Analysis of the electricity consumption in Dayton (Ohio)

Charles Dufour and Cécile Trottet

June 2019

1 Introduction

The data set [1] comes from PJM Interconnection LLC. It is a power transmission system operator distributing part of the electricity to the eastern side of the USA. PJM provides the amount of electricity consumed in megawatts per hour for several states or cities. We chose to study the consumption of Dayton city. The consumption was measured from 01.10.2004 until 03.08.2018, giving a total of 121275 observations. Given the large amount of observations and the complexity of the resulting time series, we chose to only use weekly electricity measurements, reducing the size of the set to 723 values. Furthermore, it is plausible that the outside air temperature [2] in Dayton is linked with the energy consumption of the inhabitants. We will see how these two data sets are connected.

2 Initial data analysis

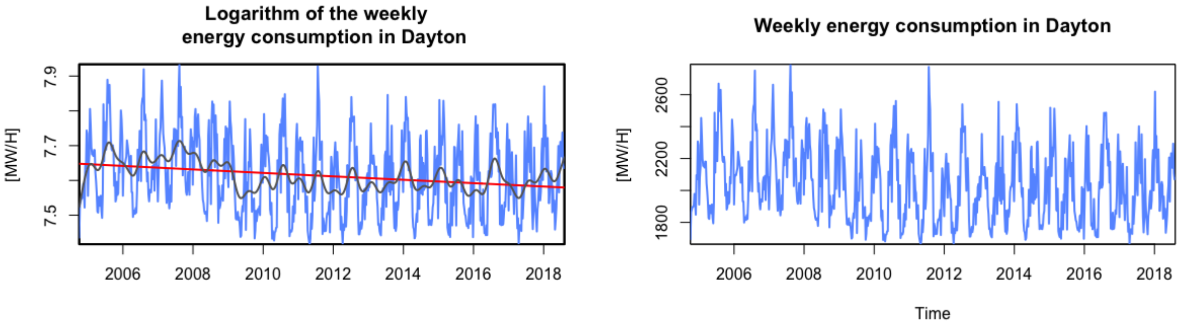
The very first step of this analysis consisted in cleaning both the electricity and temperature data sets. The electricity time series is shown in Figure 1b. Because of its large fluctuations, the logarithm of the values was taken. From now on, when speaking of the energy consumption, it should be understood that the logarithm of the weekly energy consumption is implicitly meant.

In order have a visual representation of the trend, smoothing splines of various degrees of freedom (Figure 1a) were fitted. One can see a slight overall downwards trend and waves suggesting some sort of seasonality.

Figure 2b allows us to see the yearly shape of the energy consumption and to observe several facts. For example, in winter the energy consumption peaks around week 5 (this corresponds to the beginning of February) and then decreases until week 17, when people start to use their air conditioning. Afterwards, the consumption peaks again at week 32 (beginning of August) in the summer. Unsurprisingly, the electricity usage is at its lowest in the early morning hours (i.e. around 5 a.m.), peaks around the beginning of the afternoon and is average in the early evening. Furthermore, it is interesting to notice that over the year, there is much less volatility in the energy consumption during the hours for which the consumption is low. More specifically, for example at 5 a.m., the amount of energy consumed is almost stable over the year, whereas at the peak hours the variance is larger.

In Figure 2a, one can observe the relationship between the energy consumption and the outside air temperature in Dayton. Two distinct phases are present in the graph. When the temperature is below 60 °F (i.e. 15.5 °C) the energy consumption decreases almost linearly with the increase of the temperature. After the shift at 60 °F, the energy consumption increases as it gets warmer outside. The slope in the second part of the graph is larger than in the first phase, showing that the air conditioning is more energy demanding than the heating.

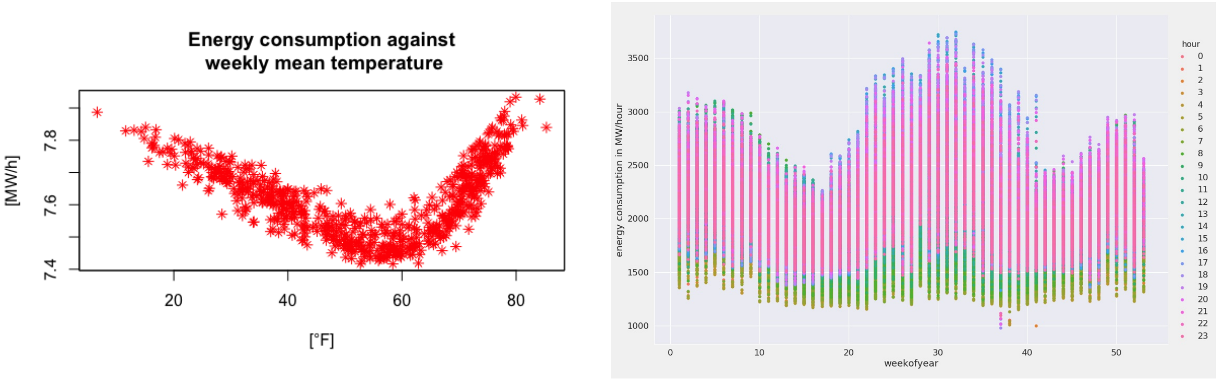
In the next sections, we used the data of the years 2004-2017, i.e. the first 560 values, to try to build an appropriate model and we kept the last 3 years (i.e. observations from 04.08.2015 until 03.08.2018) to compare with the forecasting results we got. We will refer to the first part of the data set as the ”training set” and to the second part as the ”test set”.



(a) Visualization of the trend, using splines

(b) Data set

Figure 1: Initial time series



(a) Temperature versus energy consumption

(b) Pairplot

Figure 2: Some initial plots

3 Model A

First, we tried to progressively develop an appropriate SARIMA model. Since we are analyzing an energy consumption data set over several years, the presence of a yearly seasonal component seems plausible. This choice of model therefore seems appropriate to fit the data. This is not our final model, therefore some details are omitted in this section and will be given during the construction of the final, more accurate, model. Let d_t be the energy consumption at time t, B be the backshift operator, $Bd_t = d_{t-1}$, and ε_t be white noise. From the lecture notes [3], we know that the general formulation of a SARIMA is given by :

$$\Phi_P(B^s)\phi(B)(I-B)^d(I-B^s)^D d_t = \Theta_Q(B^s)\theta(B)\varepsilon_t, \tag{1}$$

where Φ_P and Θ_Q are respectively the seasonal autoregressive and moving average operators, and $\phi(B)$ and $\theta(B)$ the ordinary autoregressive and moving average operators. We will start by firstly estimating the ordinary and difference components $(I-B)^d$ and $(I-B^s)^D$. In a second time, we will fit models with different parameters p and q that seem appropriate for the ordinary operators, and orders P and Q for the seasonal operators.

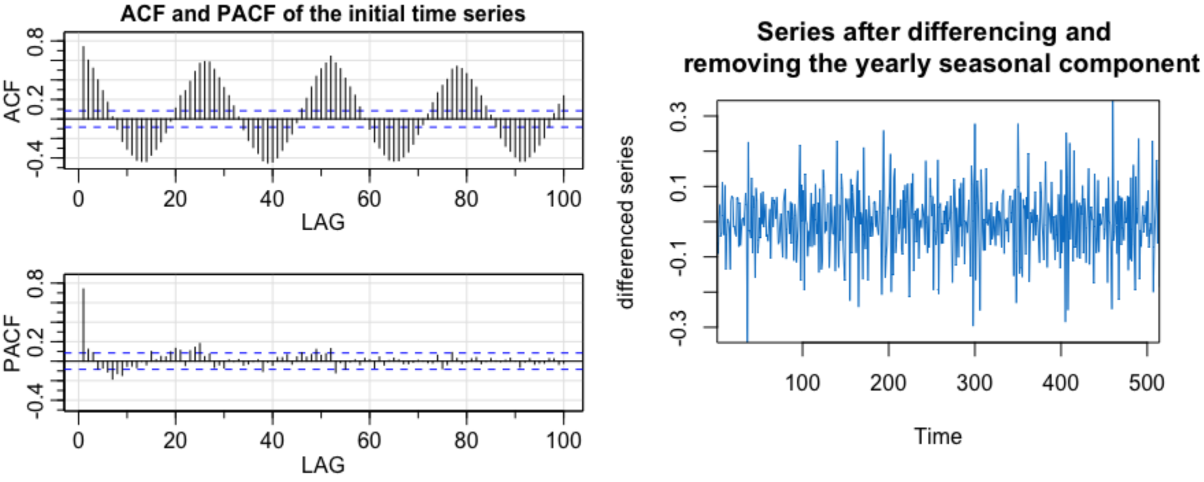
In order to select a model with appropriate p, d, q and P, D, Q, the correlogram (ACF) and the partial correlogram (PACF) are widely used. We briefly describe what these two plots represent.

Let c_h be the sample autocorrelation of the residuals at lag h. Then the ACF is the graph of $\frac{c_h}{c_0}$ against lag h. For some arbitrary h, it can be thought of as the degree of dependence between a given observation and the observation following it after h units of time.

When computing the partial autocorrelation function at lag h, we remove the effects due to linear dependencies with shorter lags. Therefore, the PACF plot gives us an idea of the dependence of a given observation only with the observation coming h steps further along.

3.1 Estimating the ordinary and seasonal difference orders

Our data set contains the values for the weekly energy consumption over several years, therefore we expect a seasonal component to be present. The most plausible assumption is that a 52-fold difference operator could be useful to remove an eventual yearly seasonality. Indeed, given two distinct years, one can imagine that the consumption should be roughly the same for example in April of each of these years. The ACF in Figure 3a falls outside of the bounds at almost every lag. Furthermore, it peaks and decays slowly every 13 weeks (corresponding to a quarter of the year). Differencing once in order to decrease the overall values (i.e. setting $d = 1$ in formula 1) and removing the yearly seasonal component (i.e. setting $s = 52$ and $D = 1$ in formula 1) seems appropriate.



(a) ACF and PACF of the original data set

(b) $(I - B)(I - B^{52})d_t$

Figure 3

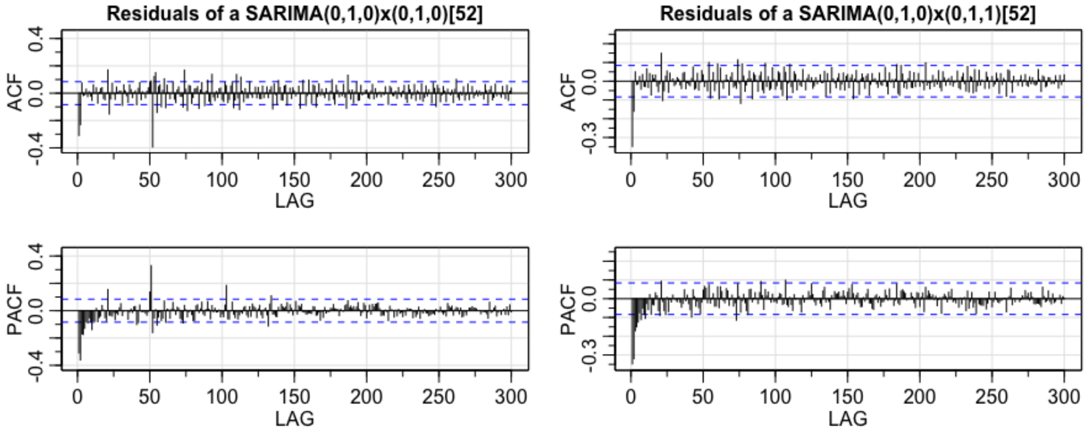
Figure 3b shows the series after differencing and removing the yearly seasonal component. Visually, this series seems to be stationary and the p-value $p = 0.1$ returned by the KPSS test confirms that the series is indeed highly likely to be stationary. For this reason, the ordinary and seasonal difference orders of 1 seem high enough. At this point, we have applied a SARIMA$(0, 1, 0) \times (0, 1, 0)[52]$ to our data set.

3.2 Estimating the orders P and Q of the SARIMA model

To find the accurate orders P and Q, we try to distinguish relevant patterns within the seasonal parts of the ACF and PACF of the residuals from the previous model (Figure 4a). We notice that the ACF cuts off at lag 52, while the PACF tails off at lags 52 and 104. This indicates to set $Q = 1$ and $P = 0$.

3.3 Estimating the orders p and q of the SARIMA model

The ACF and PACF of the residuals of the SARIMA$(0, 1, 0) \times (0, 1, 1)[52]$ don't show any patterns anymore at lags 52 and 104 (Figure 4b). It seems that setting $Q = 1$ and $P = 0$ describes our data accurately. We now look at the first lags to identify the ordinary components. Since the ACF cuts off after the first lag, and the PACF tails off suggests taking $p = 0$ and increasing q. After setting $q = 2$, the ACF and PACF of the residuals of a SARIMA$(0, 1, 2) \times (0, 1, 0)$ (Figure 5) mostly fall within the limits and don't show any strong patterns. We will perform a few additional tests on this model in order to assess whether its residuals are indeed white noise.



(a) Residuals of a SARIMA$(0,1,0) \times (0,1,0)[52]$ (b) Residuals of a SARIMA$(0,1,0) \times (0,1,1)[52]$

Figure 4

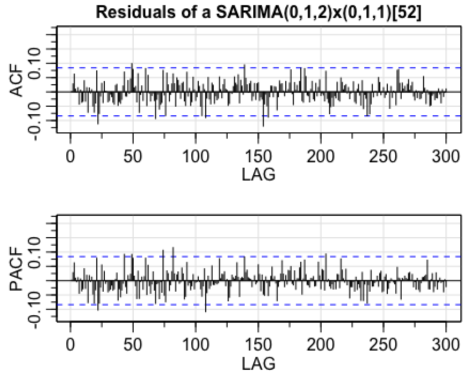


Figure 5: Residuals of a SARIMA$(0,1,2) \times (0,1,1)[52]$

3.4 Testing for white noise

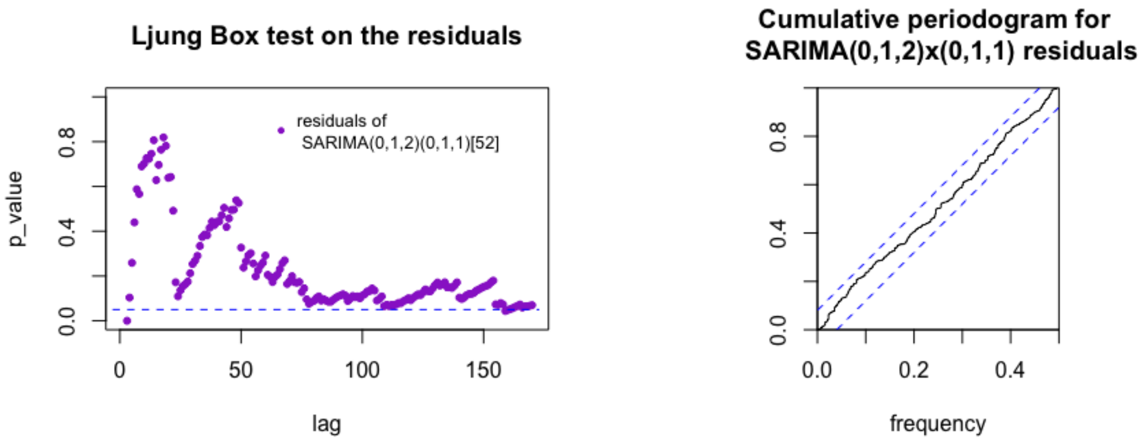
3.4.1 Ljung-Box test and cumulative periodogram

We will now perform some additional test on the SARIMA$(0,1,2) \times (0,1,1)[52]$ in order to assess whether its residuals are indeed white noise. Let us first perform a Ljung-Box test. The hypothesis we want to test is whether the residuals are independently distributed. From the lecture notes [3], we know the test statistic :

$$Q_m = n(n+2) \sum_{h=1}^{m} (n-h)^{-1} \hat{\rho}_h^2,$$

with n the number of values in the dataset, m the number of lags being tested and $\hat{\rho}_h^2$ the sample autocorrelation. Under the null hypothesis (independent residuals), Q_m follows a $\chi^2_{m-(p+q)}$ distribution for large n, where p and q are the ordinary parameters of the SARIMA model. We performed the test for lags up to 170. We notice that the value for the first lag falls under the 0.05 threshold. However, since all other values are greater than 0.05, the residuals seem to be independent. Maybe a slightly more complicated model would suppress the correlation with the first lag.

Figure 6b shows the cumulative periodogram of the residuals. Since it falls within the 95% bounds for all frequencies, it is reasonable to think that the residuals are white noise.



(a) Ljung-Box test

(b) Cumulative periodogram

Figure 6

3.5 Forecasting

The model developed on the training data set is used to forecast the values of the test data set. The predictions are accurate since the red line (representing the real values of the test data) falls within the 95% grey prediction interval bounds for the model. However, one can notice that the forecasts (in blue) fail to account for the peaks in the data and that the confidence intervals have to become quite large in order to capture the test set.

To make up for this lack of precision in the forecasting, we will build an additional model based on the temperature.

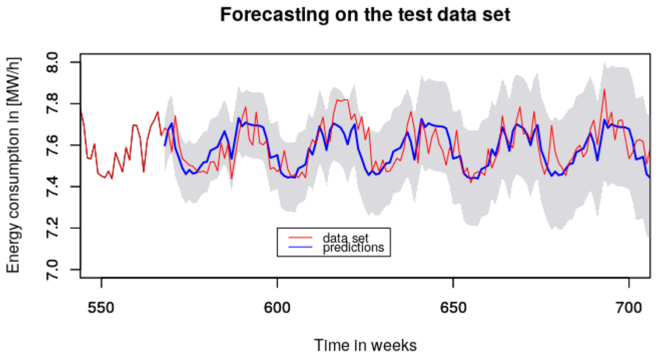


Figure 7: Predictions on the last 3 year with 95 percent confidence intervals

4 Model B

4.1 Modelling of the trend of the energy consumption

In this section we take another approach to model our time series. In order to get better forecasting results, we use the temperature data to model the trend of the energy consumption, instead of differencing. In other words, we will fit a linear model with the temperature as the exogenous variable to the energy consumption, the dependent variable. In a second phase, we will model the residuals of the linear model, using again a time series analysis.

4.1.1 Selection of the best polynomial regression

We want to assess whether the energy consumption at time t could be expressed as a linear combination of polynomials of maximum degree d, evaluated in the temperature at the same time. In order to do that, we fitted polynomial regressions up to degree $d = 4$ of the temperature to the energy consumption. Since these models are nested, we performed likelihood ratio tests in R to assess which parameters are significant. The results can be found in Table 1. The p-values indicate that a model of degree 2 is not accurate enough, while the regression of degree 4 might overfit the data. Therefore we fitted a degree 3 linear model of the temperatures to the electricity consumption.

	$H_0 : d = 1$ $H_1 : d = 2$	$H_0 : d = 2$ $H_1 : d = 3$	$H_0 : d = 3$ $H_1 : d = 4$
p-value	$< 2.2 \times 10^{-6}$	$< 2.2 \times 10^{-6}$	0.4168

Table 1: p-values for the comparison of the nested models

Let d_t be the energy consumption at time t and T_t be the temperature at time t. The parameters a_0, a_1, a_2 and a_3 have been computed so that :

$$d_t \approx a_o + a_1 \cdot T_t + a_2 \cdot T_t^2 + a_3 \cdot T_t^3.$$

The details of the fitted polynomials can be found in figure 19 of the appendix.
In Figure 8 the computed polynomial regression was used to perform a prediction of the energy consumption for the year succeeding the last year of the training data. The forecasted values are compared with the actual values of our test data set. We can see that the linear model fits well the overall trend of the consumption but fails to account for the local deviations from the average. It remains to model the part of the energy consumption that could not only be explained by a degree 3 regression of the temperature.

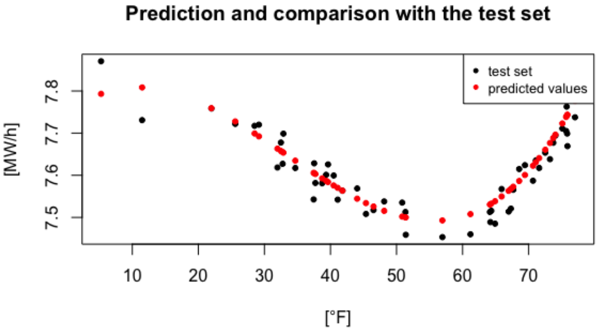


Figure 8: Prediction of future energy consumption by a degree 3 regression of polynomials and comparison with the test data set

4.2 Modelling the residuals of the linear model

Let d_i be the energy consumption at time i given by the training data set and let f_i be the fitted value of the linear model at time i. We will perform a time series analysis in order to model the residuals e_i :

$$e_i := d_i - f_i, \quad i = 1, \ldots, 671.$$

Graphically, the series is shown in Figure 10a. Similarly to what was done during the analysis of model A, we develop step by step an accurate SARIMA model to fit e_t. In order to avoid any confusion, when we refer to the "series", we mean e_t, i.e. the residuals of the linear model.

4.2.1 Estimating the appropriate ordinary and seasonal difference orders

Figure 10a gives us a first glance at the shape of the residuals. Graphically, they do not appear stationary as we can observe at least two distinct parts of the graph. There seems to be a downwards trend before 2009 and afterwards, visually, the series looks stationary. The KPSS test on the time series of the residuals returned a p-value of 0.01, confirming the non stationarity of the residual series.
When looking at the ACF and the PACF of the residuals (Figure 9), we notice that the ACF decays slowly. An ordinary difference of order at least one seems thus appropriate. Furthermore, at the lags $0, 52, 104$ there is a peak in both the ACF and the PACF. This suggests a cycle of one year in the data and it leads to believe that we need to remove the yearly seasonal component from the yearly temperatures. In mathematical terms, these two observations translate as

$$(I - B)(I - B^{52})e_t = \varepsilon_t.$$

Figure 10b shows the series after differencing and removing the yearly seasonal component. Visually, this series seems to be stationary and the p-value $p = 0.1$ returned by the KPSS test confirms that the series is indeed highly likely to be stationary. From now on, when speaking of the "differenced series", we mean the series of residuals with ordinary difference component $(I - B)$ and seasonal 52-fold difference component $(I - B^{52})$.

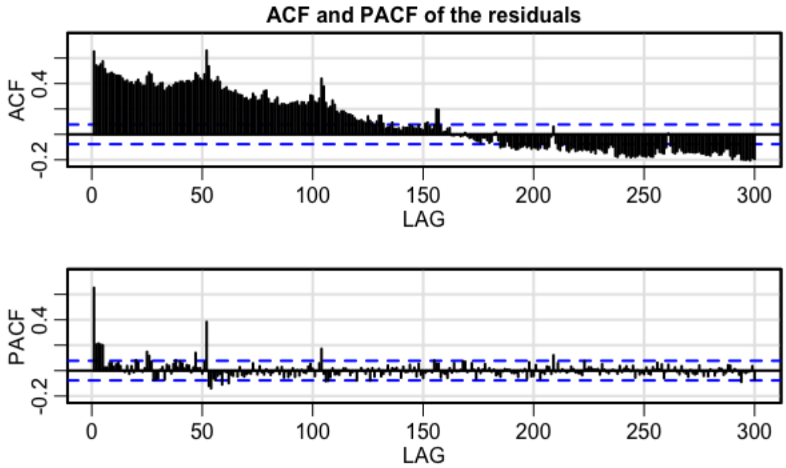
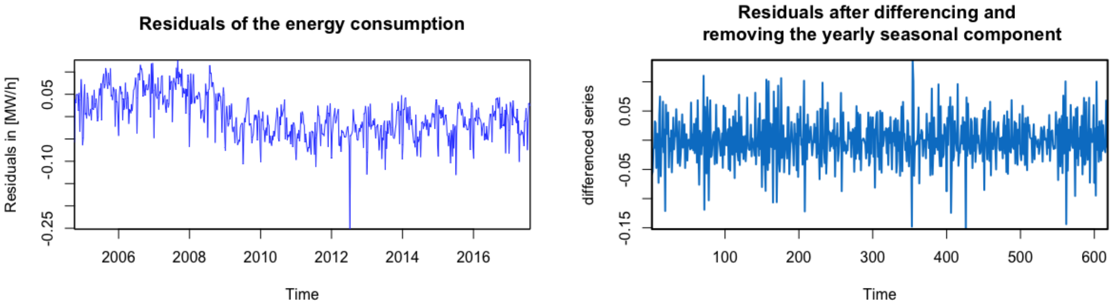


Figure 9: ACF and PACF of the residuals

4.2.2 Estimating the orders P and Q of the SARIMA model

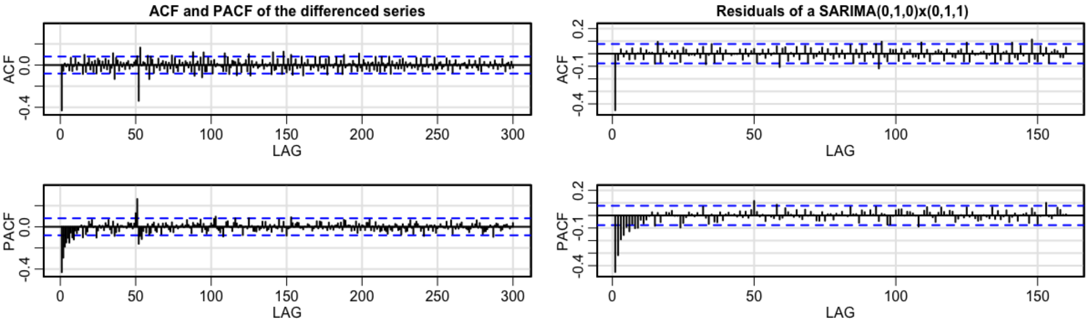
This time, we analyze the ACF and PACF of the differenced series (Figure 11a). Since we first want to remove the seasonality, lag 52 is of interest. The PACF tails off, while the ACF cuts off. This suggests trying to fit a model with a seasonal moving average component of degree Q at least 1 and an autoregressive component P of order 0.
Figure 11b shows the ACF and the PACF of the residuals after applying a SARIMA$(0,1,0) \times (0,1,1)[52]$ to the series. The graphs confirm that the yearly seasonality has been removed, since there is no visible pattern of the ACF and PACF falling outside the limits beyond lag 20.



(a) Time series of the residuals of model 1

(b) $(I-B)(I-B^{52})e_t$

Figure 10



(a) Differenced series

(b) Residuals of SARIMA(0,1,0) × (0,1,1)[52]

Figure 11

4.2.3 Estimating the orders p and q of the SARIMA model

We can now fix the ordinary parameters of the model. The ACF of residuals of the SARIMA$(0,1,0)\times(0,1,1)[52]$ model (Figure 11b) cuts of after the first lag, while the PACF tails off. These facts suggest increasing the moving average parameter, i.e. setting $q=1$. Figure 12 shows the ACF and PACF of the residuals after applying a SARIMA$(0,1,1)\times(0,1,1)[52]$ model to the series. The values fall within the limits and there is no pattern emerging, suggesting that the residuals could be independent white noise.

4.2.4 Selection of the best models according to AIC and BIC criteria

Based on the ACF and PACF results, the SARIMA$(0,1,1)\times(0,1,1)[52]$ model seems to be an appropriate model to fit the residuals of the linear model. We will now test whether it is also a good model according to the AIC and BIC criteria. These two estimators compute a tradeoff between the goodness of fit and the simplicity of the model. A model with the lowest possible AIC and BIC should be selected. We compared the values for SARIMA models with parameters p and q taking values between 0 and 3. The lowest value for the AIC was obtained for a SARIMA$(0,1,2)\times(0,1,1)[52]$, and the lowest value for BIC for a SARIMA$(0,1,1)\times(0,1,1)[52]$. Furthermore, there was a difference of less than 0.5 between the AIC values for the SARIMA$(0,1,2)\times(0,1,1)[52]$ and the SARIMA$(0,1,1)\times(0,1,1)[52]$. To sum up, the analysis of the ACF and PACF graphs led us to select the model which also appears to have the lowest BIC. A slightly more complicated model has the lowest AIC. We will proceed our analysis on the models selected by the AIC and BIC criteria. We now know that they both provide an adequate fit, even when compared to more complicated models. The mathematical formulations of the models are :

$$(I-B)(I-B^{52})e_t=(1+\Theta_{B,1}B^{52})(1+\theta_{B,1}B)\varepsilon_t \qquad \text{(BIC model)}$$

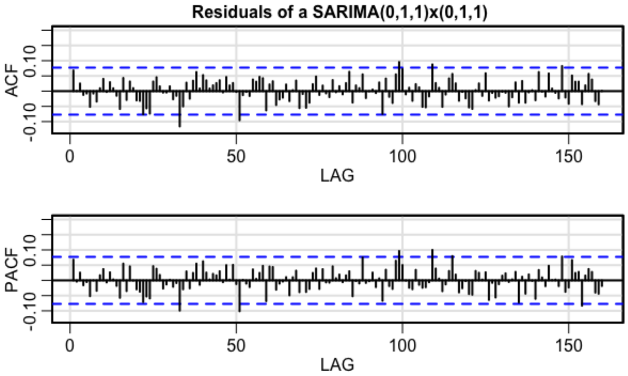


Figure 12: Residuals of $SARIMA(0,1,1) \times (0,1,1)[52]$

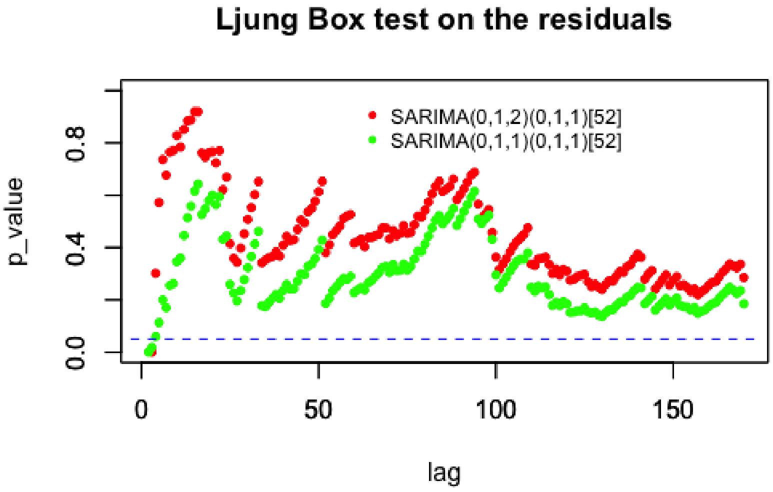


Figure 13: Ljung Box test on the two models

$$(I - B)(I - B^{52})e_t = (1 + \Theta_{A,1}B^{52})(1 + \theta_{A,1}B + \theta_{A,2}B^2)\varepsilon_t, \qquad \text{(AIC model)}$$

with $\Theta_{B,1}$ and $\theta_{B,1}$ the seasonal and ordinary moving average components of the BIC model and $\Theta_{A,1}$, $\theta_{A,1}$ and $\theta_{A,2}$ of the AIC model. The details of the models can be found in appendices 20 and 21. Remember that

$$e_t := d_t - \left(a_o + a_1 \cdot T_t + a_2 \cdot T_t^2 + a_3 \cdot T_t^3\right),$$

where d_t is the energy consumption at time t, T_t the temperature and a_o, a_1, a_2 have been estimated by fitting the linear model. It remains to determine whether the residuals of these two models are indeed white noise.

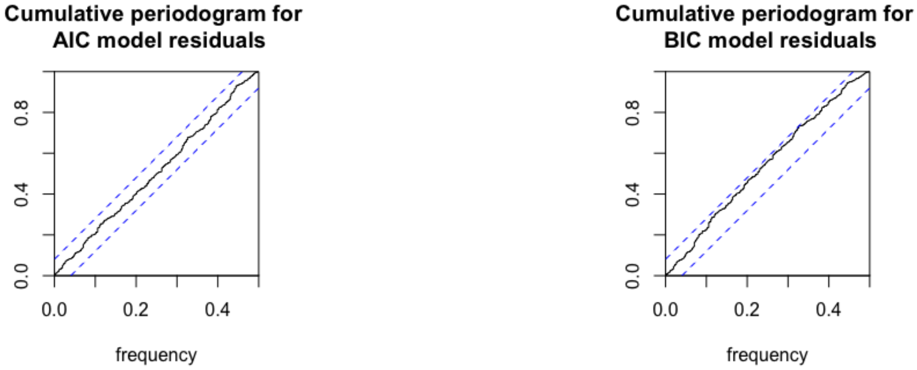
4.2.5 Testing for white noise in the residuals

- Ljung-Box test : We first performed a Ljung Box test on the residuals of the AIC and BIC models for lags up to 170.

  Overall, the p-values for the two models fall considerably above the 0.05 threshold, indicating that the residuals are likely to be independent. That being said, we notice that the two first p-values for the SARIMA$(0,1,1) \times (0,1,1)[52]$ model are smaller than 0.05. When fitting a slightly more complicated model (namely the SARIMA$(0,1,2) \times (0,1,1)[52]$) only the first p-value falls within the bounds. This points in the direction of a slight correlation with the first lag.

- Cumulative periodogram :

  Figures 14a and 14b show the cumulative periodograms for the two models. The values for both models fall within the 95% confidence bounds, suggesting that the residuals behave like white noise. Nevertheless, the plot for the BIC falls very close to the band. If there was still some serial correlation left in the BIC model, it has been removed by the more general AIC model.



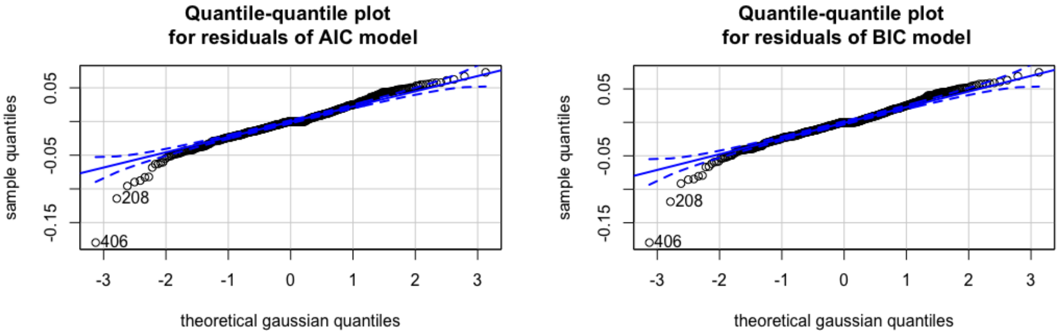
(a) AIC model

(b) BIC model

Figure 14: Cumulative periodograms

- Checking for normality :

  From the results in the previous discussion, the residuals of both models seem to behave like white noise. The following plots were made so that one may visually assess if the white noise is normally distributed or not. The Quantile-Quantile plots for both models are quite similar. The distributions seem to be slightly left skewed. One could ask the question whether observation number 406 is an outlier, since it lies well below the normality line. It corresponds to a week in July 2012. The outside air temperature was of 85 degrees Fahrenheit (29 degrees Celsius). This isn't an abnormal temperature, it just happens to be the highest temperature throughout our whole dataset. This enlightens the limits of our model. It fails to explain quite accurately some extreme observations.



(a) AIC model

(b) BIC model

Figure 15: Quantile-quantile plots

4.3 Forecasting on the test data set

We finally used the remaining three years of our energy consumption data set to compare them with the predictions provided by the developed models. Our final model is composite; a time series analysis was performed on the residuals of a linear model. We took that into account when we did the predictions. We used the last three years of temperature data to perform a prediction on the energy consumption using the linear model. The time series model was used to forecast the difference between the initial values and the fitted linear model. Therefore our final prediction is given by adding up the two separate predictions. We plotted these predictions together with the actual test data set. We also plotted the values of the predictions obtained by only using the linear model. The test data set lies within the 95% prediction intervals provided by our forecast. Furthermore, in general, the predictions of the complete model are closer to the test data set than the predictions provided only by the linear model.

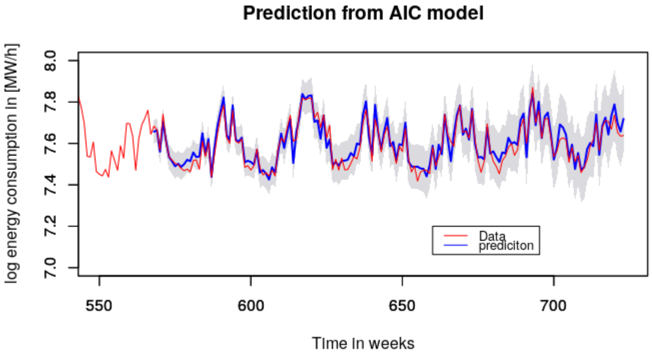


Figure 16: Predictions of the AIC model on the last 3 year with 95 percent confidence intervals

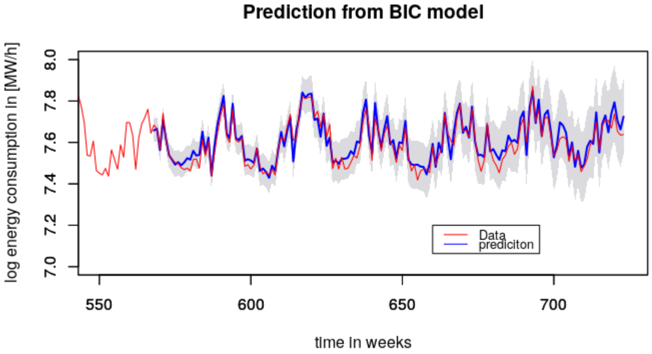


Figure 17: Predictions of the BIC model on the last 3 year with 95 percent confidence intervals

5 Conclusion

When considering only the energy consumption, after taking the logarithm of the values, a SARIMA $(0,1,2)\times(0,1,1)[52]$ seemed to fit the data quite well. This is a satisfying result, since this model isn't a very complicated one, and has only a few parameters. However, due to the lack of precision of the forecasting, it was convenient to also have the temperature data at hand. Even though it makes sense from a behavioural aspect, it was still quite surprising to observe how well the temperature fitted the electricity consumption. The remaining residuals could afterwards easily be modeled. The ACF and PACF graphs were useful to find the accurate parameters and provided an insight consistent with the theory behind the AIC and BIC estimators. Two models were selected based upon these criteria, a SARIMA $(0, 1, 1) \times (0, 1, 1)[52]$ and a SARIMA $(0, 1, 2) \times (0, 1, 1)[52]$. The statistical tests made on the residuals of these models supported the assumption that their residuals are white noise.

One could think of several other ways to model the trend of the electricity consumption hadn't the temperature data been available or if it had to be used in a different manner. For example, still using the temperature data, orthogonal polynomials could have been fitted instead of applying a regular polynomial regression. Similarly, we could also have used smoothing splines. Without the temperature, we could have tried to fit functions of the time as smoothers.

References

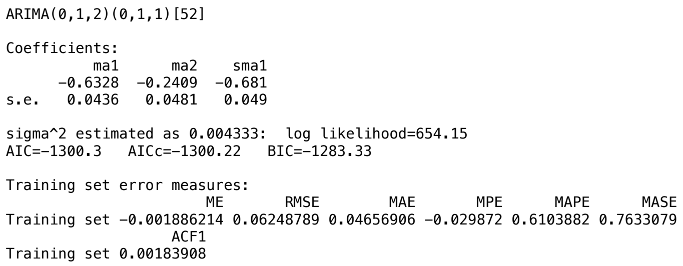
[1] https://www.kaggle.com/robikscube/hourly-energy-consumption

[2] University of Dayton - Environmental Protection Agency Average Daily Temperature Archive, http://academic.udayton.edu/kissock/http/Weather/default.htm

[3] Davison, A.C. (2019) Time Series. Course notes, Ecole polytechnique federale de Lausanne.

A Theory

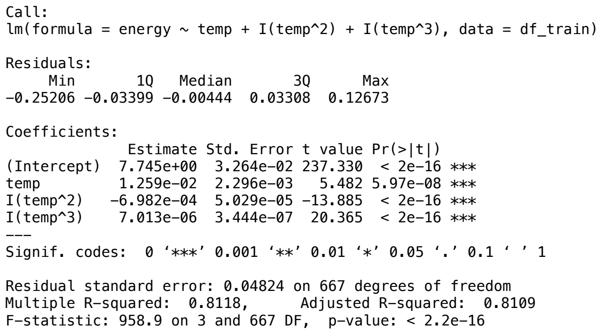
B Model A



```
ARIMA(0,1,2)(0,1,1)[52]

Coefficients:
          ma1      ma2     sma1
      -0.6328  -0.2409   -0.681
s.e.   0.0436   0.0481    0.049

sigma^2 estimated as 0.004333:  log likelihood=654.15
AIC=-1300.3   AICc=-1300.22   BIC=-1283.33

Training set error measures:
                       ME       RMSE        MAE       MPE      MAPE      MASE
Training set -0.001886214 0.06248789 0.04656906 -0.029872 0.6103882 0.7633079
                   ACF1
Training set 0.00183908
```

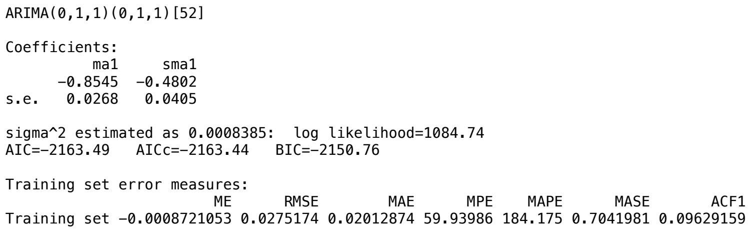
Figure 18: Details of the SARIMA(0,1,2) × (0,1,1)[52] model

C Model B



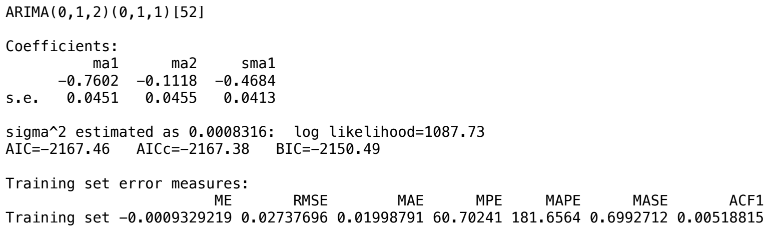
```
Call:
lm(formula = energy ~ temp + I(temp^2) + I(temp^3), data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25206 -0.03399 -0.00444  0.03308  0.12673 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.745e+00  3.264e-02 237.330  < 2e-16 ***
temp         1.259e-02  2.296e-03   5.482 5.97e-08 ***
I(temp^2)   -6.982e-04  5.029e-05 -13.885  < 2e-16 ***
I(temp^3)    7.013e-06  3.444e-07  20.365  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04824 on 667 degrees of freedom
Multiple R-squared:  0.8118,	Adjusted R-squared:  0.8109 
F-statistic: 958.9 on 3 and 667 DF,  p-value: < 2.2e-16
```

Figure 19: Details of the polynomials fitted to the energy consumption



```
ARIMA(0,1,1)(0,1,1)[52]

Coefficients:
          ma1     sma1
      -0.8545  -0.4802
s.e.   0.0268   0.0405

sigma^2 estimated as 0.0008385:  log likelihood=1084.74
AIC=-2163.49   AICc=-2163.44   BIC=-2150.76

Training set error measures:
                        ME      RMSE        MAE      MPE    MAPE      MASE       ACF1
Training set -0.0008721053 0.0275174 0.02012874 59.93986 184.175 0.7041981 0.09629159
```

Figure 20: Details of the BIC model



```
ARIMA(0,1,2)(0,1,1)[52]

Coefficients:
          ma1      ma2     sma1
      -0.7602  -0.1118  -0.4684
s.e.   0.0451   0.0455   0.0413

sigma^2 estimated as 0.0008316:  log likelihood=1087.73
AIC=-2167.46   AICc=-2167.38   BIC=-2150.49

Training set error measures:
                        ME       RMSE        MAE      MPE     MAPE      MASE       ACF1
Training set -0.0009329219 0.02737696 0.01998791 60.70241 181.6564 0.6992712 0.00518815
```

Figure 21: Details of the AIC model