Data analysis using the ARMA model

March 21, 2024

1 Introduction

The purpose of the report is to analyze real data using the ARMA model. The data is from the website kaggle.com[1]. It concerns the weather in London from 1979 to 2020. The data was collected by a weather station near Heathrow Airport in London, UK. The report will focus on the average daily temperature from 1998 to 2001.

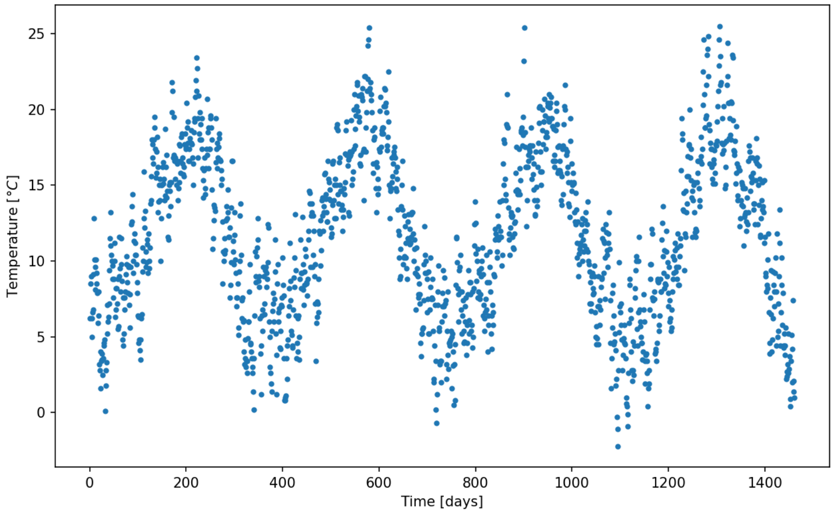

Figure 1: The scatter plot of the data.

The plot 1 visualizes the data we will be analyzing, specifically the measurements of average daily temperature on consecutive days from 1998 to 2001.

2 Model ARMA

We will fit the ARMA(p, q) model to the data:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + ... + \theta_q Z_{t-q},$$

where $Zt, t \in \mathbb{Z}$ is white noise.

The polynomial

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p$$

is called the autoregressive polynomial and

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + ... + \theta_q z^q$$

is called the moving average polynomial.

To fit the model, we estimate p and q and find the coefficients of the polynomials $\theta(z)$ and $\phi(z)$. The autocorrelation function (ACF) is defined as

$$ACF(h) = \frac{\gamma(h)}{\gamma 0},$$

where h is the lag, $\gamma(h)$ is the autocovariance function for lag h.

The partial autocorrelation function (PACF) is defined as

$$PACF(h) = \mathbf{\Gamma}_h^{-1}\gamma(h),$$

where $\gamma_h = [\gamma(1), \gamma(2), \ldots, \gamma(h)]'$, $\mathbf{\Gamma}_h = [\gamma(i-j)]_{i,j=1}^{h}$, $\gamma(h)$ is the autocovariance function for lag h.

3 Data Preparation for Analysis

The examined data have no missing values, and all measurements have reliable values. Therefore, there is no need for any changes.

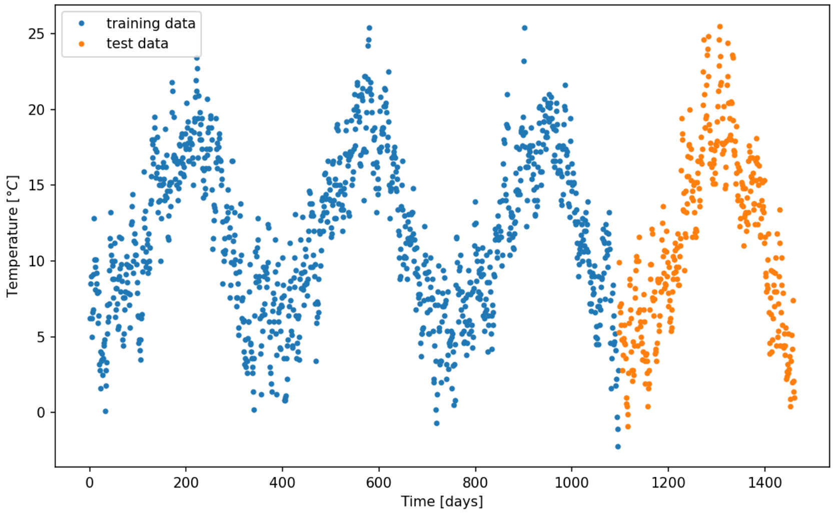

Figure 2: The scatter plot of the data divided into training and testing data.

From the data, observations for the test set from the last year considered in the analysis, that is 2001, are extracted. They will be used to check the quality of forecasts for future observations. The training data are of length 1096 and the test data are of length 365.

3.1 Time Series Decomposition

We will proceed with the Wald decomposition of the time series. That is, finding such deterministic functions $s(t)$ (periodic function) and $m(t)$ (deterministic function) in the equation

$$Y_t = s(t) + m(t) + X_t,$$

where $\{Y_t\}_{t=1}^{\infty}$is the time series, whose realizations are observed raw data, so that $\{X_t\}_{t=1}^{\infty}$ is a weakly stationary time series. We will also check whether the data are stationary using the ADF test.

The fitted trend function is a linear function of the form

$$m(t) = 8.59e - 4 \cdot t + 11.26,$$

and the seasonality function has the form

$$s(t) = -6.339 \cos(0.017t + 24.809).$$

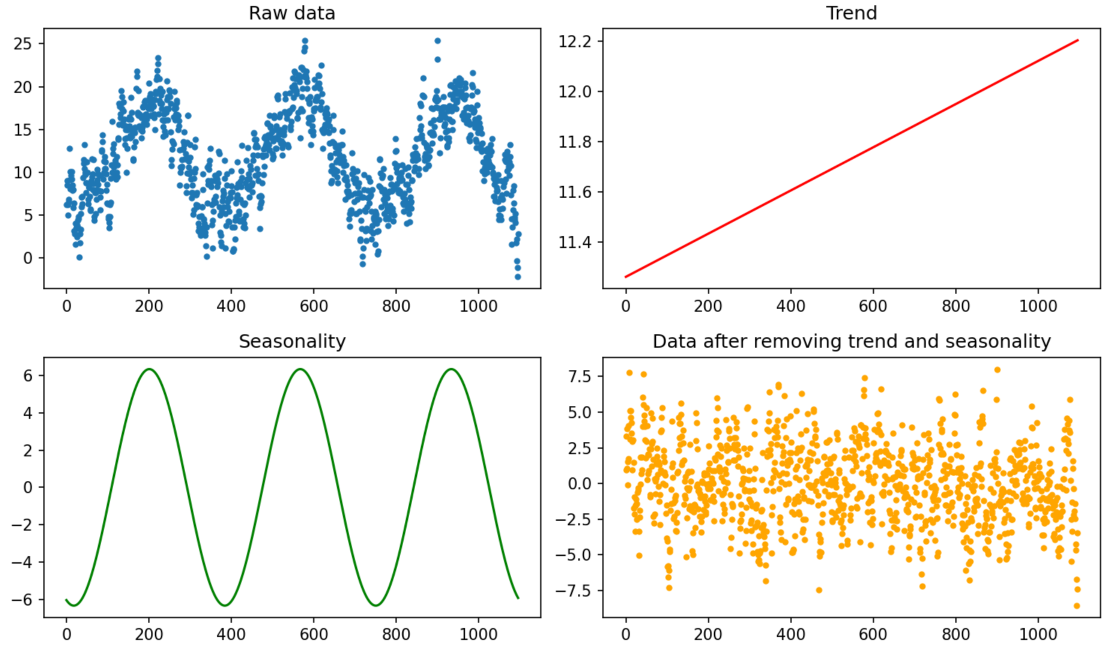

Figure 3: Plots of raw data, trend, seasonality and data after removing trend and seasonality.

In figure 3 plots of raw data, deterministic functions $m(t)$ and $s(t)$ and data after removing trend and seasonality are presented.

In figure 4, plots of ACF and PACF for $h \in \{0, 1, \ldots, 50\}$ for data at various stages of removing deterministic factors are shown. For raw data and data after removing the trend, ACF and PACF assume similar values. It can be seen that for all h, ACF functions take values significantly different from 0. The PACF values quickly approach 0 and are much closer to it. After removing the trend and seasonality, ACF values quickly converge to 0; already for $h = 8$, its value does not significantly differ from 0. The PACF plot does not differ much from those before removing deterministic functions; its values quickly approach 0 (already from $h = 4$, most values are not significantly greater than 0).

We will check whether the time series before and after decomposition is stationary. We will use the ADF test for this purpose, for which:
H_0: the time series is non-stationary,
H_1: the time series is stationary.

The ADF test		
data	statistic	p-value
raw	-2.21	0.20
without trend	-2.17	0.22
without trend and seasonality	-11.41	7.25e-21

Table 1: The ADF test for raw data and after removing trend and seasonality.

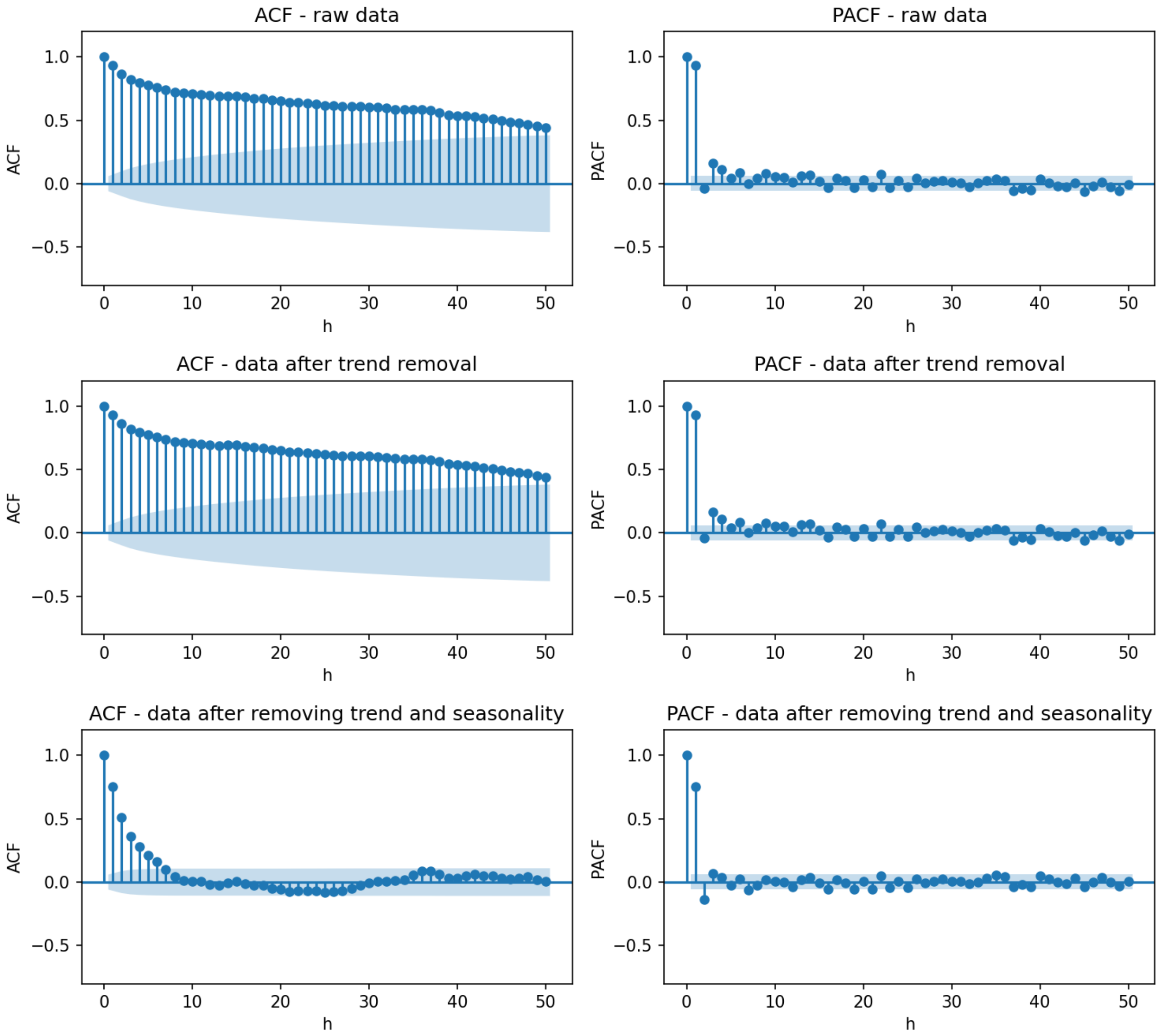

Figure 4: The plots of ACF and PACF for raw data, data after removing trend, and data after removing trend and seasonality.

Table 1 presents the test statistic and p-value for raw data, after removing the trend, and after removing both trend and seasonality. It can be seen that in the first two cases $p - wartosc > 0.05$, so we do not have evidence to reject the null hypothesis, and thus, we accept that raw data and data after removing only the trend are non-stationary.

After removing both deterministic factors $p - value < 0.05$, so we reject H_0 in favor of H_1, that is we accept that these data are stationary.

4 Modeling Data Using the ARMA Model

4.1 Selecting the Model Order

We will start by selecting the order of the ARMA(p, q) model. We will use the following information criteria for this purpose:

- AIC with the test statistic: $AIC = -2\ln L + 2(p+q)$,
- BIC with the test statistic: $BIC = (p+q)\ln(n) - 2\ln L$,

where L is the maximized value of the model likelihood function.

The selected order of the model	
p	3
q	6
Information criterion	
criterion	statistic value
AIC	4366.775
BIC	4421.768

Table 2: Selected ARMA(p, q) model order and the values of selected information criteria statistics.

All considered criteria achieved the smallest value for the ARMA(p=3, q=6) model. The selected model order and the values of the statistics are presented in table 2.

4.2 Parameter Estimation

After selecting the model order, we will proceed to estimate the parameters of the model. We will use the method of least squares for this purpose.

Estimated parameters	
ϕ_1	0.9191
ϕ_2	-1.0613
ϕ_3	0.6620
θ_1	-0.0548
θ_2	0.8329
θ_3	0.1006
θ_4	-0.0073
θ_5	-0.0666
θ_6	0.0839
σ^2	3.0805

Table 3: The estimated parameters of the ARMA(3, 6) model and the variance of the white noise.

Table 3 presents the values of the estimated parameters of the ARMA(3, 6) model and the variance of the white noise.

5 Model Fit Assessment

5.1 Confidence intervals for ACF

We will determine confidence intervals for the ACF at the $\alpha = 0.1$ level for the ARMA series with the previously estimated parameters. To do this, we will generate 1000 trajectories of the series with such p, q and for each, we will calculate the autocorrelation function values for lags $h = 0, 1, 2..., 30$. Then, for each h based on these 1000 values, we will determine quantiles of order $1 - \frac{\alpha}{2} = 0.95$ and $\frac{\alpha}{2} = 0.05$. If our model is well fitted to the data, we expect the values of the sample ACF to fall within the determined intervals 90% of the time.

Figure 5 shows the plot with the determined intervals. All sample autocorrelation values fall within their confidence intervals. We expected 90%, but we only check this for a small number (30) of h.

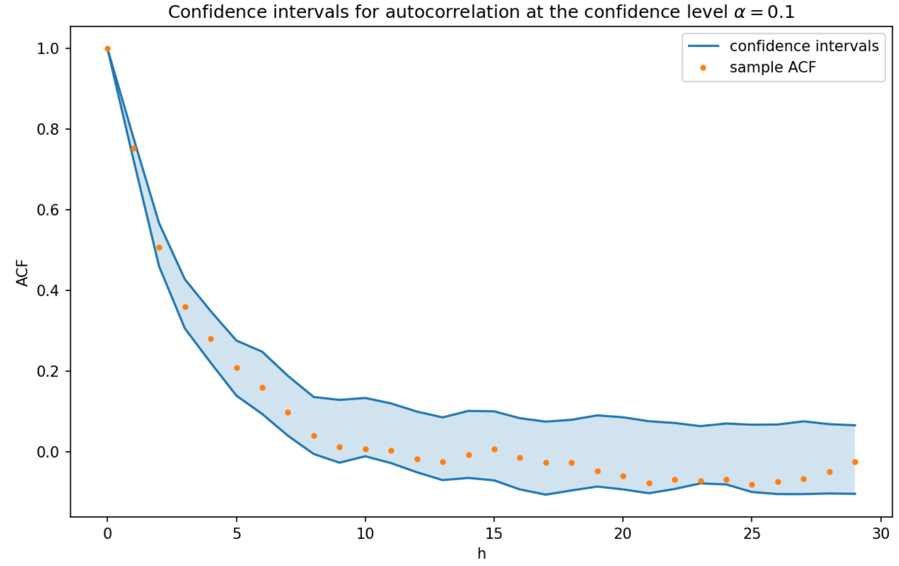

Figure 5: Confidence intervals for ACF.

5.2 Confidence intervals for PACF

Now we will determine confidence intervals for PACF for a series with the selected p, q and parameters. We will proceed similarly as in section 5.1, this time calculating PACF values, also for $h = 0, 1, 2, ..., 30$ and for 500 realizations of the ARMA series.

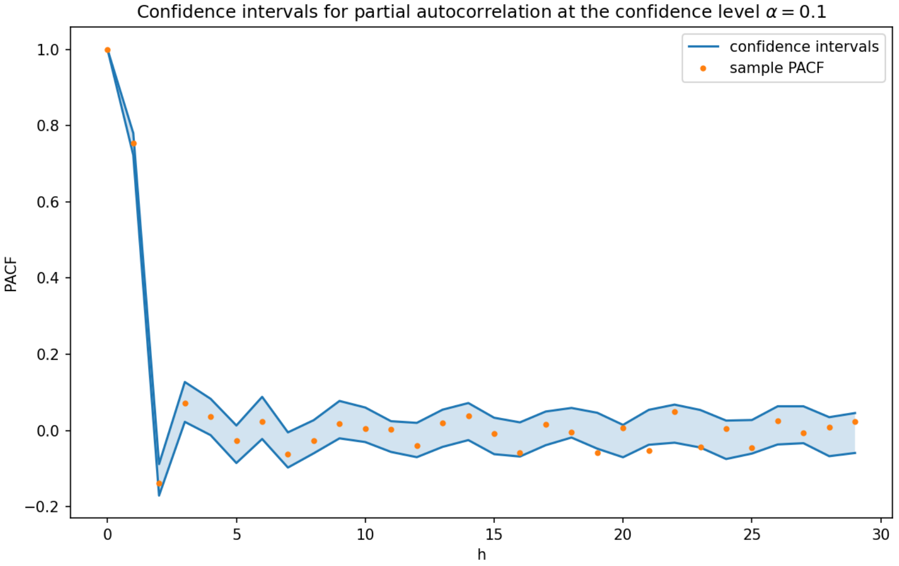

Figure 6: Confidence intervals for PACF.

Figure 6 presents the determined confidence intervals. Just like with ACF, the sample PACF values fall within the intervals, with only the values at $h = 19$ and $h = 21$ being slightly lower than the lower bound.

93% of observations fall within the confidence intervals, indicating a good fit of the model to the data.

5.3 Comparison of quantile lines with trajectory

We will also estimate quantile lines for the trajectory. We will generate 2000 trajectories of the ARMA(p,q) series with the estimated parameters, each of length n equal to the length of the test data. Then, for each $i = 0, 1, ..., n$, we will determine quantiles of order 0.05 and 0.95, as well as 0.25 and 0.75. For the first pair of values, we expect 90% of the data to fall between the determined lines, and for the second pair, it should be 50%, if the model is well chosen.

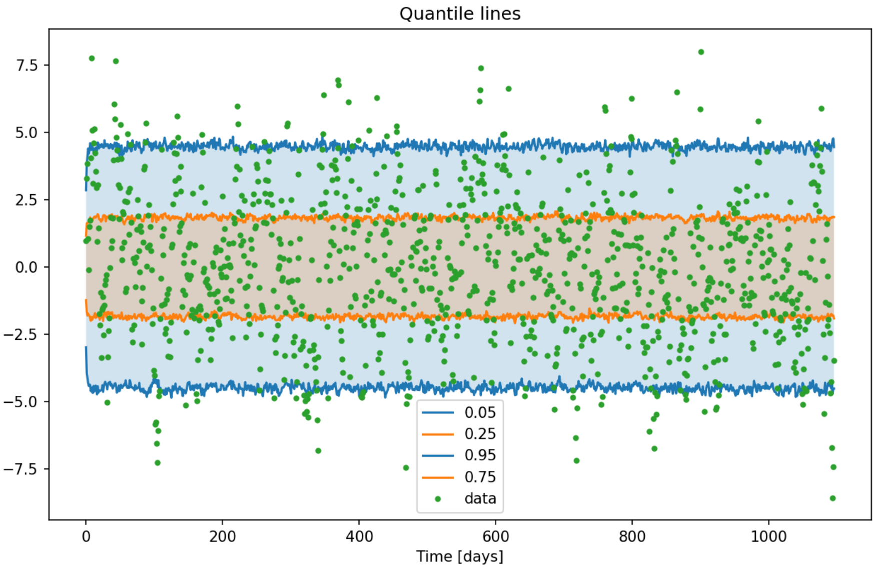

Figure 7: Quantile lines of trajectory.

The plot of the data with overlaid quantile lines is shown in Figure 7. Between the quantile lines of order 0.05 and 0.95, 977 observations fall, which is 89.1% of the total, roughly what was expected. Between the lines of order 0.25 and 0.75, there are 549 observations, or 50.1%, again in line with expectations. We can assume that our data follows the trajectory of the selected ARMA model.

5.4 Forecasting for future observations

To forecast future observations, we will use confidence intervals of the form:

$$[\hat{X}_{t+h} - q_{1-\alpha/2} \cdot \hat{\sigma}^2,\ \hat{X}_{t+h} + q_{1-\alpha/2} \cdot \hat{\sigma}^2],$$

where $\hat{\sigma}^2$ is the estimated variance of the white noise, $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile from the standard normal distribution N(0, 1) and $\hat{X}_{t+h}$ is calculated for successive h as follows

$$\hat{X}_{t+h} = \phi_1\hat{X}_{t-1+h} + \phi_2\hat{X}_{t-2+h} + \ldots + \phi_2\hat{X}_{t-p+h} + \hat{Z}_{t+h} + \theta_1\hat{Z}_{t-1+h} + \theta_2\hat{Z}_{t-2+h} + \ldots + \theta_q\hat{Z}_{t-q+h},$$

where for $t \leq n$ we substitute $\hat{X}_t$ and $\hat{Z}_t$ with X_t and Z_t respectively, for $t > n$ we replace $\hat{Z}_t$ with the expected value of white noise, which is zero. The provided confidence interval is not exact, it is only an approximation of the exact confidence intervals. The larger n, the more accurate the confidence intervals are. Since we have $n = 1096$, we can consider the obtained result to be close to the exact one.

On the plot 8 confidence intervals for future observations are presented along with the actually observed values over this time interval (test data), while on the plot 9, the same data and intervals are depicted but

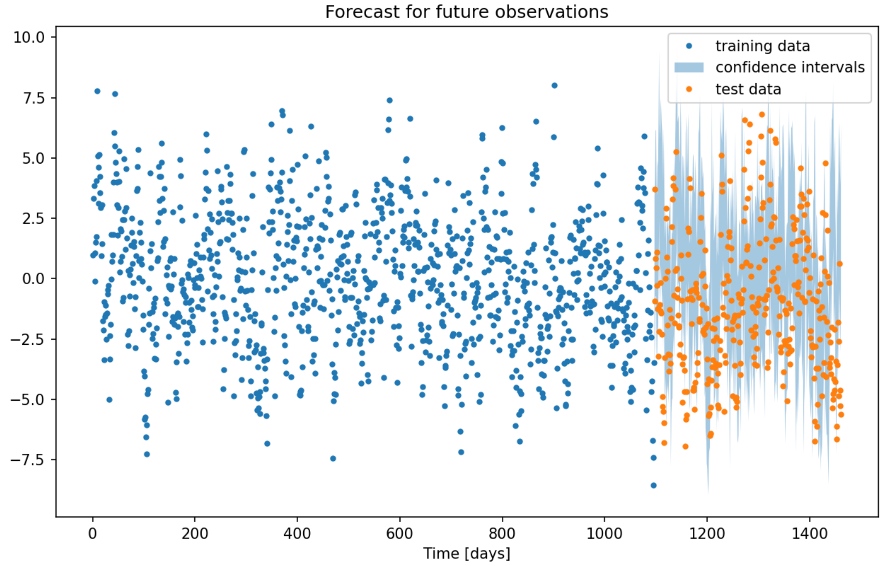

Figure 8: Confidence intervals for future observations compared to values from the test set.

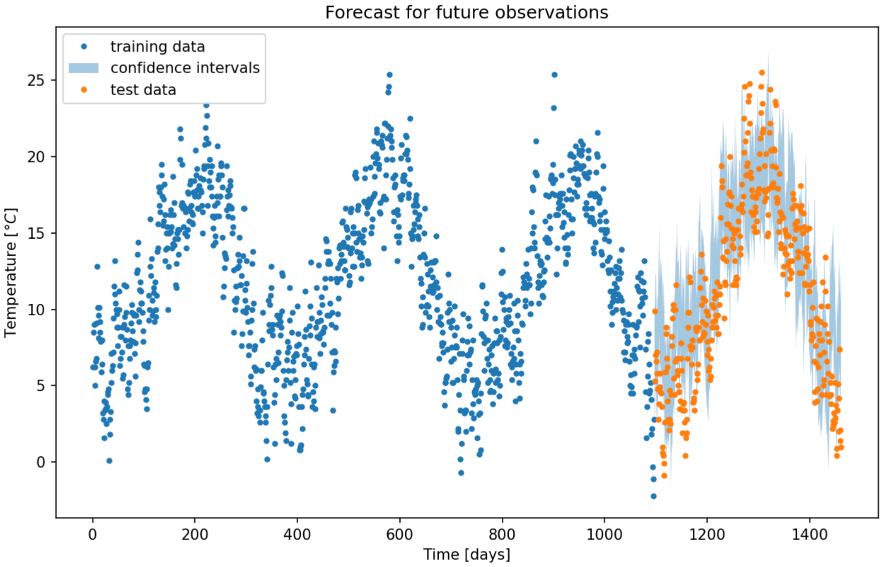

Figure 9: Confidence intervals for future observations compared to values from the test set before decomposition.

in the pre-decomposition form. At first glance, it's evident that many observations indeed fall within the confidence interval, but only around 60% of the observed values lie within them. However, it's worth noting that even those not falling within the confidence intervals are close to them.

6 Verification of Noise Assumptions

We'll now proceed to analyze the residuals of the model, that is the analysis of Z_t. We'll check whether they fulfill our assumptions:

- mean equals 0,
- constant variance,
- uncorrelated residuals,
- the normal distribution.

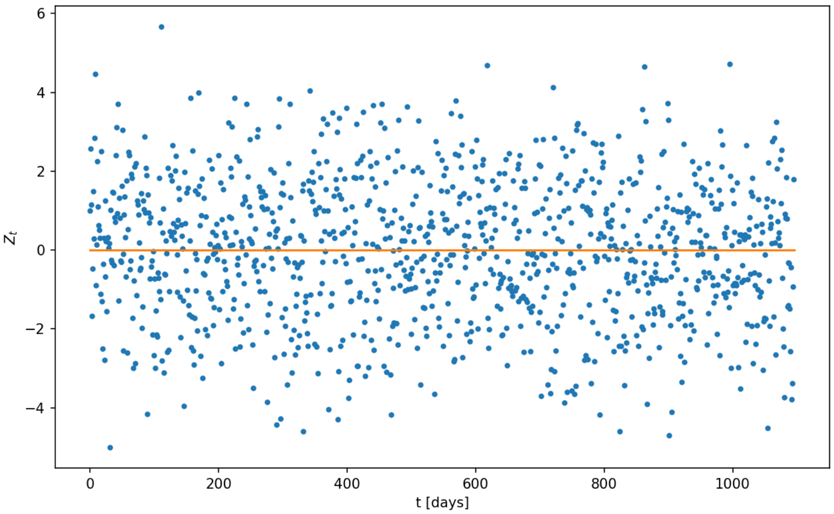

Figure 10: Scatter plot of residual values.

6.1 Mean

The mean value of the residuals is 0.0019. From the plot 10 we observe that the residual values are evenly distributed around 0. We'll further verify this using a Student's t-test at a 95% confidence level. H_0 : sample mean is μ, H_1 : sample mean is different from μ. The test statistic is:

$$T = \frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}},$$

where $\overline{X}$ is the sample mean, $\hat{\sigma}$ is the sample standard deviation, n is the sample size.
We want to check if the mean is 0, so $\mu = 0$. The test results are shown in table 4. The p-value is greater than 0.05, so we have no grounds to reject the null hypothesis.

statistic value	-0.010
p-value	0.992

Table 4: The results of the t-Student test for the mean.

The assumption of a mean equal to 0 is satisfied.

statistic value	0.561
p-value	0.454

Table 5: Results of Levene's test

6.2 Variance

From the scatter plot (figure 10) we can observe that the spread of values around the mean is relatively constant with no significant outliers.
We'll conduct Levene's test for equality of variance at a significance level of 0.05. This test checks the equality of variances in k groups of observations.

$$H_0 = \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$

$$H_1 = \sigma_i^2 \neq \sigma_j^2 \text{ for at least one pair } (i, j)$$

The test statistic is

$$L = \frac{N-k}{k-1} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2} [3]$$

where:
k - number of groups, N_i - number of observations in group i,
N - total number of observations,
Y_{ij} - value of observation j in group i,
$|Y_{i\cdot}|$ - mean of group i,
$Z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$,
$Z_{i\cdot} = \frac{1}{N_i} \sum_{i=1}^{N_i} Z_{ij}$,
$Z_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$.

We'll divide the data into two groups and check the equality of variances in those groups. The test results are shown in Table 5. The p-value is greater than 0.05, so we have no grounds to reject the null hypothesis. The variance in these groups is the same.

Based on this, we can conclude that the variance is constant, which meets the second assumption.

6.3 Uncorrelated and Independent

To check the independence of residual values, we'll use ACF and PACF plots as well as the Ljung-Box test.

From the plots in Figures 11 and 12, we observe that for all $h > 0$, these functions have values not significantly different from zero, indicating that they are uncorrelated.

We'll also conduct the Ljung-Box test. H_0: the data are uncorrelated, H_1: the data are not uncorrelated. The test statistic is

$$Q = n(n+2) \sum_{k=1}^{h} \frac{\hat{\rho}(k)}{n-k},$$

where n is the number of observations, $\hat{\rho}(k)$ is the empirical correlation for lag k.

From table 6, it can be seen that for all lags, the p-value is greater than 0.05. Therefore, we reject the alternative hypothesis, thus concluding that the residuals are uncorrelated.

Based on the gathered information, we conclude that the residuals are uncorrelated and may be independent.

6.4 Normality of Distribution

We will check if the residual values follow a normal distribution. In figures 13, 14 and 15 we compared the distribution function of $N(0, \hat{\sigma}^2)$ with the empirical counterpart, similarly with density and quantile plots. For each plot, the empirical value aligns closely with the theoretical one.

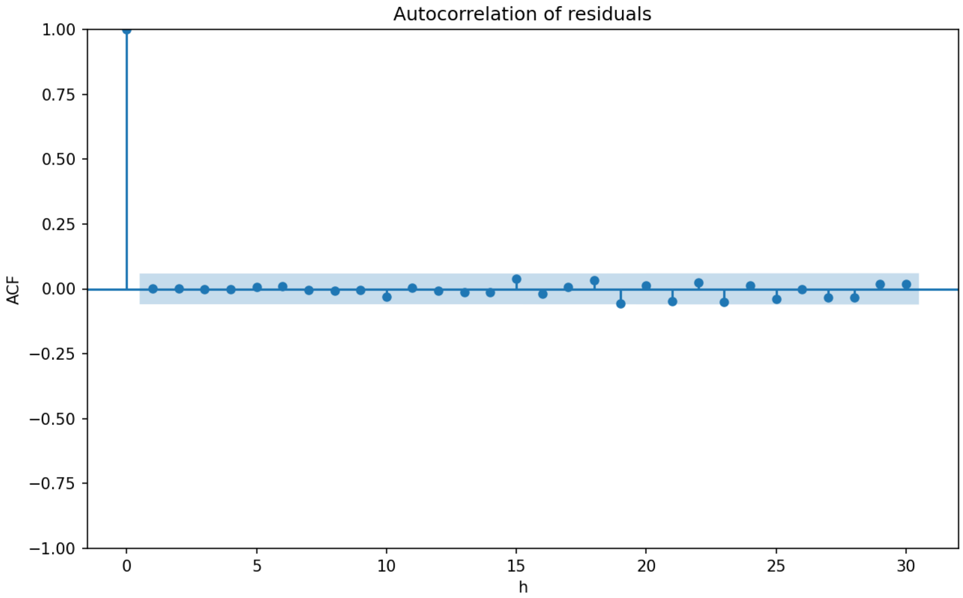

Figure 11: ACF plot for residual values.

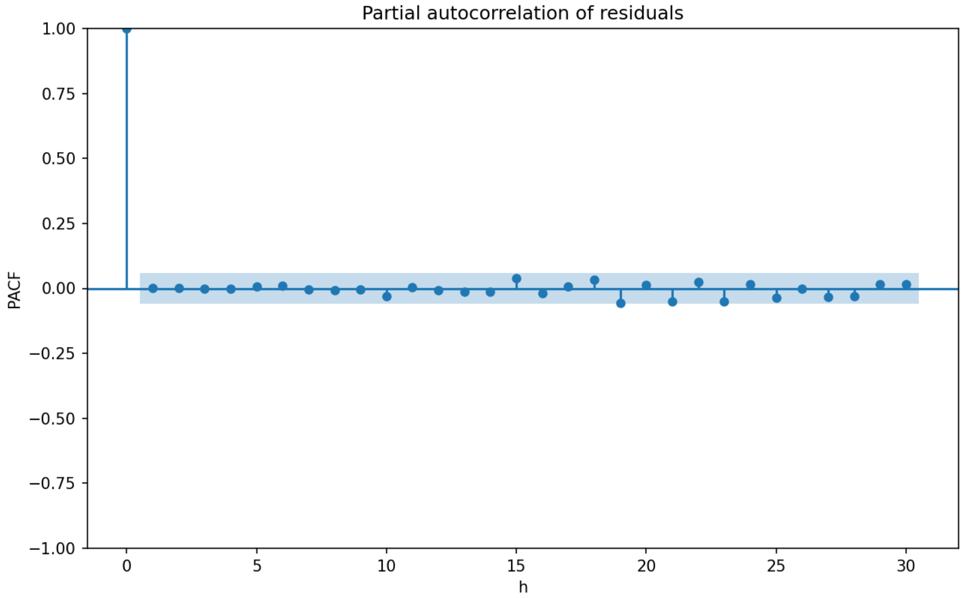

Figure 12: PACF plot for residual values.

We will further verify the normality of the distribution by conducting the Shapiro-Wilk test. H_0 : the sample distribution is normal. H_1 : the sample distribution is not normal. The test statistic is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, [2]$$

Results of the Ljung-Box test		
h	statistic value	p-value
1	0.002514	0.960015
2	0.003961	0.998022
3	0.006745	0.999853
4	0.008661	0.999991
5	0.068813	0.999936
6	0.192002	0.999863
7	0.210112	0.999970
8	0.286599	0.999984
9	0.311311	0.999996
10	1.380784	0.999261

Table 6: Results of the Ljung-Box test.

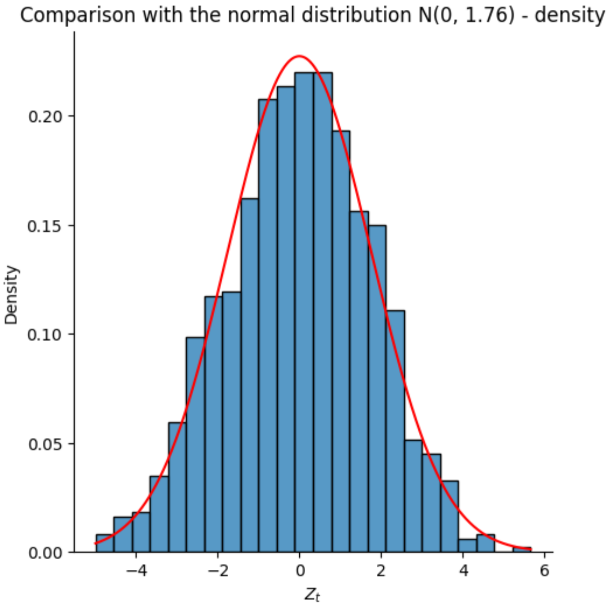

Figure 13: Comparison of normal distribution density with residuals.

where $x_{(i)}$ are the ordered values of x_i, a_i are constants generated from the means, variances, and covariances of ordered samples of length n from a normal distribution, n is the number of observations.

Results of the Shapiro-Wilk test	
statistic value	0.998
p-value	0.237

Table 7: Results of the Shapiro-Wilk test.

From table 7, it is evident that the p-value is greater than 0.05, thus we lack evidence to claim that the residuals do not follow a normal distribution. It can also be noted that the test statistic is close to 1, further confirming the normality of the residuals. Based on this information, we can conclude that the assumption of normality in the distribution of residuals is satisfied.

Since $\{Z_t\}$ are uncorrelated and follow a normal distribution, it is very likely that they are independent. All assumptions are met, and thus we can conclude that the selected ARMA model is appropriate and accurately describes the data.

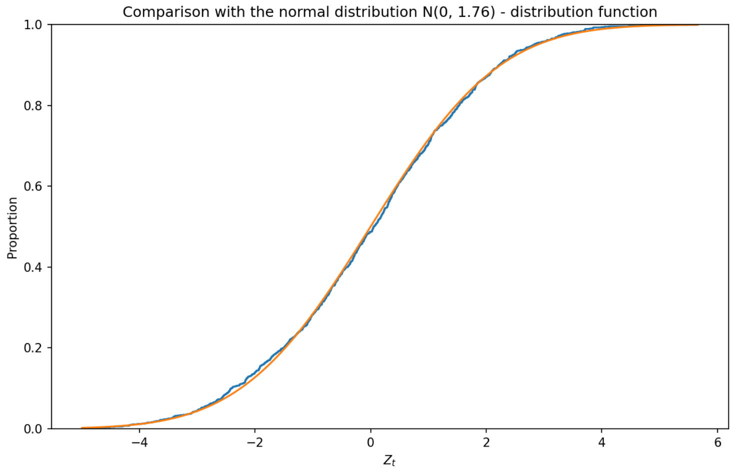

Figure 14: Comparison of the cumulative distribution function of the normal distribution with the residuals.

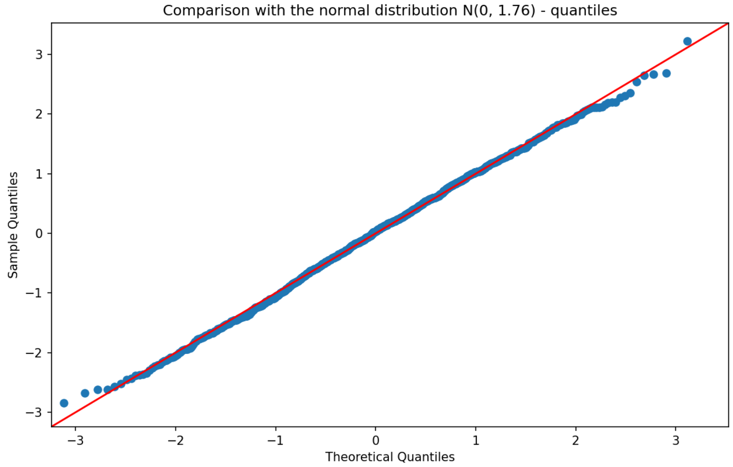

Figure 15: Q-Q plot of normal distribution and residuals distribution.

7 Conclusions

The report analyzed data concerning the average temperature in London from 1998 to 2001. The data underwent a Wald decomposition, and an ARMA(3, 6) model was fitted to it. The obtained model proved to be a good fit for the data and satisfied all assumptions regarding the residual values. Additionally, a forecast for future observations was conducted. Although only approximately 60% of the observations fell within the confidence intervals, the values outside did not deviate significantly.

References

[1] https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

[2] https://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm

[3] https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm