Projet R - Prédiction de la consommation éléctrique

Camille Palmier - Arnaud Valladier

10 Mars 2016

Présentation

Nous souhaitons prévoir la consommation en éléctricité de l'année 2008 aux Etats-Unis à partir des données recoltées au court des années 2004 à 2007. Nous concentrerons nos efforts sur la zone 2 du jeu de données fourni.

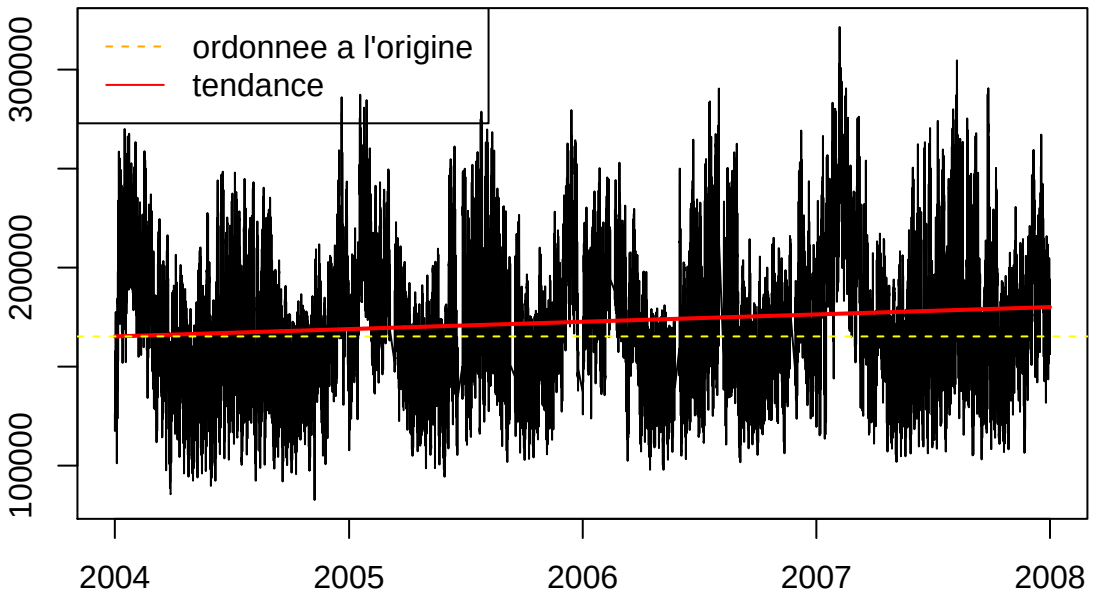
Dans un premier temps, nous allons effectuer une analyse descriptive afin de mettre en lumière la présence d'une tendance et de plusieurs saisonnalités emboîtées. Nous exhiberons également les variables explicatives de la consommation, telles que la température de l'air (données météorologiques), l'heure de la journée, le mois, le jour de la semaine etc.

Dans un deuxième temps, nous construirons un modèle de prévision par régression sur les variables explicatives. Ces observations nous permettront de bâtir un modèle de régression. Pour tester ce modèle, nous découperons notre jeu de données en deux. Les années 2004 à 2006 seront nos données d'apprentissage, et 2007 sera notre année test.

À chaque régression, nous évaluerons le taux d'erreur entre les données réelles de 2007 et celles prédites par notre modèle de régression. Nous ne bâtirons pas notre modèle sur toutes les variables explicatives, nous chercherons un modèle qui prédise le mieux possible avec le moins de données possibles. Ces variables seront sélecionnées en partant d'un modèle très simple puis en l'ettoffant au fur et à mesure en sélectionnant des variables qui amèliorent sensiblement la qualité de la prédiction.

1.Analyse desciptive des données

Les données sont issues de relevés de consommation éléctrique et de température effectuées toutes les heures pendant 4 ans. Jetons un oeil sur les données de la zone 2 :



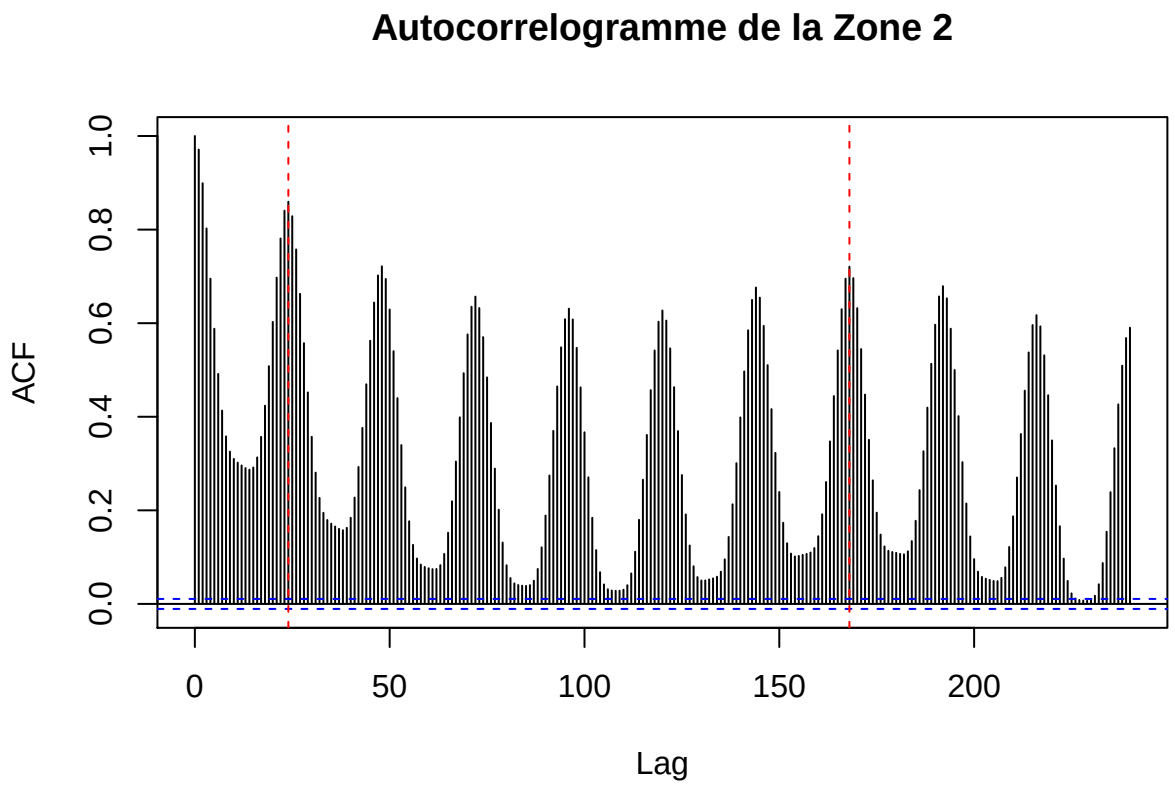
???? Tendance à l'origine ????

Nous avons 33720 relevés. Nous remarquons que certains intervalles de données sont manquants. Nous tâcherons de compléter ces relevés à l'aide d'un modèle de prédiction.

On constate une faible tendance à la hausse ainsi qu'une saisonnalité annuelle. La consommation éléctrique va dépendre des températures et des indices de luminosité et d'humidité. Les consommations élevées en hiver doivent correspondre à l'utilisation du chauffage et de l'éclairage éléctrique provoqués par des températures faibles et une luminosité faible. Les consommations élevées en été doivent correspondre à l'utilisation de la climatisation provoquée par des températures hautes et un taux d'humidité possiblement élevé.

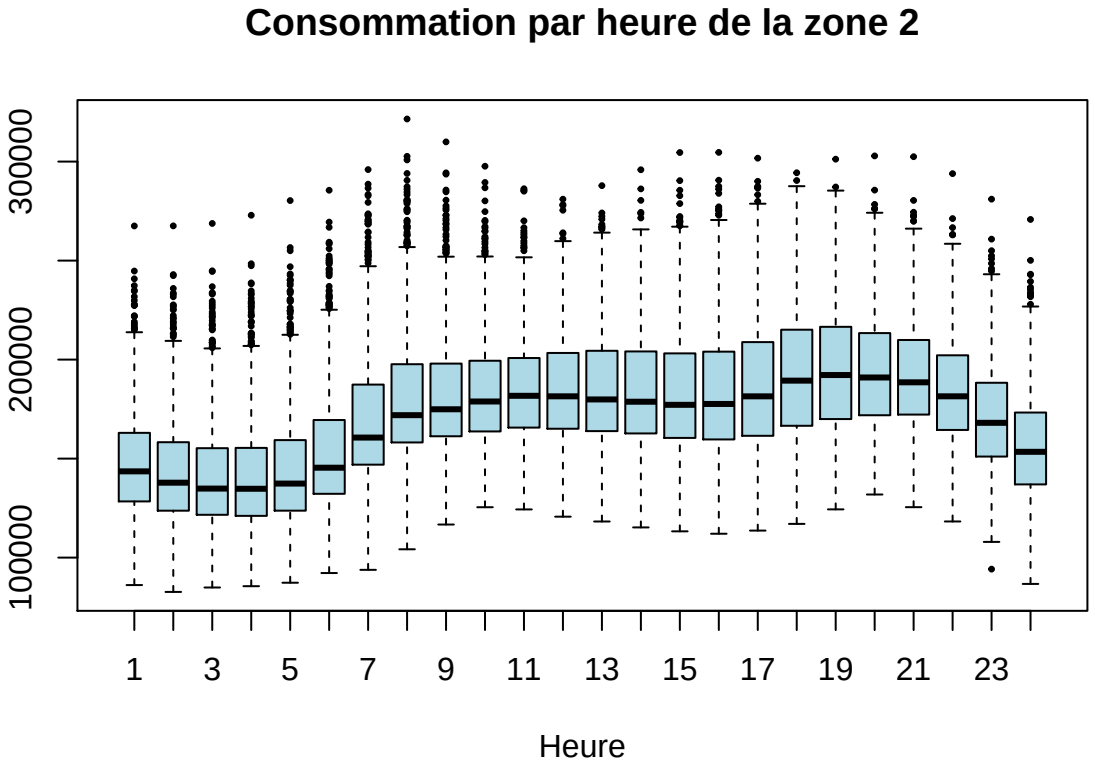
Ne connaissant pas la position géographique de la zone sur laquelle nous travaillons, nous n'avons pas pu compléter nos données avec les indices de luminosité et d'humidité. Nous allons cependant utiliser l'influence des températures pour nos prévisions.

Afin d'avoir une idée précise des saisonnalités, nous avons tracé l'autocorrélogramme de nos données :



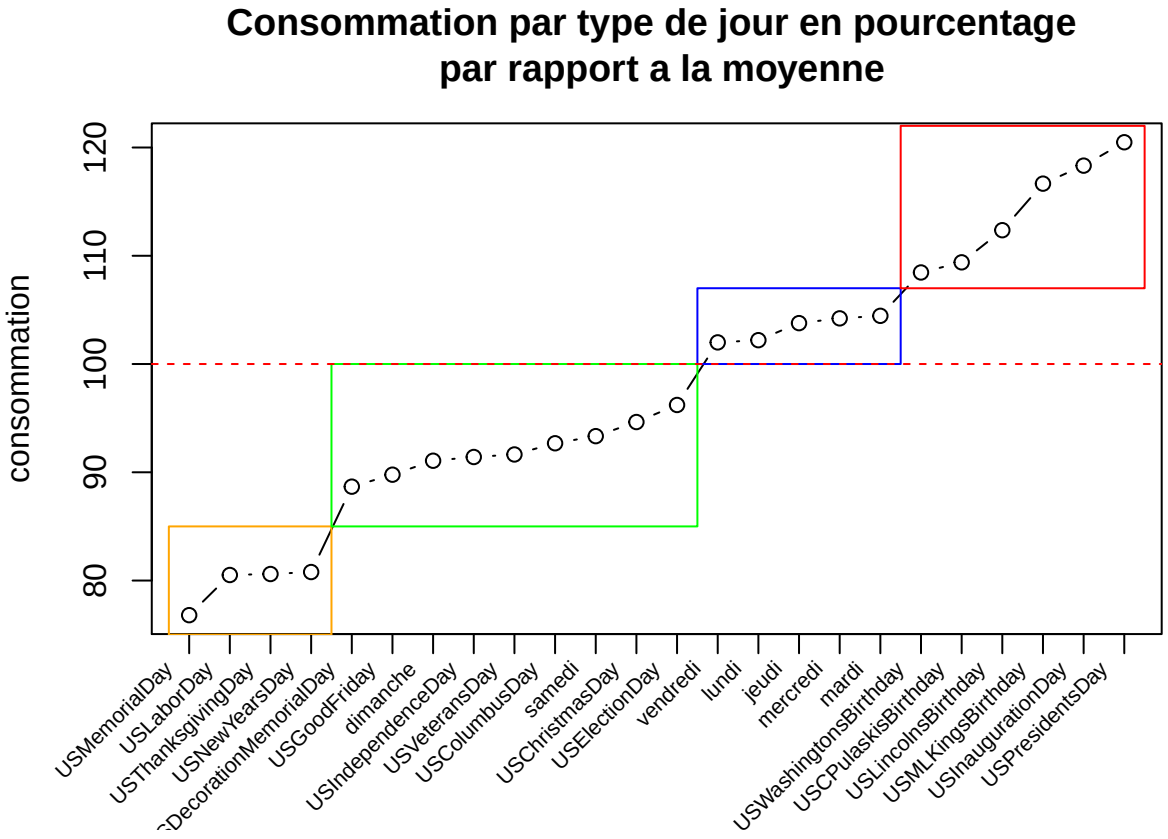
Ce graphique nous a permis de confirmer qu'il y a peu de tendance dans nos données. Le graphe de l'autocorrélation partielle confirme également une double saisonnalité : par heure et par jour.

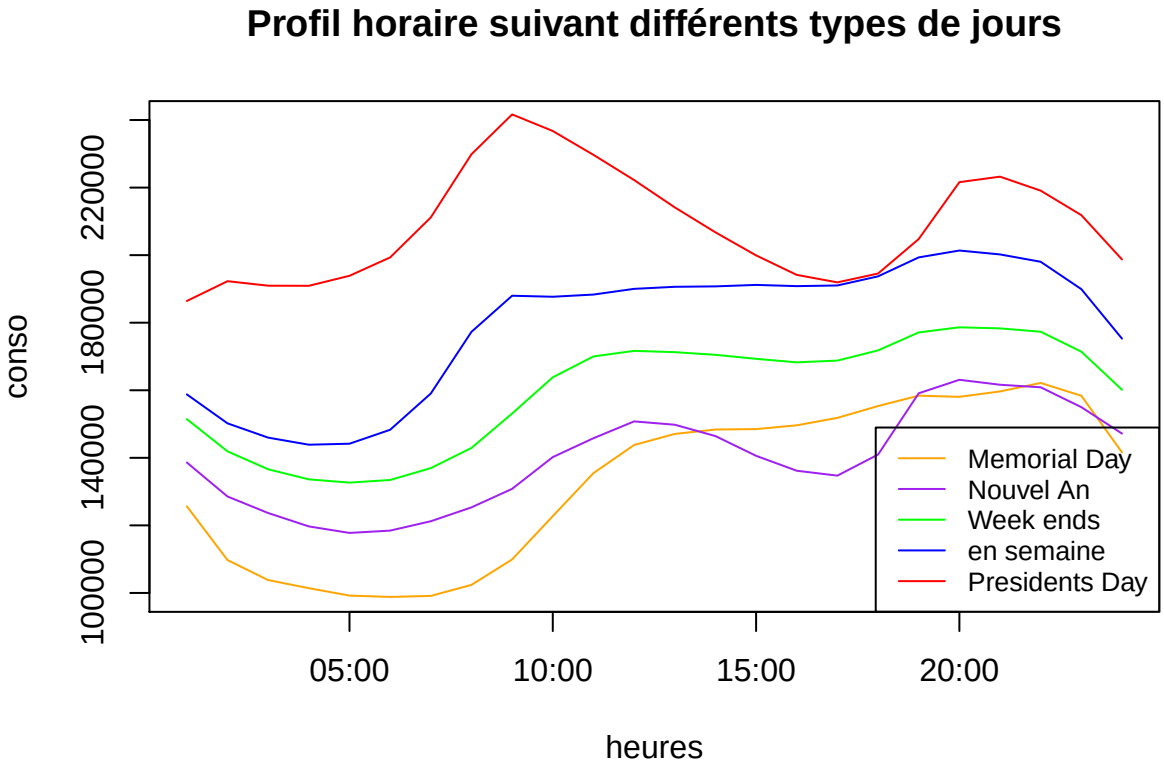
Annalysons maintenant nos données à l'aide de certaines statistiques descriptives. Nous allons regarder l'influence des différentes échelles temporaires (heure, jour, mois).



Nous constatons une évolution de la consommation par heure ainsi que de grands écarts de consommation suivant les heures de la journée. Nous allons devoir prendre en compte l'influence de cette variable dans nos modèles de prédiction.

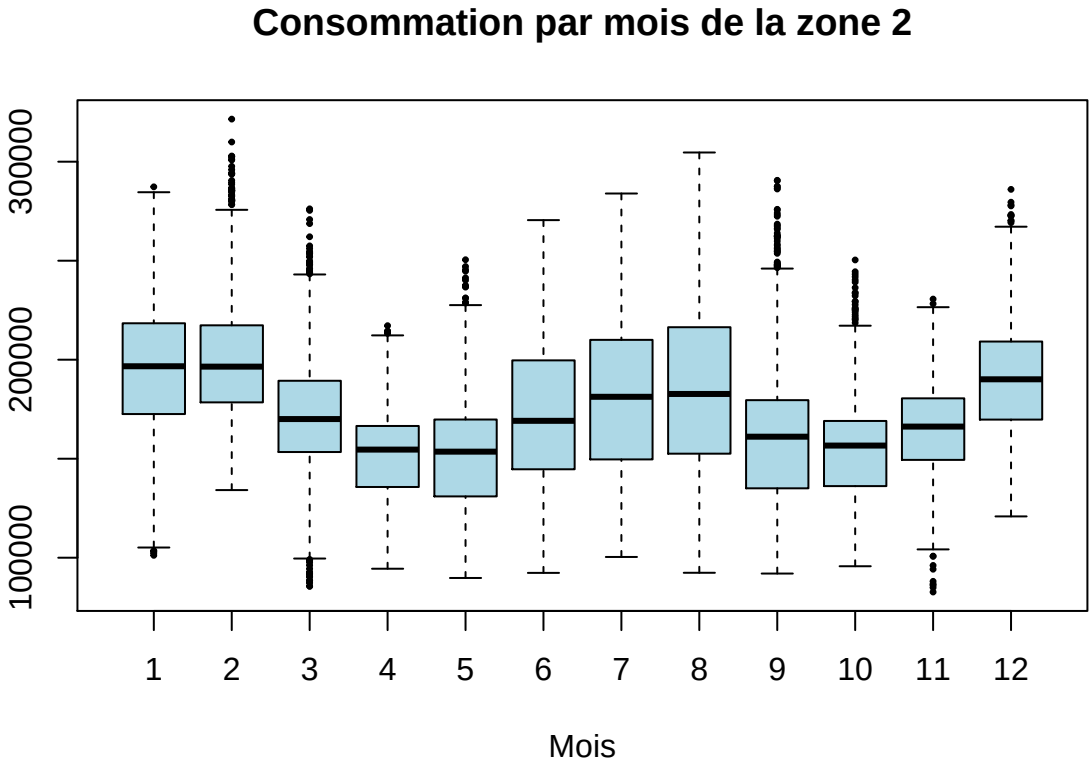
Regardons à present la consommation selon les jours :





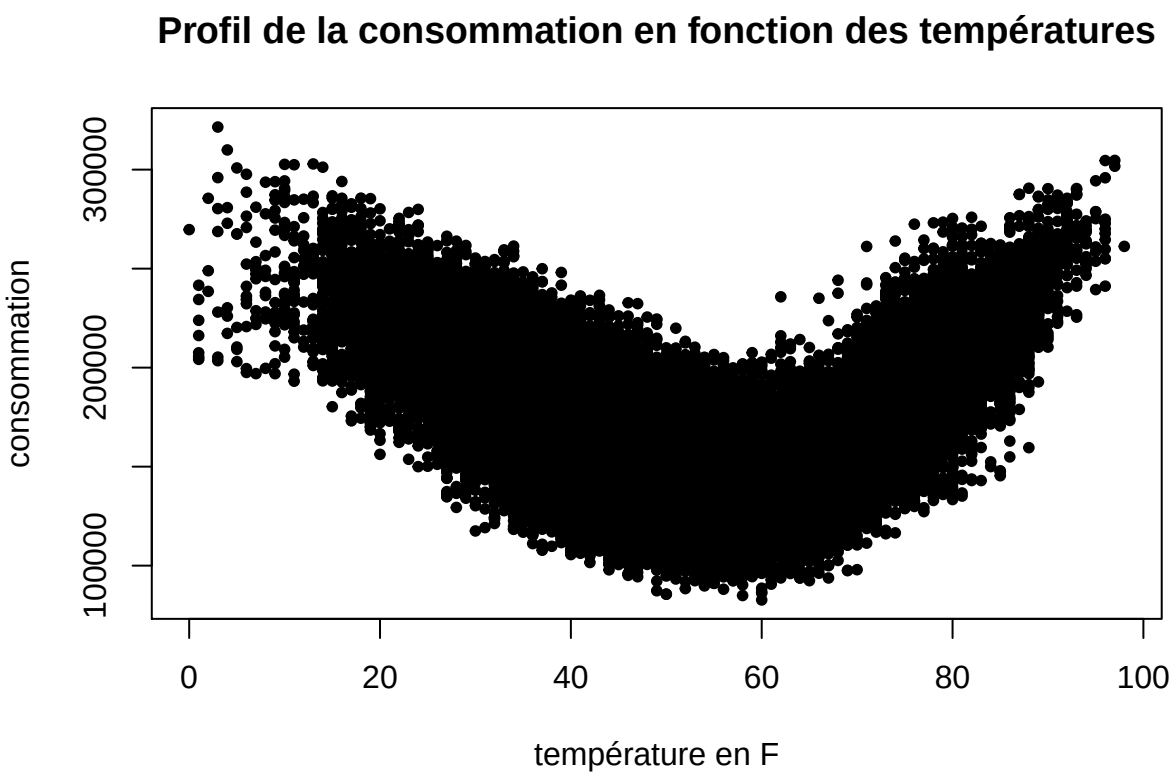
On met en évidence un profil de consommation suivant les évènements. Par exemple, les jours fériés et les week-end montrent une plus faible consommation tandis que le President Day montre une consommation plus élevée. On en déduit qu'il faudra prévoir des pics de consommation durant ces jours spéciaux, et donc connaître leurs profils horaires. Ces différences de profils sont visibles sur le deuxième graphe.

Voici la consommation par mois :



On remarque une plus grande dispersion en été qu'en hiver avec des valeurs extrêmes concentrées dans les période de transition entre hiver/printemps et été/automne. Ces écarts peuvent s'expliquer par des jours anormalement froids ou anormalement chauds. La grande dispersion en été s'explique par de plus grandes oscillations de consommation dans une même journee dûe au comportement des Americains vis-à-vis de la climatisation. Il sera pertinent de prendre en compte ce phénomène.

Regardons le graphe de la consommation par rapport à la température (on montrera à posteriori que la station 11 est une bonne station de travail) :



Nous remarquons que c'est une fonction linéaire par morceaux qui passerait le mieux dans le nuage de points. Nous allons utiliser le modèle GAM pour la régression sur les températures.

2. Modèle de prédiction

2.1 Influence de la tendance et des températures

Nous séparons nos données dans deux tables distinctes. La première contient les informations des années 2004 à 2006, la deuxième celles de 2007. Dans un premier temps, nous voulons modeliser l'influence de la température. Nous effectuons une simple régression lineaire sur les jours et non linéaire sur les données d'un couple de station. Ainsi nous prenons en compte l'influence de la tendance et des températures. Le code suivant permet de trouver le couple de station qui va minimiser l'erreur de prédiction :

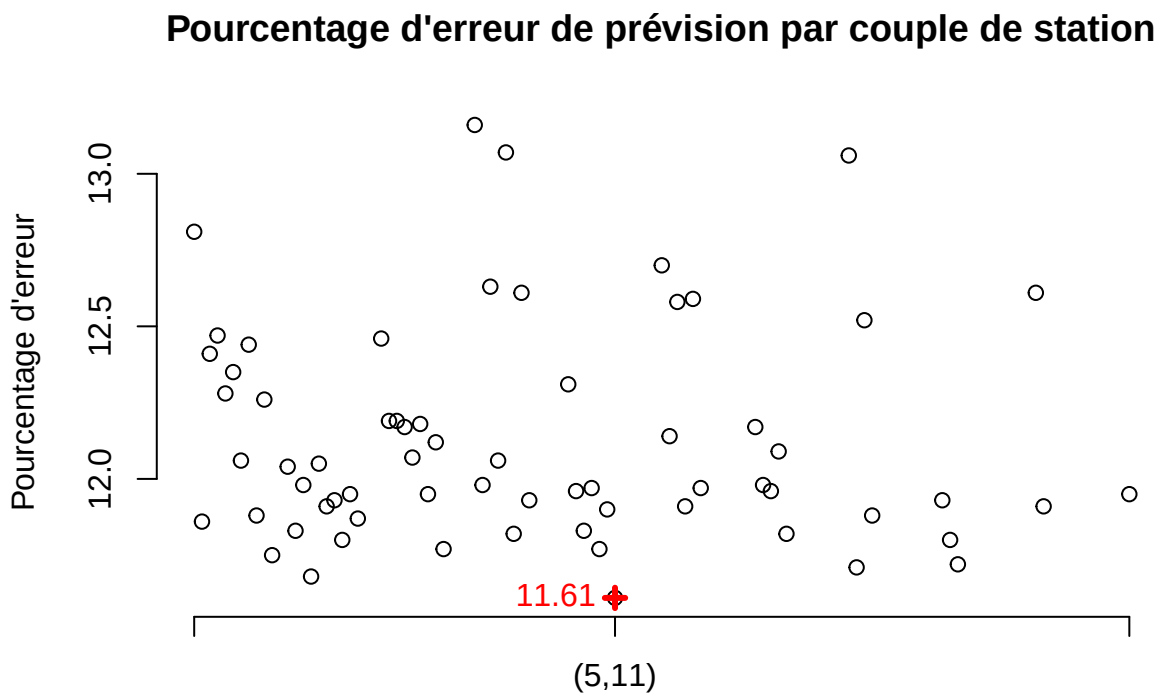
```
Station      <- matrix(nrow = NStation, ncol = NStation)
reg.station  <-list()
reg.forecast <-list()
eq           <-list()


for (i in c(1:NStation)){
  for(j in c(1:NStation)){
```

```
      Station[i,j] <- paste("s(Station",i,",k=10, bs='cr')+s(Station",j,",k=10, bs='cr')",
                            sep = "")
  }
}


R.squ           = matrix(ncol=NStation, nrow=NStation)
fit.err         = R.squ
forecast.err    = R.squ
fit.map         = R.squ
forecast.map    = R.squ


for(i in c(1:NStation)){
  for(j in c(1:i)){
    eq                  <- as.formula(paste("Zone2~Time+",Station[j,i],sep=""))
    reg.station         <- gam(eq, data=data0a)
    reg.forecast        <- predict(reg.station, data0b)
    R.squ[i,j]          <- summary(reg.station)$r.sq
    fit.err[i,j]        <- rmse(data0a$Zone2 - reg.station$fitted)
    forecast.err[i,j]   <- rmse(data0b$Zone2 - reg.forecast)
    fit.map[i,j]        <- mape(data0a$Zone2, reg.station$fitted )
    forecast.map[i,j]   <- mape(data0b$Zone2, reg.forecast)
  }
}
```

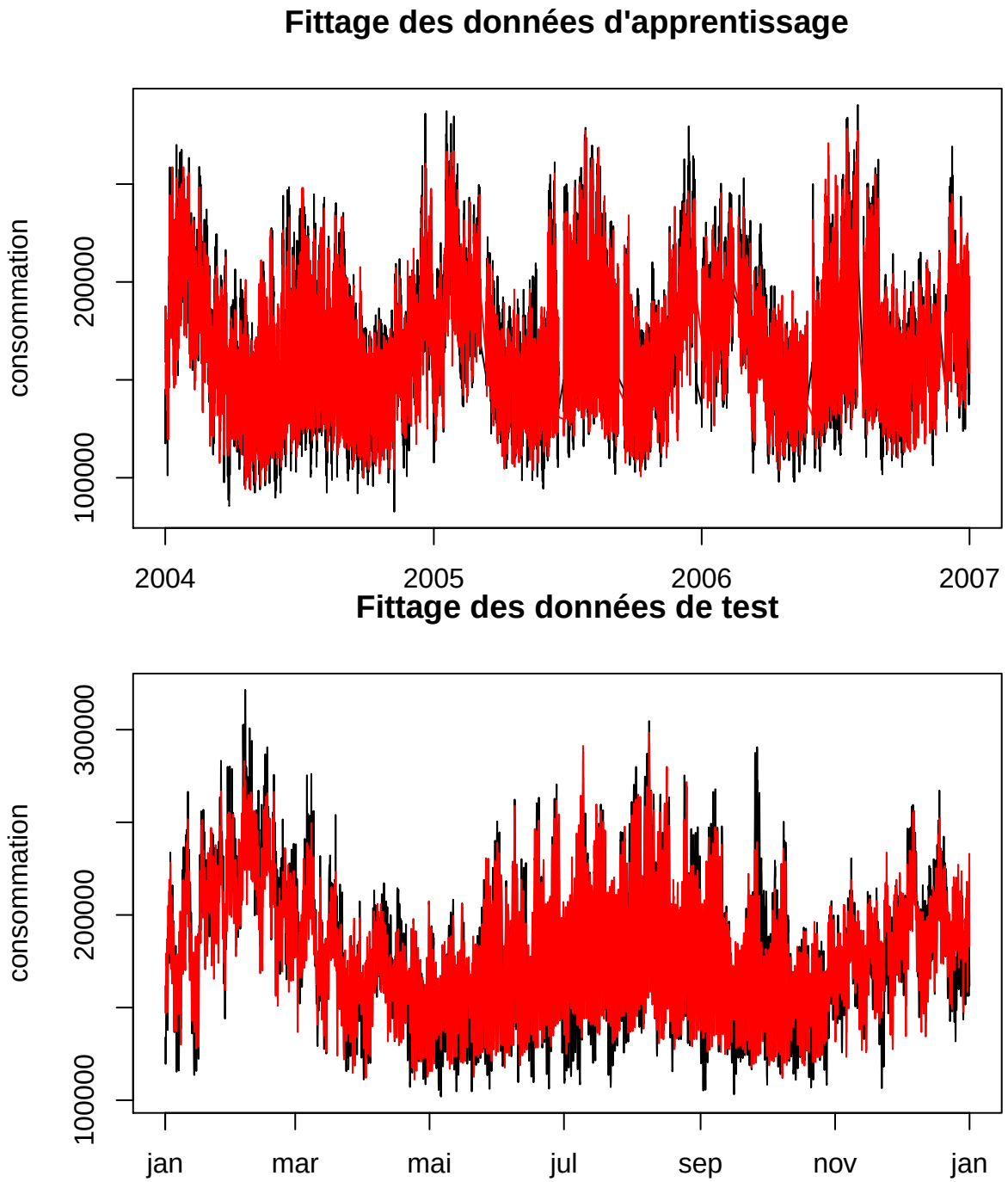


On trouve que le couple (5,11) minimise l'erreur de prédiction. C'est ce couple que nous utiliserons par la suite. On obtient un coefficient de corrélation R2 égal à 54%.

2.2 Influence de la saisonnalité annuelle et de la saisonnalité journalière

Nous allons désormais prendre en compte l'influence des composantes temporaires dans notre modèle. La variable Hour expliquera la saisonnalité journalière. Nous avons vu que les profils horaires changent selon le type de jour, nous allons donc rajouter cette information à la régression. La variable Dow du jeu de données décrit le numéro de la mesure faite au cours de l'année.

```
eq        <- as.formula(paste("Zone2~Time + s(Hour, k=10, by=daytype) + s(Toy, bs='cc') +",
                              formule.gam.double, sep=""))
```



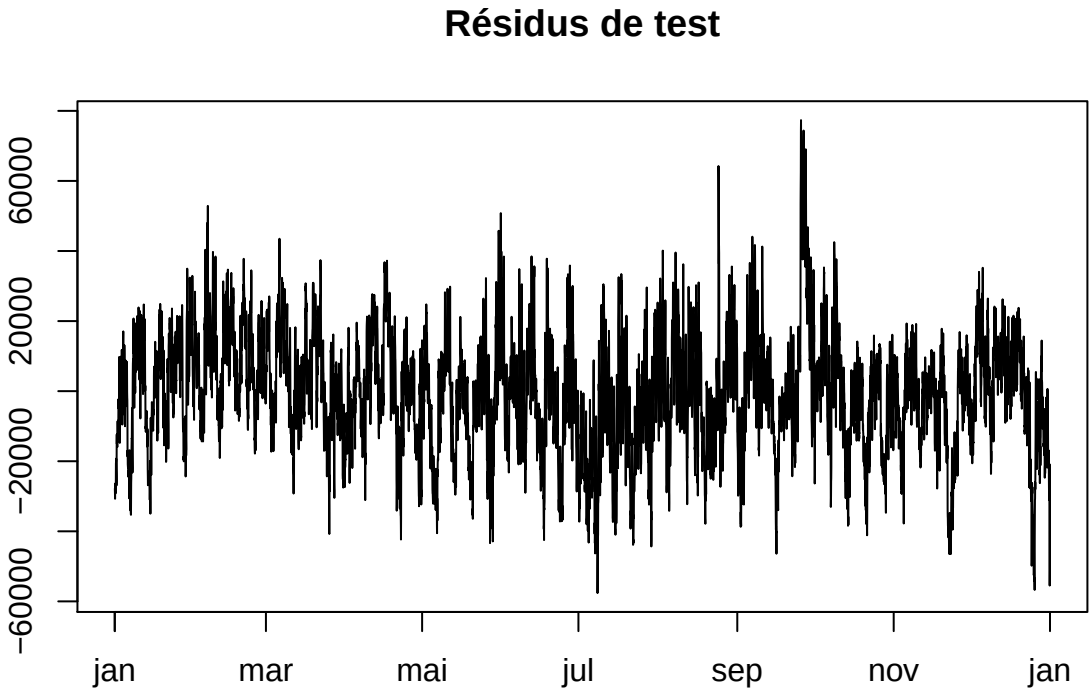
Nous avons significativement améliore notre modèle. Nous sommes passé de 11.61% à 7.35% en terme d'erreur

de prédiction et la variance des données expliquée passe de 54% à 81%.

Nous pouvons encore améliorer notre modèle. On constate qu'on ne s'adapte pas tout à fait aux valeurs extrêmes de nos données d'apprentissage, nous ne prédisons pas certains pics de nos données de test. Nous souhaiterions avoir une erreur de prévision proche des 5%. Pour cela nous allons étudier les résidus de notre modèle actuel.

3 Etude des résidus - modèle SARIMA

Nous avons récuperé les résidus de notre modèle gam. Leur étude va nous permettre de savoir s'il reste de l'information non exploitée dans notre modèle. Voici le graphique des résidus :



Sur l'autocorrélogramme, on remarque qu'il reste une tendance ainsi que les saisonnalité journalière et hebdomadaire. En effet, on voit des pics à 24h et à 7*24h, de plus l'autocorrélation ne tend pas exponentiellement vers 0.

Nous allons construire un modèle SARIMA en différentiant nos résidus. Pour que le modèle fonctionne bien sur nos ordinateurs, nous avons coupé en deux les résidus de la partie test.

Le modèle SARIMA que nous avons tenté sur les résidus différenciés à l'ordre 7*24 n'a pas abouti à cause du nombre de paramètre dans le modèle.

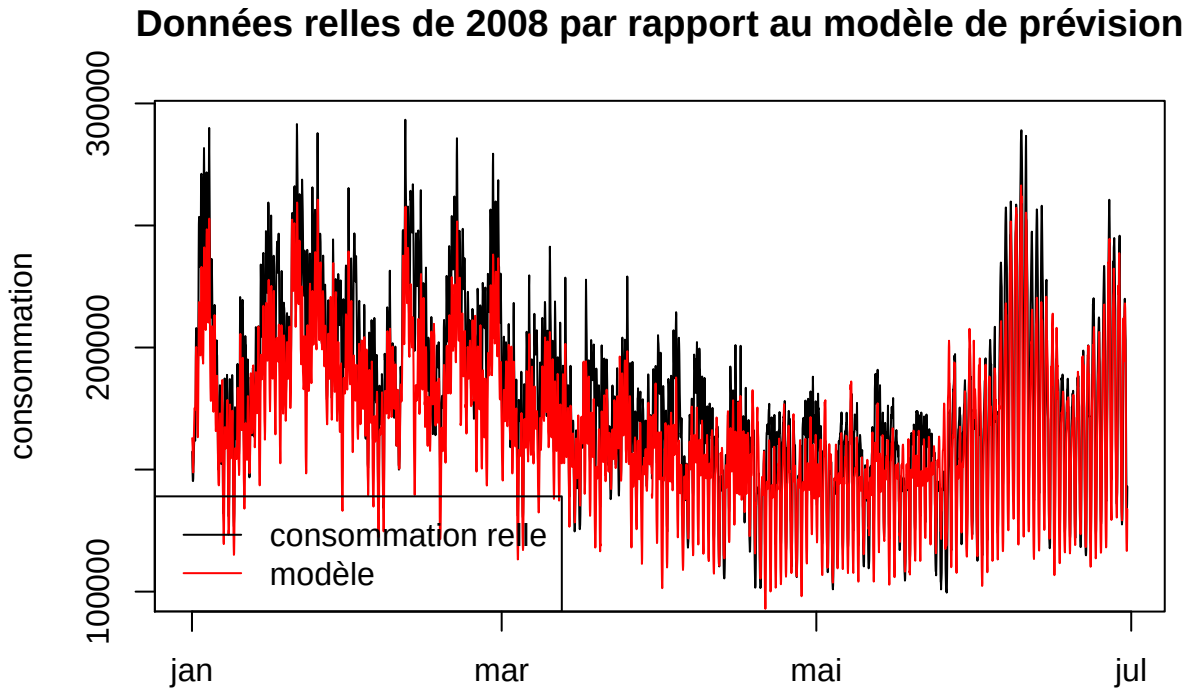
Nous avons eu l'idée de considérer la saisonnalité hebdommadaire sans prendre en compte la saisonnalité journalière. Pour cela, on a créé vingt-quatre modèles SARIMA soit un par heure de la journée. Cependant, la variance du modèle que nous avons obtenu est beaucoup trop grande, notre modèle n'est pas utilisable. Le modèle par heure est trop sensible aux valeurs extrêmes (présentes en grand nombre dans nos données). Avec plus de temps, nous aurions pû corriger de façon efficace ces valeurs extrêmes et ainsi faire un modèle de prévision prenant également en compte les erreurs présentes dans les résidus.

4. Prédiction sur l'année 2008

Nous allons maintenant appliquer notre modèle de prévision sur toutes les données de l'année 2008. Voici l'équation de notre modèle :

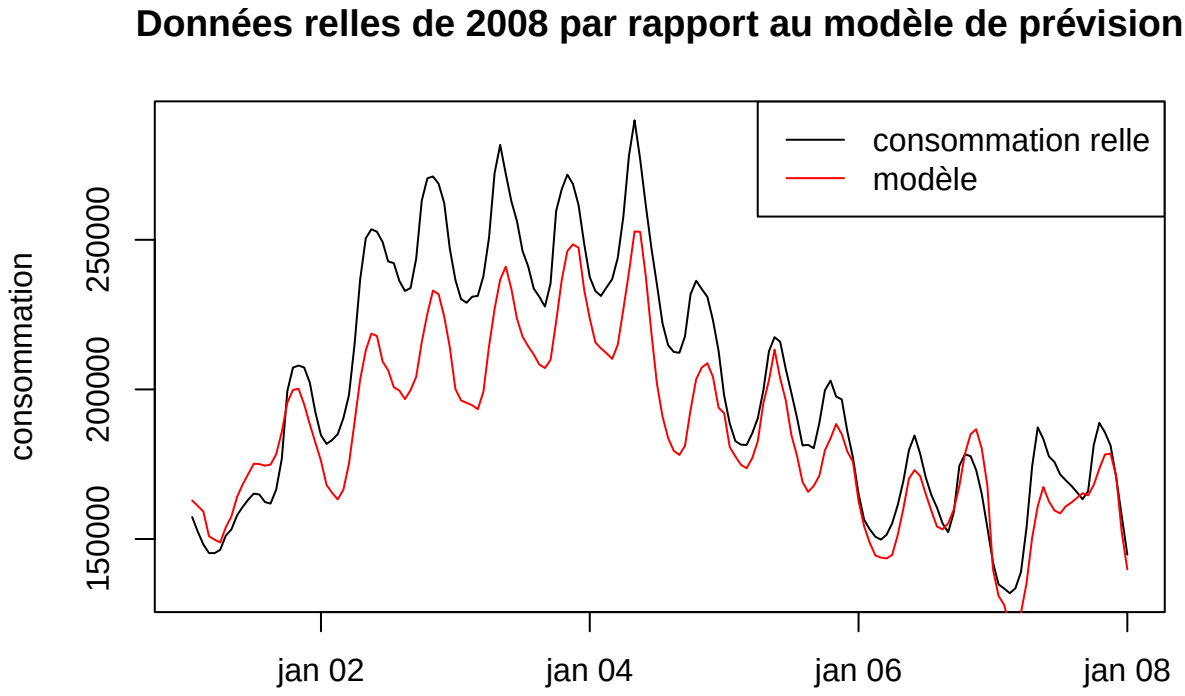
```
eq
```

```
## Zone2 ~ Time + s(Hour, k = 10, by = daytype) + s(Toy, bs = "cc") +
##     s(Station5, k = 10, bs = "cr") + s(Station11, k = 10, bs = "cr")
```



On constate que nous sous-estimons la tendance de l'année 2008, nos prévisions sont décalées vers le bas. Notre indice R2 est similaire à celui de nos test et est à 82%. Notre erreur de prédiction passe à 8.7% ici contre 7.3% lors des tests.

Regardons la qualité de nos prévisions à horizon une semaine:



On remarque un écart important du 2 au 4 janvier, même si notre erreur de prévision n'est que de 8.27%.

Conclusion

Notre modèle nous permet d'avoir une bonne idée du profil de la consommation de la zone 2 au cours du temps. Cependant, nos prédictions ne sont pas assez précise pour être utilisées au jour le jour. De plus, nous n'avons pas anticipé la hausse de la tendance sur les six premiers mois de 2008.

Si nous avions eu plus de temps, nous aurions essayé une autre méthode de prédiction itérative, avec par exemple un pas de temps de l'ordre du jour ou de la semaine. Nous aurions alors complété notre modèle avec l'information donnée par les résidus, puis nous aurions continué nos prévisions de manière itérative tout en le complétant au fur et à mesure. Construire ce modèle nous aurez permis de déduire au fur et à mesure l'augmentation de la tendance sur l'année 2008, et donc d'avoir une meilleure prévision.

L'étude de ces résidus aurait été beaucoup plus facile, car dans un laps de temps plus court, la fonction arima n'aura pas le souis du trop grand nombre de données. De plus, la variance de la consommation aurait été beaucoup plus petite, les prévision des modèles SARIMA aurait été de bien meilleurs qualite. ?????????

De plus, nous n'avons pas pris le temps de compléter les données manquantes dans notre table d'apprentissage. Cela nous aurait permis d'avoir un modèle sans doute plus cohérent. Pour finir, il aurait été interessant d'utiliser des méthodes de bootstrap pour quantifier les erreurs successives de notre modèle.