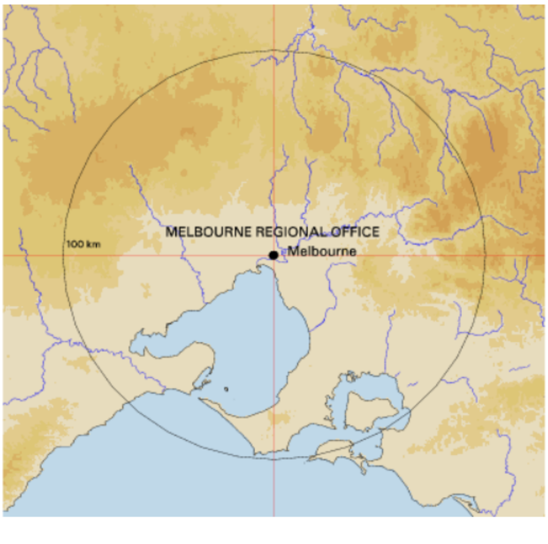
Report: Time Series

Analysis of Daily Maximum Temperature at Melbourne in Australia

Due on June 08, 2018

Fengyue Zhan, Wenyan Wong
fengyue.zhan@epfl.ch, wwong@epfl.ch

8 juin 2018



1 Introduction

Our chosen dataset is a record of the daily maximum air temperature[1] taken in Melbourne Regional Office (site opened in 1908 and closed at January, 6th 2015), Australia, with -37.81 DD (decimal degree) as latitude and 144.97 as longitude. The daily maximum air temperature is nominally recorded at 9 am local clock time. It is the highest temperature for the 24 hours leading up to the observation, and is recorded as the maximum temperature for the previous day. The dataset spans more than 100 years, from 1908 to 06 Jan 2015.

The dataset consists of eight columns : Product Code, Station Number, Year, Month, Day, Maximum Temperature, Period over which temperature is measured, Quality of recorded Temperature. The columns that are of interest to our study are Year, Month, Day, Maximum Temperature and Quality of recorded temperature.

The main objectives of our study are :

1. Construct a forecasting model, try to understand main seasonal cycles within the temperature series.
2. Recover data with bad quality by estimating actual value using a trained forecasting model.

2 Data cleaning

As mentioned in introduction, not all of the temperatures recorded in the dataset are equivalent in terms of quality. There are a total of 216 bad instances in the entire time series $\mathfrak{D}$, which is approximately 0.3% of all available instances. It can be seen in figure 1 that low quality instances are most prevalent at the beginning and end of the dataset. If we measure the longest segment of continuous data points without low quality data (ie. number of instances between the two blue lines), this is approximately 89.05% of the entire dataset.

Since our dataset is large (58444 instances), we have decided that the easiest solution would be to remove all instances at the beginning, $\mathcal{D}_b$ (segment before first blue line) and at the end, $\mathcal{D}_e$ (segment after second blue line), with remaining 42340 instances, $\mathcal{D}$, for further analyses. The points in $\mathcal{D}_e$ will be considered as a test sample and we'll use it to check the consistence of our model.

3 Removing seasonal component

3.1 Visualizing the time series

Due to the large number of instances, it is difficult to visualize the entire daily series with limited pixel resolution in R. Thus, we have produced another plot by focusing in the first 5% and 10% (figure 2). We observe that there is clearly a seasonal component with a period $T = 1$ year $= 365$ days approximately. This is reasonable and most likely an instance of climate

[1] http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_display_type=dailyDataFile&p_stn_num=086071&p_nccObsCode=122&p_startYear=2003&p_c=-1481645376.

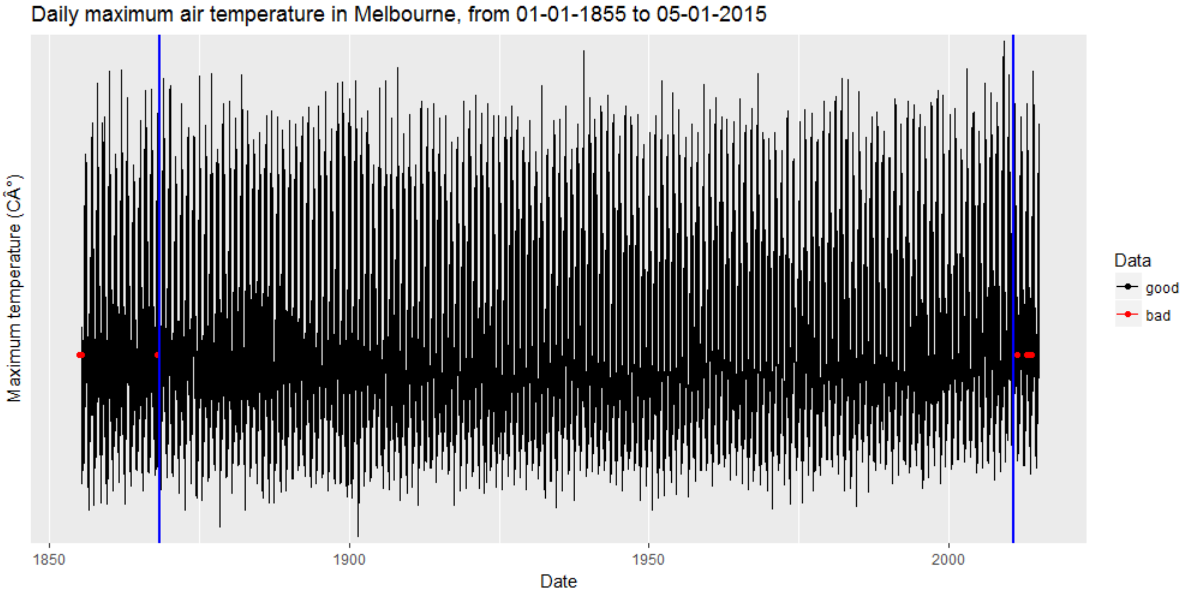


Figure 1 – Overview of the time series $\mathfrak{D}$. To ensure that our training data is clean, we simply identify the longest continuous segment of time series $\mathfrak{D}$ without bad quality data (blue lines indicate start and end of longest segment). The dataset is thus split into $\mathcal{D}_b$, $\mathcal{D}$, and $\mathcal{D}_e$. All of our further analysis are done using $\mathcal{D}$ with length $N = 42340$ days $= 116$ years, the data between the blue lines.

oscillation.

To better visualize the seasonal component, and confirm the value of the period, we split 5 consecutive years of daily data into 5 plots. If 1 year is effectively a period, the 5 plots should be very similar, and from figure 3, we observe that it is effectively the case.

These observations are strong indications that our temperature series is not a stationary time series. Since the theory taught in this course is based on stationary data, the first step of our study is to remove the seasonal trend. For this purpose, we tested three approaches : STL, differencing, and periodogram.

3.2 Seasonal decomposition of time series by Loess (STL)

The STL function in R decomposes a given time series (in an additive fashion) into *irregular*, *seasonal* and *trend components*. However, the output of STL is dependent on the the frequency attribute of the input time series (which is initially specified by user). We experimented with various commonly used frequency values (365, 7, 30, etc) for STL but in each case the returned *trend* component was not visually insightful and could not be interpreted well.

3.3 Differencing

Using differencing, we were able to transform the initial time series into a stationary series, but this is only achievable via at least two orders of differencing ($d > 2$). This eventually led to more difficulties in latter stages of our analysis, such as 'exploding' confidence bands during the forecasting step of ARMA models. Additionally, differencing introduce dependencies in our data, especially when d is big (which is our case). We conclude that this method is not well-suited for our dataset.

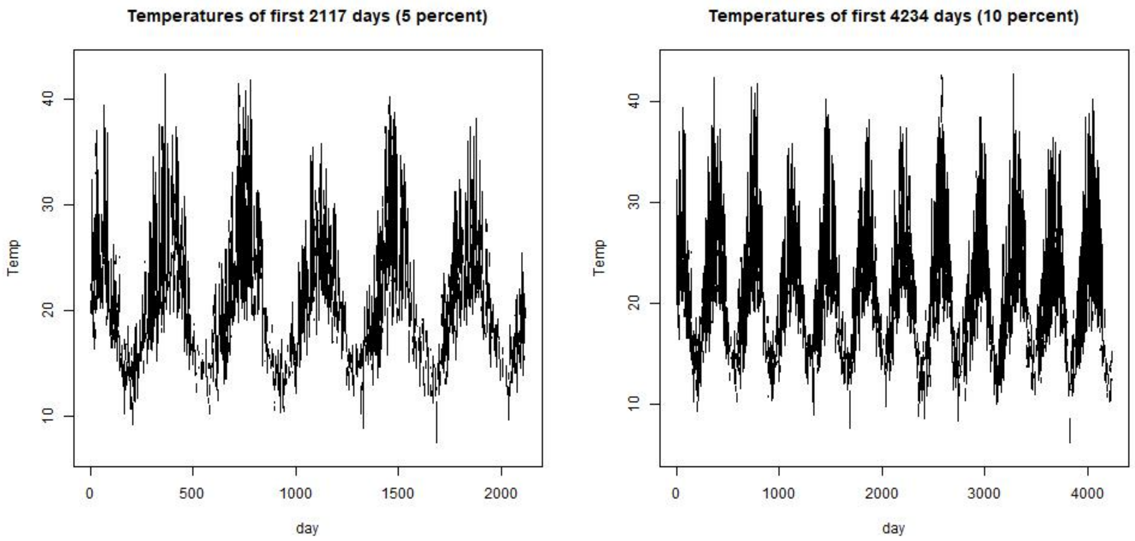


Figure 2 – Zoom on the time series : first 5% (left) and first 10% (right). We observe that there is clearly a trend due to a main seasonal component with a period $T = 1$ year $= 365$ days approximately. To apply time series models such as ARMA, we need to remove this seasonal component and obtain a stationary time series. For this, we decide to use the periodogram.

3.4 Spectral filter

Failing to obtain satisfactory results in the time domain, we proceeded to working with the frequency domain. The main idea of this approach can be summarized as : First, perform Fourier transform on original time series. Then, use a *spectral filter* to filter out frequencies that correspond to the seasonal components in the data. Finally, perform inverse Fourier transform to obtain data without seasonal trend.

To study the data in frequency domain, as periodogram is an approximation of spectral density, which is squared Fourier transform, we used both Fourier transform and periodogram to identify peaks that correspond to strong cycles in the time series.

A plot of the periodogram of the time series $\mathcal{D} = \{d_1, d_2, ...d_N\} = (d_t)_{t=1,...,N}$ with $d_{t_0} = \mathcal{D}(t = \text{day } t_0)$ and N its length (figure 4) shows two significantly strong peaks that stand out from the noise, at around $f_1 = 2,73.10^{-3}$ Hz and around $f_2 = 5,5.10^{-3}$ Hz. Quick calculations show that this is most likely due to annual and biannual seasonalities in the time series ($T_1 = 366$ days and $T_2 = 182$ days, and the amplitude of the second one is approximatively 10 times smaller than the previous one according to figure 5).

Let's note $S_{1T} = (s_t^1)_{t=1,...,N}$ and $S_{2T} = (s_t^2)_{t=1,...,N}$, respectively the time series of the annual and the biannual seasonal components, and $\mathcal{R}_{filter} = (r_t^f)_{t=1,...,N}$ the residual of filtering. The filtering operation described above corresponds to the operation 1 :

$$\mathcal{R}_{filter} = \mathcal{D} - (S_{1T} + S_{2T}) \quad \Leftrightarrow \quad \forall t = 1, ..., N, \quad r_t^f = d_t - (s_t^1 + s_t^2) \tag{1}$$

$S_{1T} = (s_t^1)_{t=1,...,N}$ and $S_{2T} = (s_t^2)_{t=1,...,N}$ can be thought of as the output of the IFFT of the signals 'discarded' by the spectral filter in the frequency domain. A plot of these quantities are given in figure 5.

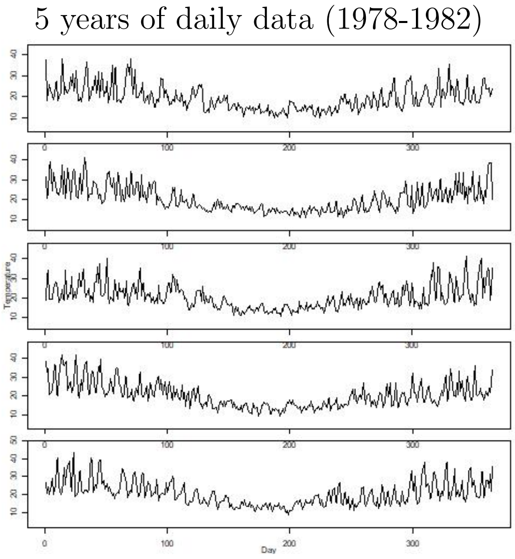


Figure 3 – To confirm the period of the seasonal component, we split 5 consecutive years of daily data into 5 plots. From top to bot, we have respectively 1978, 1979, 1980, 1981 and 1982. We observe that the 5 plots is very similar, that means we have effectively 1 year = 365 days

4 Normalization

4.1 Approach : Daily mean and daily variance

We obtained $\mathcal{R}_{filter}$ from expression 1, and in figure 5, we note that the seasonal component is effectively removed but our data is still clearly not stationary : the variance is dependent on time. Nonetheless, we observe that the variance is periodic (ie. high variance during the start and end of a year, low variance during midyear). Thus, the variance can be modelled using a deterministic function which takes day of the year as an input. Since our goal is to obtain a time series with constant variance (equals to 1), the idea is to normalize our data using the daily variance (obtained from the deterministic function).

Let $D_k = (d_t)_{t \equiv k[T_1]}$ be a subset of $\mathcal{D}$ with $k = 0, ..., T_1 - 1$. We define the daily variance $V = (v_k)_{k=1,...,T_1}$ and the daily mean $M = (m_k)_{k=1,...,T_1}$ with v_k and m_k respectively the variance and the mean of D_k. Let's note $\mathcal{Y} = (Y_t)_{t=1,...,N}$ the normalized version of $\mathcal{R}_{filter}$ such that :

$$\mathcal{Y} = \frac{\mathcal{R}_{filter} - M}{V} \quad \Leftrightarrow \quad \forall t = 1, ..., N, \quad Y_t = \frac{r_t^f - m_k}{v_k} \quad \text{with} \quad t \equiv k\,[T_1] \tag{2}$$

We have plotted the daily variance and the daily mean in figure 6. The variance, effectively, seems to be periodic and deterministic, and the daily mean constant. The last point is coherent because the seasonal trends is already removed via filtering, so we are supposed to observe just a constant which is the mean of daily temperature. The difference between the center (points are more concentrated) and the border (points are more dispersed) is due to the fact that temperature variate differently in winter and in summer.

4.2 Result : Stationary time series

From expression 2, we obtained our data before normalization, $\mathcal{R}_{filter}$, and after normalization, $\mathcal{Y}$ in figure 7. We observe that the resulting time series have approximately constant variance that is not dependent on time, which is the purpose of our normalization.

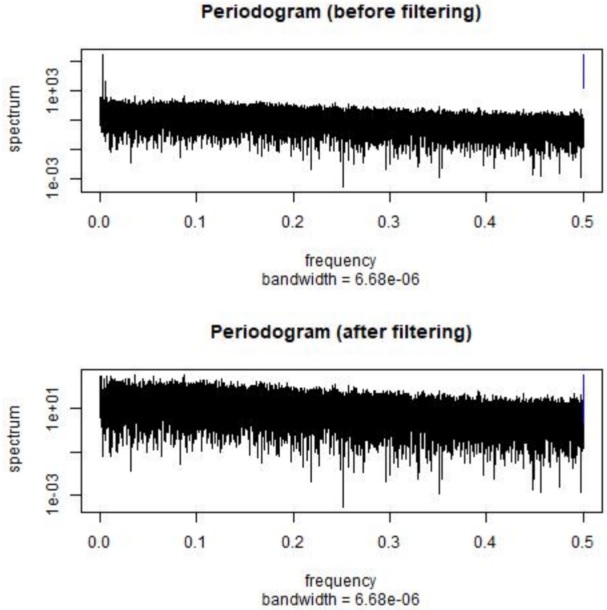


Figure 4 – Periodogram before (top) and after (bot) filtering. The two peaks around zero characterise our seasonal components, thus, we filtered it to remove the trend. The filter we used give us directly the removed signal in time space, so this operation is perfectly invertible.

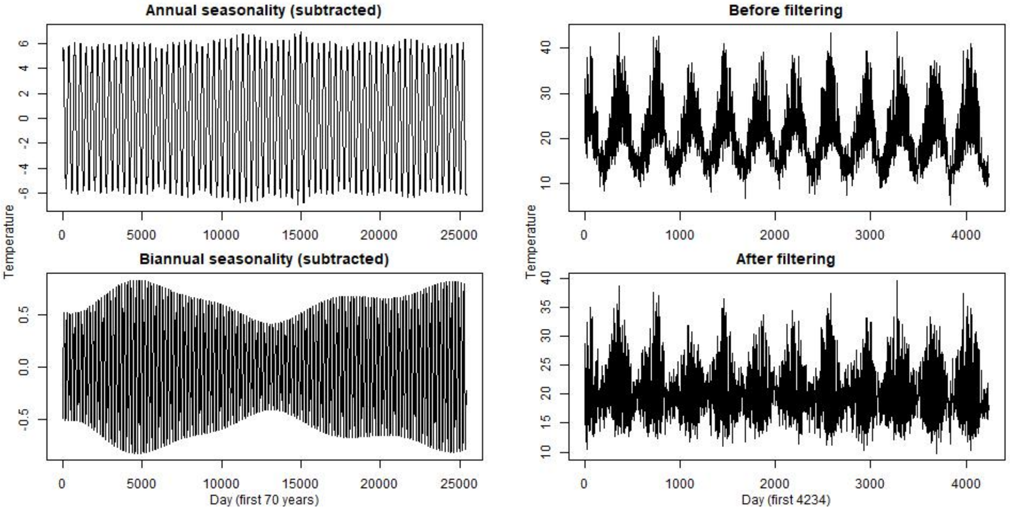


Figure 5 – We subtracted $S_{1T} = (s_t^1)_{t=1,\ldots,N}$ and $S_{2T} = (s_t^2)_{t=1,\ldots,N}$, respectively the annual (left-top) and the biannual (left-bot) seasonal components (with 10 times smaller amplitude), obtained from the periodogram, from the data. The right plots show our first 10% data before (right-top) and after (right-bot) filtering. We note that the seasonal component is removed but our data is still clearly not stationary : the variance depend on time. As it seems to be a deterministic and periodic (T=365 days) function, we can normalize our data to have a variance constant equals to 1.

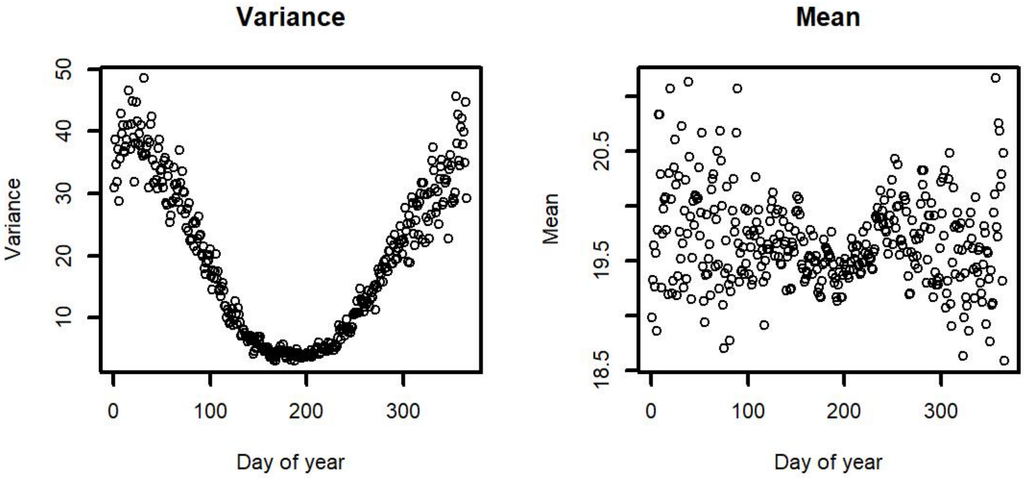


Figure 6 – Daily variance $V = (v_k)_{k=1,\dots,T_1}$ (left) and daily mean $M = (m_k)_{k=1,\dots,T_1}$ (right) of the dataset $\mathcal{D}$. As we have a long time series (116 years), each v_k and m_k are computed over 116 points. We observe that the variance is effectively periodic as predicted from previous figures.

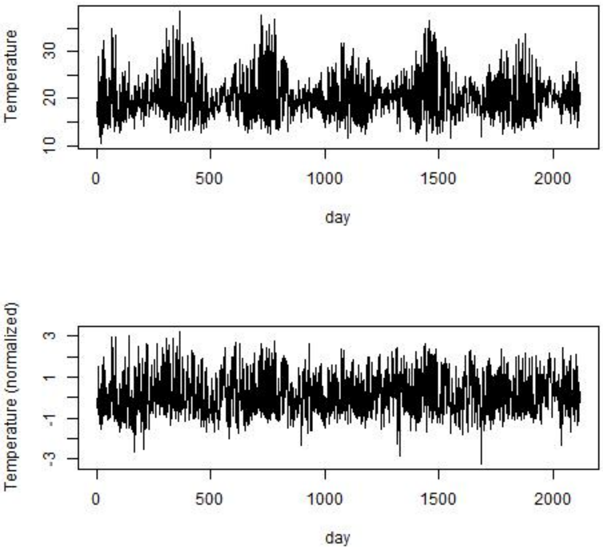


Figure 7 – Our data before normalization, $\mathcal{R}_{filter}$ (top), and after normalization, $\mathcal{Y} = (Y_t)_{t=1,\dots,N}$ (bot). For data point, we subtract the corresponding daily mean and then divide by the corresponding standard deviation according to expression 2. We observe that the resulting time series have approximately constant variance that is not dependent on time, but there is strong dependence in the time series as shown in the ACF (figure 8). We can now model this resulting series using time series models such as ARMA.

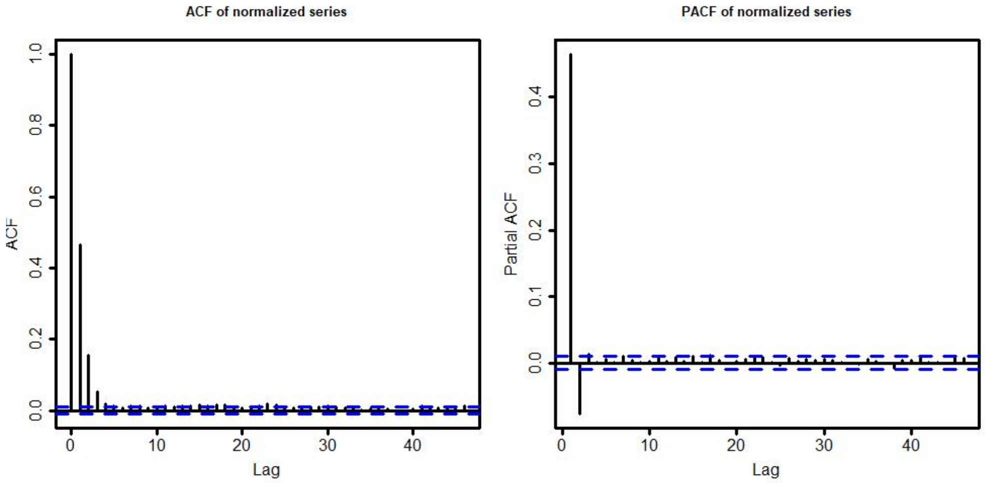


FIGURE 8 – Autocorrelation function, ACF (top), and partial autocorrelation function, PACF (bot) of our normalized data. We observe a tail-off in the ACF and a cut-off at $lag = 2$ in the PACF. These characteristics suggest that an AR(2) could be a good model for our data.

5 Modeling

5.1 AR(2), AR(3) and AR(4)

From the previous normalized residual, we wondered if there is any dependence in the times series, and the answer is positive as shown in the ACF (figure 8). Now, we can proceed to model these dependences in time series using models such as ARMA, GARCH, etc.

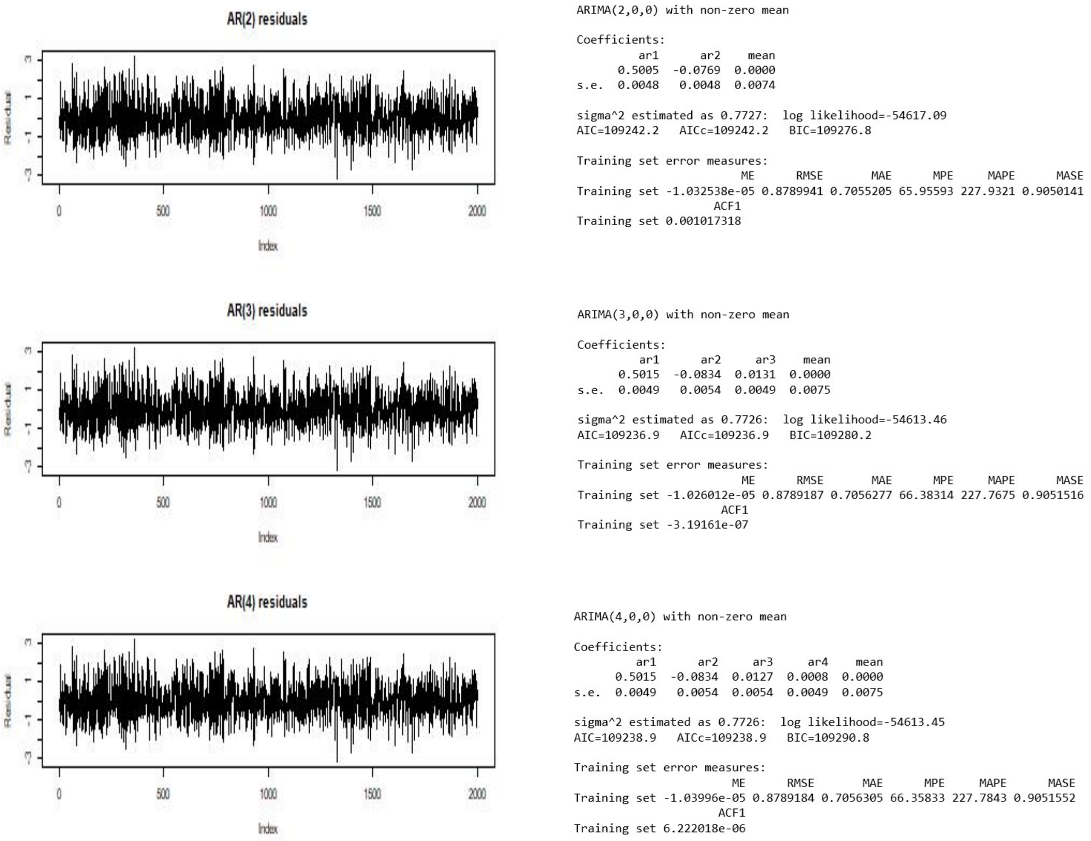
To choose our model, we plot the autocorrelation function (ACF) and the partial autocorrelation (PACF) of $\mathcal{Y}$ in figure 8. We observe a tail-off in the ACF and a cut-off at $lag = 2$ in the PACF. These characteristics suggest that an AR(2) could be a good model for our data. Nonetheless, we tested also AR(3) and AR(4) model because as its are more complex, its should give us better results. The results of fits are shown in figure 9.

In this figure 9,we note that the two first coefficients are very similar from one model to another and each supplementary coefficient is approximatively 10 times smaller than the previous one. In addition, the AIC indexes suggest that there are few difference between the three models in term of lost information ($\simeq 10^{-5}\%$). Finally the simplest model, AR(2), seems to be the best model.

5.2 Stationarity of the residuals : Ljung-Box test

According to all our previous analysis on the detail of fits, an AR(2) model for $\mathcal{Y} = (Y_t)_{t=1,\ldots,N}$ seems to be the best choice. To confirm this point, we plotted the ACFs and the PACFs of the residuals in figure 10.

We observe that most of the dependence in the mean of the time series can be well explained by all three AR models mentioned above, thus, the residuals are uncorrelated for the three models. So since the three models give us the same result, our choice of simplest model AR(2) is justified. Further comparisons between the three models and have been done but all the plots we obtained are similar to the ones from AR(2). These supplementary plots can be found in Annexe.



```
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1      ar2    mean
      0.5005  -0.0769  0.0000
s.e.  0.0048   0.0048  0.0074

sigma^2 estimated as 0.7727:  log likelihood=-54617.09
AIC=109242.2   AICc=109242.2   BIC=109276.8

Training set error measures:
                        ME      RMSE       MAE      MPE     MAPE      MASE
Training set -1.032538e-05 0.8789941 0.7055205 65.95593 227.9321 0.9050141
                    ACF1
Training set 0.001017318
```

```
ARIMA(3,0,0) with non-zero mean

Coefficients:
         ar1      ar2     ar3    mean
      0.5015  -0.0834  0.0131  0.0000
s.e.  0.0049   0.0054  0.0049  0.0075

sigma^2 estimated as 0.7726:  log likelihood=-54613.46
AIC=109236.9   AICc=109236.9   BIC=109280.2

Training set error measures:
                        ME      RMSE       MAE      MPE     MAPE      MASE
Training set -1.026012e-05 0.8789187 0.7056277 66.38314 227.7675 0.9051516
                     ACF1
Training set -3.19161e-07
```

```
ARIMA(4,0,0) with non-zero mean

Coefficients:
         ar1      ar2     ar3     ar4    mean
      0.5015  -0.0834  0.0127  0.0008  0.0000
s.e.  0.0049   0.0054  0.0054  0.0049  0.0075

sigma^2 estimated as 0.7726:  log likelihood=-54613.45
AIC=109238.9   AICc=109238.9   BIC=109290.8

Training set error measures:
                       ME      RMSE       MAE      MPE     MAPE      MASE
Training set -1.03996e-05 0.8789184 0.7056305 66.35833 227.7843 0.9051552
                     ACF1
Training set 6.222018e-06
```

Figure 9 – Even if the tail-off in the ACF and the cut-off in the PACF suggest an AR model, since we are not sure about the cut-off at $lag = 2$, we decided to include AR(3) and the AR(4) models in our analyses. The residuals of models are presented on left, and their respective descriptions on the right. We note that the two first coefficients are very similar from one model to another and each supplementary coefficient is approximatively 10 times smaller than the previous one. In addition, the AIC indexes suggest that there are little differences between the three models in terms of lost information ($\simeq 10^{-5}\%$). The simplest model, AR(2), appears to be the best model.

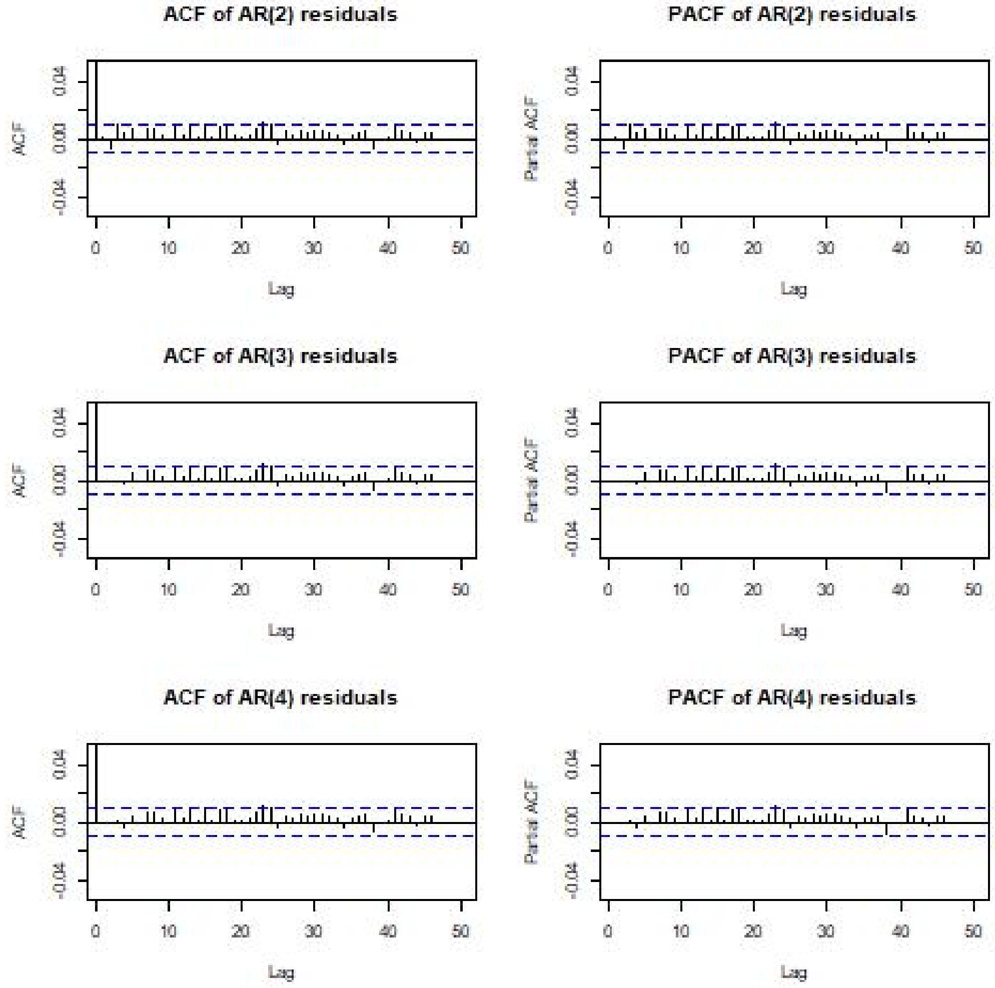


Figure 10 – We plot the ACF (right) and the PACF (left) of the residual obtained by applying AR(2), AR(3) or AR(4) model. We observe that the structures inside the data are almost totally removed, thus, the residual is effectively uncorrelated for the three models. We note also that since the three models give us the same result, our choice of simplest model AR(2) is justified.

Now, we define $\mathcal{E} = (\epsilon_t)_{t=1,...,N}$, as the residuals of AR(2) plotted in figure 9, and the coefficients $\alpha = 0.5005$ and $\beta = -0.0769$ such that :

$$\forall t = 2, ..., N, \quad Y_t = \alpha Y_{t-1} + \beta Y_{t-2} + \epsilon_t. \tag{3}$$

Before proceeding to forecasting with our AR(2) model, we need to verify that :

1. $\mathcal{E}$ fulfils weak stationarity, ie. it has constant mean, and its ACF is only dependent on the lag
2. ϵ_t are independent and identically distributed (iid) random variables

To test for stationarity of $\mathcal{E}$, we make use the Ljung-Box Test. The Ljung-Box Test would test whether the residual is a white noise series. Since a white noise must be stationary, it follows that if the $\mathcal{E}$ passes the Ljung-Box Test, it must be stationary. The results of our test (figure 11) shows that the residuals of AR(2) passes the Ljung-Box Test for different ranges of lag values. We conclude that the residuals of AR(2) is stationary.

To check whether the ϵ_t are independent and identically distributed, we try verify that it is normally distributed and uncorrelated.

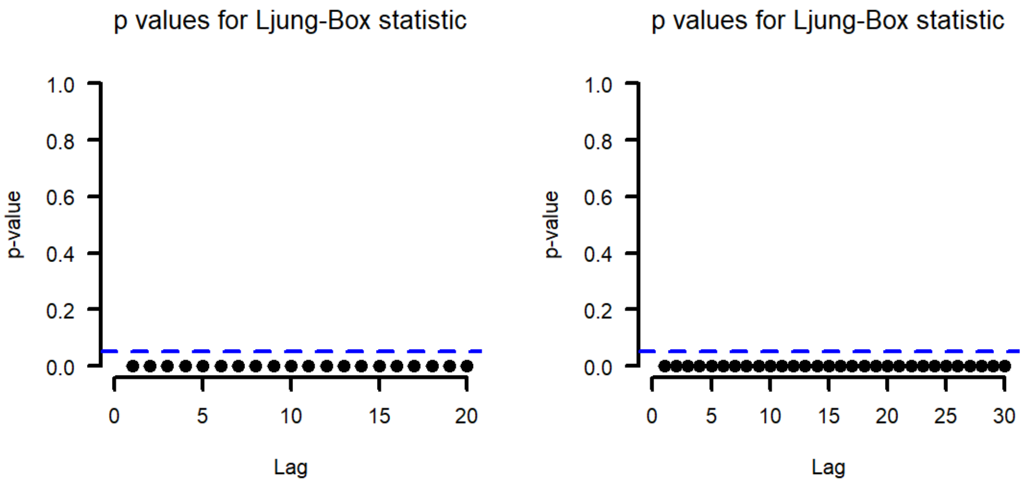


Figure 11 – We applied our AR(2) model and we tested if the residual is stationary or not. We plotted the Ljung-Box test with different range of lag. We note that all the p-values are under the threshold, thus the data seems to be stationary.

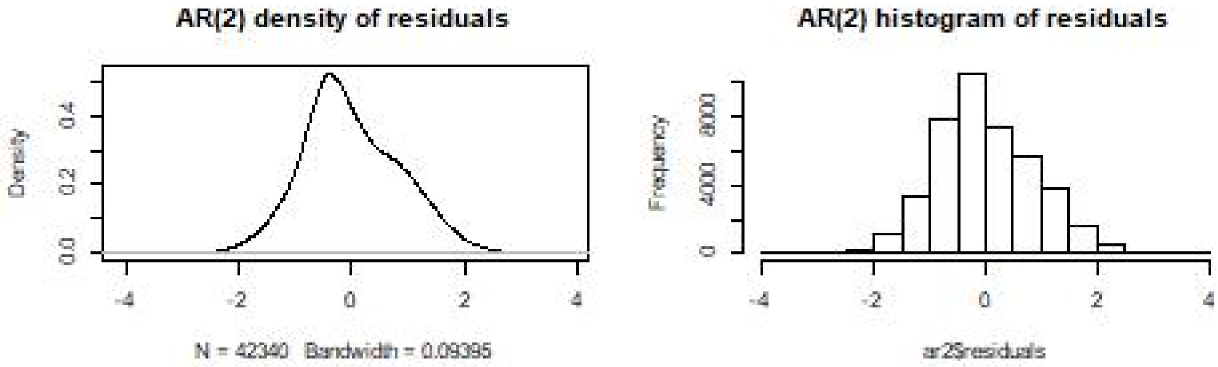


Figure 12 – The aim now is to verify if our ϵ_t is independent and identically gaussian distributed or not. For that, we first plot the density (left) and the histogram (right) of the residual. We observe that it is clearly not completely gaussian due to the asymmetry. Q-Q plots of the AR(2) residuals (figure 13) verify that they are indeed not fully gaussian.

For the correlation, we can already verify this using the ACF and PACF plots presented in figure 10. Clearly, there is no noticeable correlation among the data, except at $lag = 0$.

To verify if the residual is normally distributed, we plotted its histogram [2] and its density in figure 12. We observe that it is not completely gaussian due to the asymmetry, but it could be acceptable with incertitude bands, thus, we plot a Q-Q plot [3] to fix this point in figure 13. We note that the residual is not gaussian distributed because even at the center, the solid black line and the solid blue line failed to collapse. Nonetheless, we accepted this result and proceeded to check for the forecasting performance of the AR(2) model.

[2] For density plots and histograms of AR(3) and AR(4) model, please contact the authors.

[3] For QQ-plot of AR(3) and AR(4) model, please contact the authors.

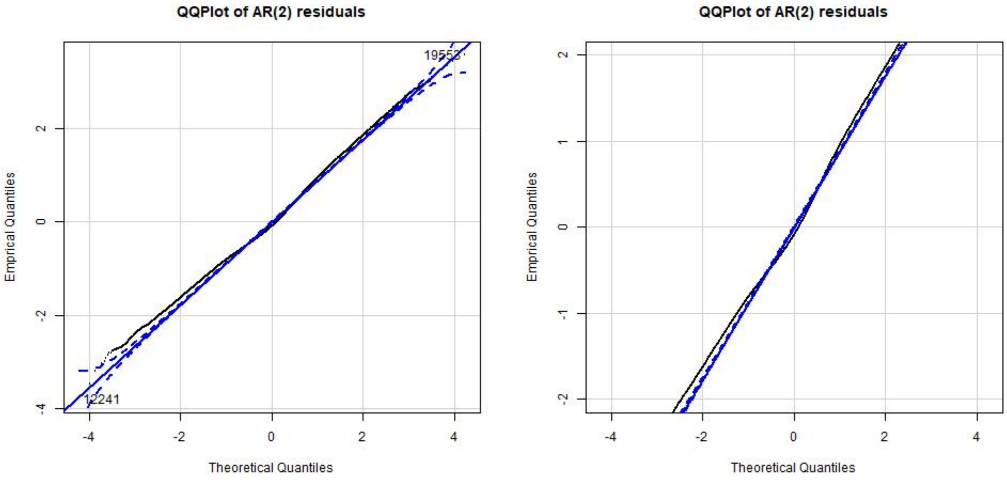


Figure 13 – The Q-Q plot of the residual (left) and its zoomed version (right). Due to technical limitation, we are just able to change the y-axis. We note that the residual is not gaussian distributed because even at the center, the solid black line and the solid blue line don't collapse.

6 Forecasting

6.1 Result of prediction

According to our AR(2) model with $\mathcal{E}$ stationary and ϵ_t approximately i.i.d (test of gaussianity failed), we predicted the future values of Y_t. Let's note $\overline{Y_t}$ the predicted value of Y_t for $t \geq N+1$.

We plot the forecasting result [4] for the residual of AR(2) model in figure 14. The blue solid line is the predicted values, the gray region is the 80% confidence band, and the light gray is the 95% confidence band. We observe that the forecasted values quickly approaches the mean after few days ($\simeq 3-4$ days), which is coherent with our model.

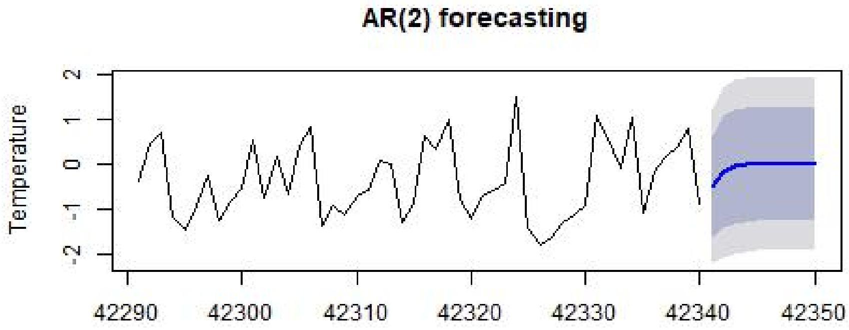


Figure 14 – The result of forecast (blue part) as a function of time t (in day). As the model is very simple, the best prediction is just the mean when t is large enough.

6.2 Recovering original time series

We can transfer the previous forecasting result to our temperature data according to the following formula :

$$\forall t \geq N+1, \quad \overline{d_t} = v_k \overline{r_t}^f + m_k = v_k \overline{y_t} + m_k + s_t^1 + s_t^2 \quad \text{with} \quad t \equiv k\,[T_1] \tag{4}$$

[4] For forecasting result using AR(3) model, please contact the authors.

The expression 4 is possible because S_{1T}, S_{2T}, V and M are deterministic functions, thus the prediction of Y_t give directly the prediction of d_t for $t \geq N+1$.

We now proceed to visualize the predictive power of our model, by comparing $(\overline{d_t})_{t \geq N+1}$ to $\mathcal{D}_e$. ($\mathcal{D}_e$ consist of the observed future values that were not used to fit the AR(2) model.)

We first plotted[5] the fitted values and the data on the same plot in figure 15. It give us an idea of how well our AR(2) model fit the data, and the result is acceptable because the fitted values and training data coincide relatively well, without collapsing totally (which is a sign of over-fitting).

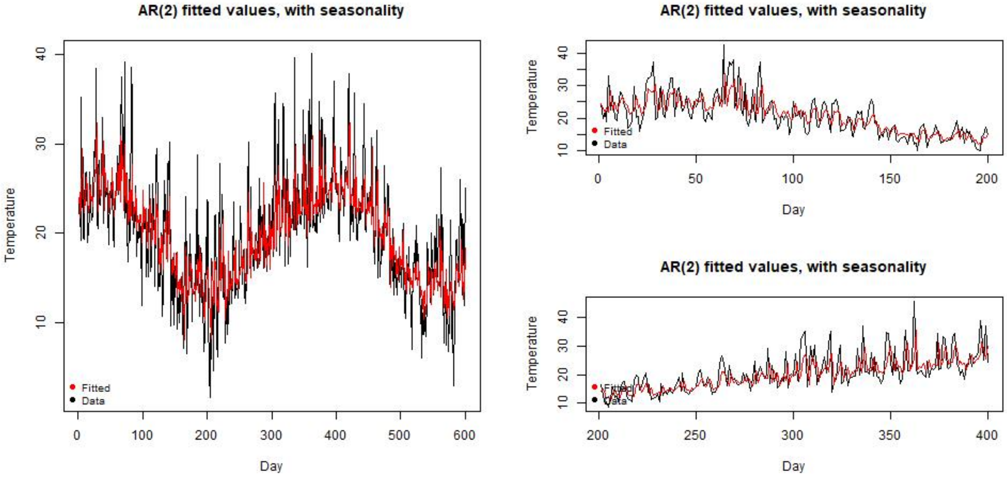


Figure 15 – The AR(2) fitted values (red) and our data (black) on the same plot. It give us an idea of how well our AR(2) model fit the data, and the result is satisfying because both coincide relatively well, without collapsing totally (which is an insight of over-fitting).

To visualize the predicted series[6], we plot the forecasted values using the expression 4 in figure 16. We note that almost all the future observed data (first values of $\mathcal{D}_e$) we removed are contained within the 95% confidence interval. Additionally, the confidence band is bigger during summer (high temperature) and smaller during winter (low temperature), which is consistent with the observation that the temperature vary with higher uncertainty in summer (thus a bigger confidence interval to account for this) than in winter. This result is also in line with our initial observation that the magnitude of variance oscillates with a predictable manner.

7 Conclusion

In this study of the daily maximum air temperature taken in Melbourne Regional Office in Australia, we have successfully constructed a consistent forecasting model. To build this model, we divided the time series into training set and test set and we made our analysis on the first set.

During the analysis, we start by removing the seasonal trend. For that, we tested several methods like differencing, STL and we finally used the Fourier transform which seems to be the most natural to deals with seasonal components. This method revealed 2 interesting periods : 366 days and 180 days.

Since the filtered series is not stationary, we used to normalization using the daily mean and

[5] For the same plot using AR(3), please contact the authors.

[6] Supplementary plots of the predicted temperature with different range of time (day) can be furnished, please contact the authors.

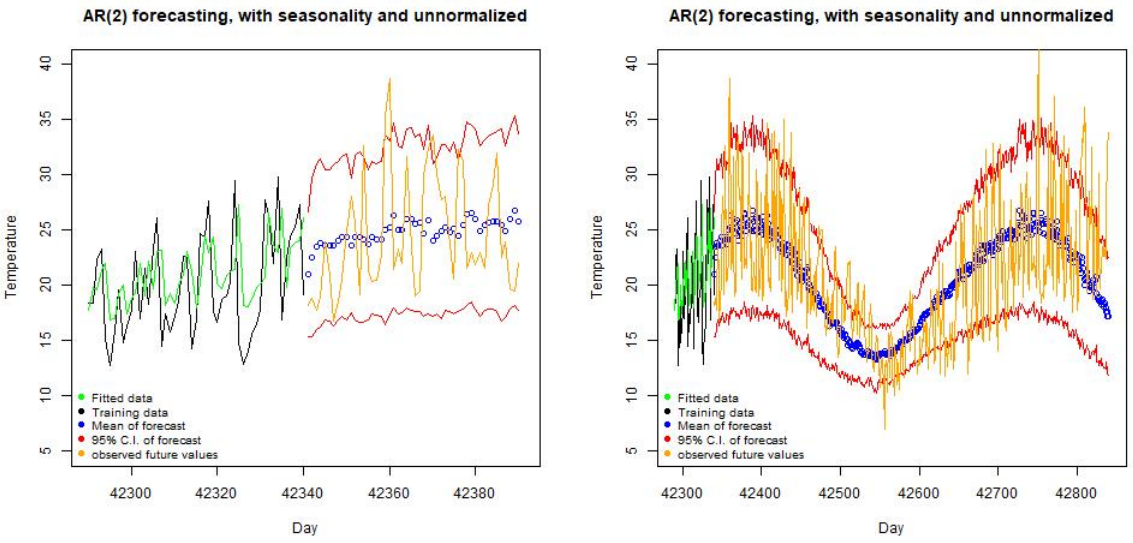


Figure 16 – The predicted maximum daily temperature as function of time (day). The left plot contain 50 predicted values and the right plot, 500 predicted values. We observe that the 95% confidence band (red lines) seems to contain effectively 95% of measured values (yellow line). Additionally the confidence band is bigger in summer (high temperature) and smaller in winter (low temperature), which is consistent with the observation that the temperature variate more in summer (so a bigger band to cover theses variations) than in winter.

the daily variance computed from the training data. By analysing the ACF and PACF of the normalized residual, we selected an AR(2) model which ended to be the best model compared to AR(3), AR(4) or GARCH model.

Before proceeding to forecasting, we conducted the Ljung-Box test and gaussianity test (QQ-plots) on the AR(2) residuals. The second test shows that the AR(2) residuals are not fully Gaussian but we believe this is an acceptable result. Forecasting results confirm that the model can still predict relatively well even though the AR(2) residuals are only approximately normally distributed. Indeed, most of the future observed vales (500 of them) are contained within the 95 percent confidence band of forecasted values.

In our future work, we can take into account the non gaussianity of our residual and bootstrapping. This method aims to generate possible future observations we can make from the distribution of the residual. And with a large number of possible realisations, we can compute the confidence band, which should be better than what we obtained without considerations of the result of gaussianity test.

8 Annexe

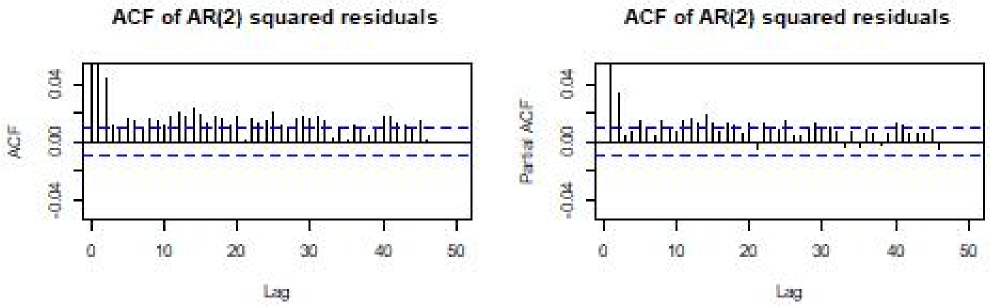


Figure 17 – We plot the ACF (left) and the PACF (right) of the square residual of AR(2), AR(3) and AR(4). We not that the plots are extremely similar. These plot show some structure in the squared data and thus suggest a GARCH model, but we choose the simplest model, AR(2), because it seems to work well.

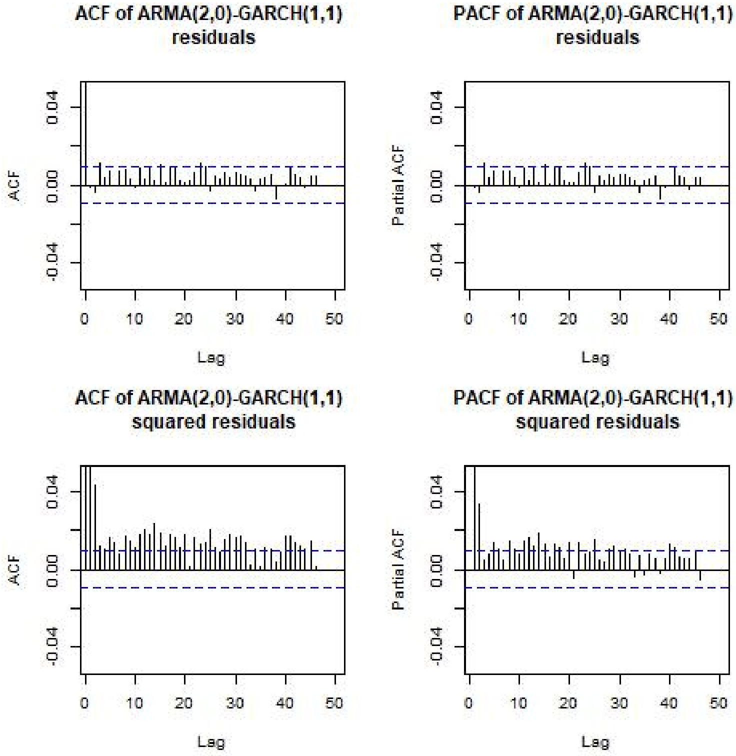


Figure 18 – The ACF and PACF of the residual of an ARMA(2,0)-GARCH(1,1) model. Similar to the AR(2), the dependence in mean of the time series is well explained by this model. However, it still fails to capture the dependence in variance, thus, we selected the AR(2) model which is far more parsimonious.