a été analysée.

Les variables explorées sont soit météorologiques ou relèvent de la saisonnalité. Les données de températures proviennent du site web du 'National Oceanic and Atmospheric Administration'

(NOAA). Elles sont journalières et n'ont aucune donnée manquante pour la station de Dallas FAA Airport[18]. Cette station a été stratégiquement sélectionnée car elle est au centre des trois régions à l'étude et elle se trouve dans la région où la demande en électricité est la plus importante. Elle a fourni les variables de températures journalières moyennes, maximales et minimales en degrés Celsius. NOAA a également mis à notre disposition une variable de vent qui a été transformée en effet de refroidissement afin d'obtenir une meilleure relation linéaire avec la demande d'électricité journalière. Les dernières données météorologiques sélectionnées sont celles de l'humidité relative qui proviennent de TexMesonet[19]. L'humidité relative n'est pas linéaire par rapport à la demande, toutefois, lorsqu'une interaction est appliquée avec la saison ou la température, elle devient plus linéaire par rapport à la demande journalière d'électricité.

La température a une relation quadratique avec la demande d'électricité d'où la nécessité de la transformer en relation linéaire avec un HDD (Heating degree-days) et un CDD (Cooling degree-days). Le HDD exprime le besoin d'utiliser le chauffage lorsqu'il fait froid tandis que le CDD exprime le besoin d'utiliser l'air climatisé lorsqu'il fait chaud. La Figure 6 en fait la démonstration.

[18] Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC) (noaa.gov) et
[19] TexMesonet | Custom Datasets

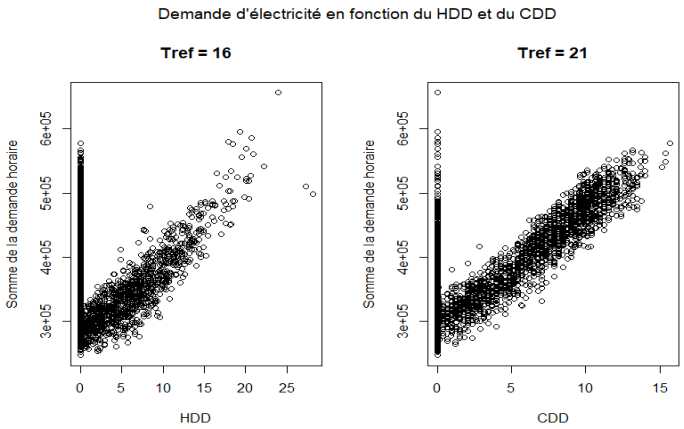


Figure 6 Graphiques de la demande d'électricité en fonction du HDD et CDD

Ce graphique montre également que la meilleure linéarité a été obtenu en utilisant un T_{ref} de 16 °C pour le HDD et un T_{ref} de 21 °C pour le CDD. La sélection du T_{ref} s'est faite grâce à des visualisations similaires en utilisant plusieurs valeurs de température de référence sur l'intervalle du minimum au maximum de $T_t = (T_{min} + T_{max})/2$. Aussi, le T_t est habituellement une moyenne pondérée, mais l'exploration effectuée a montré que la linéarité la plus marquante est atteinte grâce à $(T_{min} + T_{max})/2$.

La deuxième transformation est celle de la moyenne de la vitesse du vent journalière en effet de refroidissement. Lorsque celle-ci est directement mise en relation avec la demande, aucune linéarité n'en ressort. Toutefois, avec la transformation proposée dans la section 6.2.1 des notes de cours, une linéarité est observée. Le nuage de point de la figure 8 le démontre assez bien.

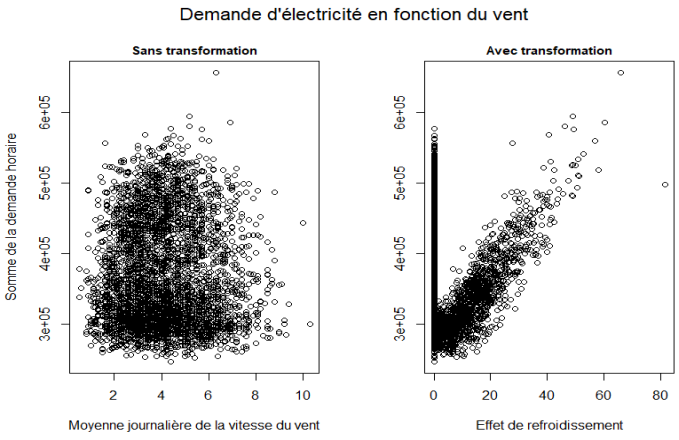


Figure 7 Graphiques de la demande journalière d'électricité en fonction du vent

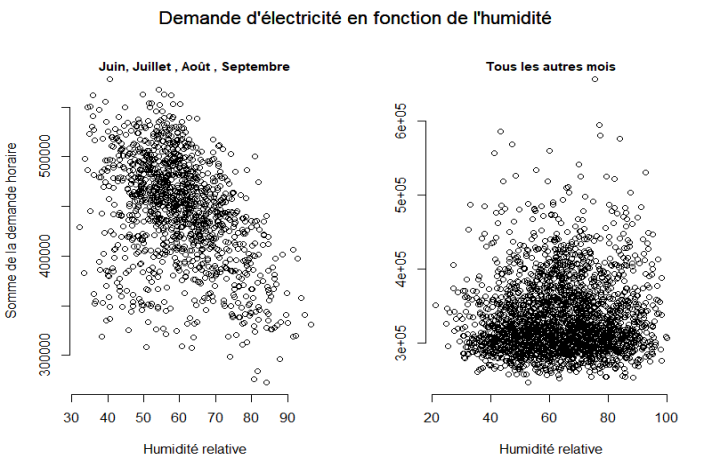


Figure 8 Graphique de la demande journalière d'électricité en fonction de l'humidité relative

La troisième variable météorologique d'intérêt est celle de l'humidité relative.

Pour améliorer la linéarité, il a fallu observer la relation entre l'humidité et la demande au cours de différents mois. Ceux pour lesquels la relation est la plus marquante sont les mois les plus chauds. La figure 9 montre la différence entre juin, juillet, août et septembre avec les autres mois de l'années. Comme il semble y avoir une amélioration, la variable humidité sera retenue.

D'autres variables météorologiques telles que les précipitations, le brouillard, la direction du vent et bien d'autres ont également été considérées mais n'ont pas été sélectionnées en raison d'une absence de relation linéaire avec la demande horaire d'électricité.

Une variable indiquant les différents mois de l'année a été ajoutée dû à la variation de la demande observée selon les mois de l'année en figure 9.

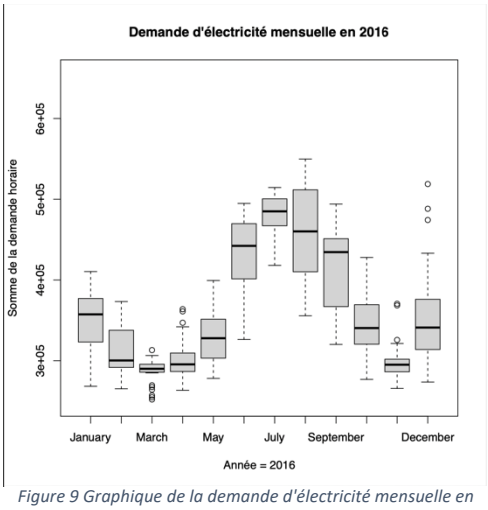


Figure 9 Graphique de la demande d'électricité mensuelle en 2016

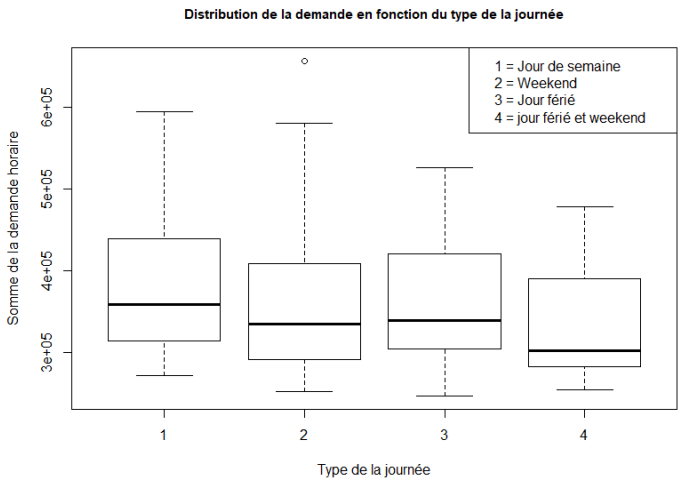


Figure 10 Boxplots de la distribution de la demande en fonction du type de journée

Une variable indiquant les fins de semaine a été créée car une légère baisse de la demande d'électricité est observée durant le weekend comme présenté dans la figure 11. Cette baisse est due à la fermeture de la plupart des entreprises et des bureaux durant le weekend.

De plus, une variable indicatrice des jours fériés au Texas[20] a été créée. On observe sur la figure 10 une diminution de la demande d'électricité durant les jours fériés car les entreprises et bureaux

[20] https://www.officeholidays.com/countries/usa/texas/2021

sont fermés durant ces jours. Au Texas lorsqu'un jour férié tombe en fin de semaine, il n'est pas déplacé à une autre date. Une variable binaire indiquant les fins de semaine qui correspondent à des jours fériés a été également créée car la demande des fins de semaine avec des jours fériés est moins important que celle des fins de semaine sans jours fériés.

Transformation des données

Afin de pouvoir travailler avec nos données, nous avons dû effectuer plusieurs modifications au jeu de données *ercotdata.RData*. En effet, nous avons commencé par modifier la date et l'heure du jeu de données à son équivalent en *Central Standard Time* CST (heure du Texas). Avec cette transformation, nous avons des données incomplètes pour le 31 décembre 2021, donc nous l'avons enlevé en entier. Nous avons ensuite regardé les valeurs manquantes. Il y en avait une seule pour les trois régions assignées en date du 6 novembre 2016 à 18h00. Nous avons imputé les valeurs manquantes par la moyenne de l'observation qui précédait et l'observation qui suivait. Finalement, nous avons pu agréger les données par journée pour retrouver la somme de la demande pour jour *t* dans une nouvelle colonne *Somme*.

Évaluation des méthodes naïves

Pour ce qui est de l'évaluation des méthodes naïves, nous voulions évaluer les méthodes suivantes : *Naïve no change*, *Naïve seasonal 7 jours*, *Moyenne mobile 3 jours* et *Moyenne mobile 7 jours*. Les méthodes naïves ont été évaluées sur les données de validation (2018-2019) puisqu'elles serviront de *benchmark* afin de comparer les modèles et les méthodes que nous allons utiliser dans les prochaines parties. La performance de ces derniers sera calculée sur l'échantillon de validation.

Le *naïve no change,* qui consiste à prendre l'observation de la veille pour notre prévision au temps t+1, est un bon point de départ pour nos recherches. Ensuite, nous voulions essayer le *naïve seasonal 7 jours* puisque hypothétiquement, la consommation d'électricité les jours de la semaine devrait être similaire d'une semaine à l'autre, cependant il y aura plusieurs facteurs que cette méthode ne prendra pas en compte, telles que les conditions météorologiques, et plusieurs autres variables explicatives. Par ailleurs, nous trouvions qu'il ne valait pas la peine d'essayer une méthode naïve avec saisonnalité de 365 jours puisque beaucoup trop de facteurs diffèreraient entre une date à l'année t et la même date à l'année t-1, par exemple le jour de la semaine serait différent et les conditions météorologiques seraient très probablement différentes aussi. Ensuite, les moyennes mobiles de 3 et 7 jours ont été retenues car nous voulions lisser la prédiction et réduire l'effet des irrégularités sur une petite période, 3 jours, et ensuite sur une plus longue période, 7 jours. On s'attendait en revanche à ce que les prévisions avec la moyenne mobile sur 7 jours soit beaucoup plus lisse que celles de la moyenne mobile à 3 jours et aussi que toutes les autres. Les moyennes mobiles tendent à réagir beaucoup plus lentement aux changements que les autres méthodes, puisqu'on prend la moyenne. Un détail additionnel : nous avons ajusté une marche aléatoire avec horizon $h = 1$ avec un drift, mais les prévisions étaient presque identiques au *naïve no change*, alors nous l'avons exclu de nos méthodes à évaluer.

Voici les différentes mesures de performance obtenues :

	Biais	RMSE	MAE	% Biais	MAPE
Naïve no change	223.90	33368.41	24338.33	0.42	6.34
Naïve seasonal 7 jours	1084.76	61970.88	46589.41	1.55	12.22
Moyenne mobile 3 jours	418.15	41554.12	31104.78	0.86	8.21
Moyenne mobile 7 jours	301.37	43909.99	33152.14	1.10	8.74

Figure 11 Mesures de performance des méthodes naïves

Concrètement, le *Naïve no change* gagne haut la main, avec des mesures de performance dominantes dans toutes les catégories. Ensuite, les moyennes mobiles 3 jours et 7 jours retournent des résultats similaires, mais la moyenne mobile 3 jours semble mieux performer que celle de 7 jours puisque la moyenne mobile 7 jours ne performe mieux qu'au niveau du biais. Le *naïve seasonal 7 jours* retourne les pires résultats pour toutes les mesures de performance. De plus, il est intéressant de noter que tous les biais sont positifs, i.e. que toutes les méthodes vont avoir tendance à faire des prévisions trop hautes par rapport aux observations. Dans notre contexte, si on avait à choisir entre un biais positif ou négatif, nous préférons avoir des biais positifs puisqu'il est mieux de prévoir plus d'électricité à offrir que pas assez, afin d'éviter les blackouts comme dans le passé. Ultimement, le *Mean Absolute Percent Error* (MAPE) nous montre que le *Naïve no change* effectue en moyenne des prévisions différentes à 6,34% des observations, tandis que la moyenne mobile 3 jours et la moyenne mobile 7 jours effectuent à leur tour, en moyenne, des prévisions différentes des observations à 8,21% et 8,74% respectivement. Le *naïve seasonal 7 jours* de son côté se trompe en moyenne à une différence de 12,22% des observations.

Pour analyser plus spécifiquement les erreurs, nous sommes allés regarder les résidus de notre meilleure méthode, le *naïve no change*. Les résidus sont définis comme étant la différence entre l'observation et la prévision au jour *t*.

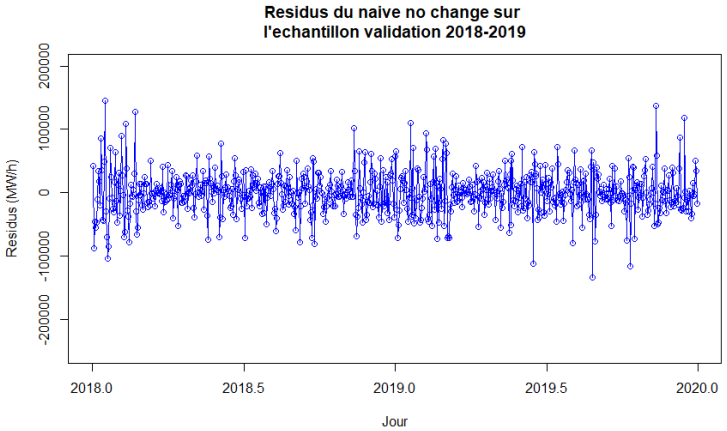


Figure 12 Résidus du naïve no change sur l'échantillon validation 2018-2019

En regardant les résidus, nous avons remarqué que cette méthode a tendance à plus se tromper durant certaines saisons que d'autres,

plus spécifiquement en hiver, dans le sens où l'amplitude et la variance des erreurs augmente durant cette saison. Nous avons donc fait une analyse du MAPE durant les différentes saisons.

	Hiver	Printemps	Été	Automne
Naïve no change	8.12	5.64	4.90	6.79
Naïve seasonal 7 jours	17.08	8.37	10.23	13.46
Moyenne mobile 3 jours	10.89	6.75	5.97	9.32
Moyenne mobile 7 jours	11.51	6.66	7.08	9.84

Figure 13 Mesure de performance MAPE sur les différentes saisons

En observant les performances des différentes méthodes par saison, nous pouvons voir que l'hiver est une saison plutôt difficile à prédire et ce, peu importe la méthode utilisée. Ce phénomène est dû au fait qu'il y a de grandes variations dans la demande . Ces changements doivent être très reliés à la variation de la température en hiver. Les méthodes naïve ne peuvent pas prendre cela en compte. Néanmoins, la méthode naïve *no change* reste *la plus* efficace à travers toutes les saisons de l'année. Le *naïve seasonal 7 jours* performe très mal à travers toutes les saisons, probablement parce que les changements de semaine en semaine sont trop importants pour qu'ils soient capturés en regardant la même journée de la semaine précédente. Les méthodes utilisant des moyennes sont légèrement meilleures que le *seasonal 7 jours*, car elles incorporent la journée précédent le jour *t* qui semble être la plus importante pour la prédiction.