AI6123 TIME SERIES ANALYSIS

INDIVIDUAL PROJECT 1

School of Computer Science and Engineering

Programme: MSAI

Date: 15 March 2024

Authored By: Tan Jie Heng Alfred (G2304193L)

Contents

1. Introduction

In this project, we analyze the time series consisting of the number of users connected to the internet through a server. Each of the data is collected at an interval of one minute and there are a total of 100 observations made. We will attempt to fit an autoregressive integrated moving average (ARIMA) model on this time series and use the model for future predictions. To do this, we will use the open-source programming language Python, together with some other libraries. In particular, we use `pandas` to handle the time series dataset in terms of dataframe; `statsmodels` for various statistical models and tests; and `matplotlib` to visualize various plots.

2. Preliminary Dataset Analysis

Before fitting an ARIMA model, we first have to determine whether the time series is weakly stationary. A time series $\{X_t\}_{t=0}^{N-1}$ is said to be weakly stationary if $\mathbb{E}(X_t) = \mu_t$ is independent of timestep t (i.e., $\mu_t = \mu$ for some μ) and $\gamma(t, k) = \operatorname{cov}(X_t, X_{t+k})$ is independent of time step t for each time lag k. For brevity, we will refer to weakly stationary time series as stationary time series.

To have an intuitive understanding of whether our time series is stationary, we will plot the observations, x_t, at time step t (see Figure 1).

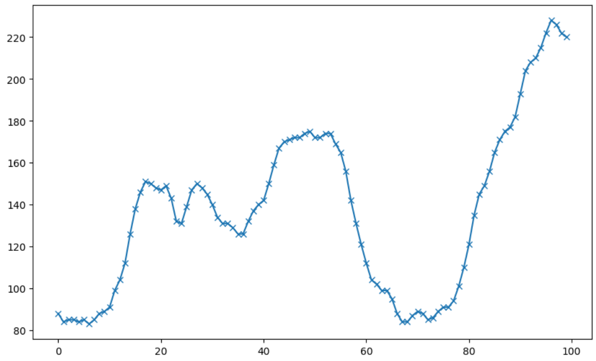


Figure 1: Visualization of the observations $\{x_t\}_{t=1}^{100}$

We can see that for timesteps $0 \leq t \leq 10$, the mean of the observations is roughly 80, whereas for timesteps $20 \leq t \leq 30$, the mean of the observations is roughly 140. Since the sample mean is an unbiased estimate of the expectation, this means that $\mathbb{E}(X_t) \neq \mathbb{E}(X_i)$, for $t \in \{0, 1, \dots 10\}$ and $i \in \{20, 21\,, \dots ,30\}$, suggesting that $\{X_t\}_{t=0}^{99}$ is not a stationary time series. Indeed, there seems to be a 'trend' component in this time series, leading to it not being a stationary time series.

To be more rigorous, we analyzed the autocorrelation function (ACF) and partial autocorrelation function (PACF) of $\{X_t\}$. The autocorrelation function ρ_k of $\{X_t\}$ at time lag $k \in \mathbb{Z}$ is given to be,

$$\rho_k = \frac{\gamma(t, k)}{\sqrt{\gamma(t, t)\gamma(k, k)}} = \frac{\operatorname{cov}(X_t, X_{t+k})}{\sqrt{\operatorname{var}(X_t)\operatorname{var}(X_{t+k})}}$$

This is equivalent to $\rho_k = \frac{\gamma_k}{\gamma_0}$.

We could use the ACF to determine whether $\{X_t\}$ is stationary. Specifically, if $\{X_t\}$ is non-stationary, then its ACF is likely to decay very slowly. If $\{X_t\}$ is determined to be stationary, then a further analysis of its PACF is required to determine the ARIMA model that would likely fit $\{X_t\}$.

The partial autocorrelation function ϕ_{kk} of $\{X_t\}$ at time lag $k \in \mathbb{Z}$, $k \geq 2$, is given to be,

$$\phi_{kk} = \operatorname{corr}(X_k - f_{k-1}, X_0 - f_{k-1}), \text{where}$$

$$\operatorname{corr}(X_i, X_0) = \rho_i \text{ and } f_{k-1} = f(X_{k-1}, X_{k-2}, \dots, X_1).$$

Here, f_{k-1} minimizes the mean square linear prediction error, i.e., $f_{k-1} = \underset{F_{k-1}}{\operatorname{argmin}}\, \mathbb{E}(X_k - F_{k-1})^2$.

Now, we will investigate whether the ACF of $\{X_t\}$ decays slowly. However, because we only have finite samples of observations, we will use the sample ACF (SACF) and sample PACF (SPACF) to estimate, up to a level of confidence (i.e., 5% level of significance), the ACF and PACF of $\{X_t\}$ respectively. Therefore, if the SACF at lag k is within a certain bound $(-\alpha_k, \alpha_k)$, then there is a 0.95 probability that the ACF of the time series at lag k is negligible (i.e., ACF ≈ 0). To visualize the SACF and SPACF, we use functions `plot_acf` and `plot_pacf` in `statsmodels.graphics.tsaplots` (see Figure 2). These functions automatically calculate the bound $(-\alpha_k, \alpha_k)$.

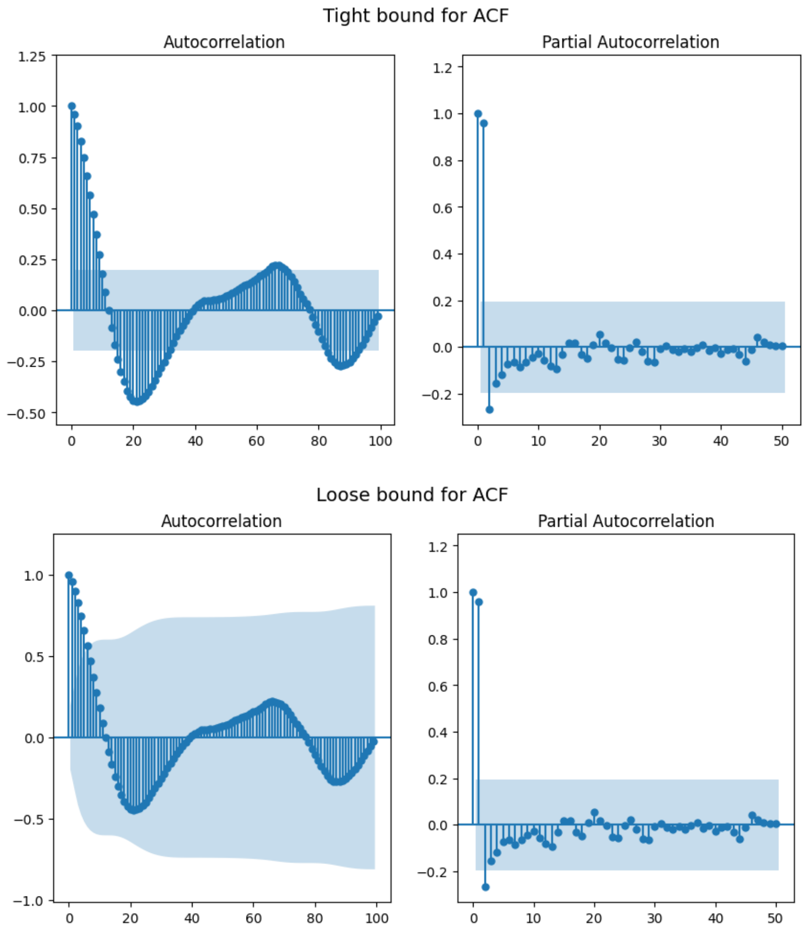


Figure 2: SACF and SPACF plots of $\{x_t\}$, with tight bounds (top) and loose bounds (bottom) of $(-\alpha_k, \alpha_k)$.

Note that the bound of SACF is a function of the time lag k. We included the tighter bound, at the constant $\frac{1.96}{\sqrt{N}} = \frac{1.96}{\sqrt{100}} = 0.196$, and the loose bound which widens with k. Disregarding the bound, we see that the SACF is fluctuating and decaying at a slow rate. Indeed, using the tight bound, we can see that the SACF is significant different from 0 even after lag 80, suggesting that the ACF is decaying slowly. However, if we use the loose bound, the SACF is negligible after time lag 6. Therefore, if we adopt the tight bound, we would conclude that the time series $\{X_t\}$ is not stationary, whereas if we adopt the loose bound, we would conclude that the time series is stationary since the ACF is statistically negligible after lag 6. However, the SACF is still fluctuating a lot, and the plot of the time series provide additional evidence that the time series is indeed non-stationary.

3. Obtaining Stationarity by Differencing

Assuming $\{X_t\}$ is non-stationary, we cannot fit an ARIMA model on it. Fortunately, we can deal with this by applying the differencing operation, ∇. In particular, we have $\nabla X_t = X_{t+1} - X_t$. Ideally, this would remove the 'trend' component, assuming the 'trend' component is a polynomial in terms of t. Even if it is not a polynomial, if its Taylor series exists (that is, the 'trend' component is a continuous and bounded function) and valid for a sufficient large range of t, then this approximation can be made arbitrarily close. Therefore, after finitely many differencing, say d times, $\{\nabla^d X_t\}_{t=0}^{N-d-1}$ would be free of the ‘trend’ component. Since there is no seasonality component, as we considered only 100 minutes of observations, this means that $\{\nabla^d X_t\}_{t=0}^{N-d-1}$ is a stationary time series. We will note that, in our time series, the values we observed have an initial upward trend, followed by a downward trend, before having a larger upward trend. This suggests that the ‘trend’ component is likely not linear (i.e., of the form $\beta_0 + \beta_1 t$), hence multiple differencing may be required.

To check whether $\{\nabla^d X_t\}_{t=0}^{N-d-1}$ is stationary, for some $d \in \mathbb{Z}^+$, we employ the same method as in section 2: For some $d \in \mathbb{Z}^+$, we take the steps:

1. Visualize the plot $\{\nabla^d x_t\}_{t=0}^{N-d-1}$, and observe that $\mathbb{E}(\nabla^d x_t) = \mu$ is independent of time step t and $\gamma(t,0) = \text{cov}(\nabla^d x_t, \nabla^d x_t) = \text{Var}(\nabla x_t)$ is independent of time step t. This is a non-rigorous, inconclusive visual inspection. Hence, we chose to observe the easier-to-see $\gamma(t,0)$ instead of $\gamma(t,k)$, for $k > 0$.
2. Plot the SACF of $\{\nabla^d X_t\}_{t=0}^{N-d-1}$, including the bounds for which we decide with 95% probability that the ACF is negligible. We include both the loose bound and tight bound.
3. Plot the SPACF of $\{\nabla^d X_t\}_{t=0}^{N-d-1}$ to determine the ARIMA model if $\{\nabla^d X_t\}$ is stationary.

3.1 Differencing operation: $d = 1$

For $d = 1$, we obtain the following plot of $\{\nabla x_t\}_{t=0}^{98}$ (Figure 3).

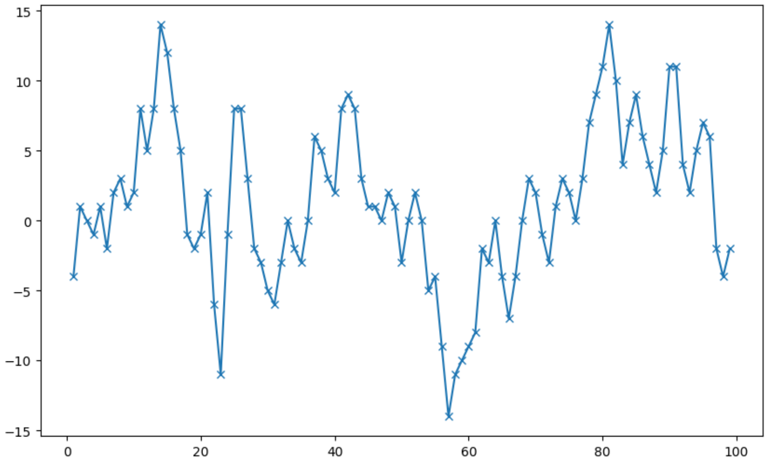


Figure 3: Visualization of the observations $\{\nabla x_t\}_{t=0}^{98}$

Although less severe than the observations $\{x_t\}$, the mean still seems to be dependent on the time period. For instance, for $0 \leq t \leq 5$, the mean is roughly 0, whereas for $10 \leq t \leq 15$, the mean is closer to 10. This suggests that $\{\nabla X_t\}$ is non-stationary. For a more conclusive test, we plot the SACF and SPACF, shown in Figure 4 below. Also, there seems to be an upward trend for $\{\nabla x_t\}_{t=0}^{98}$ between $60 \leq t \leq 80$, which explains the fluctuating trend seen in $\{x_t\}_{t=0}^{99}$.

The magnitude of SACF for $\{\nabla x_t\}$ is smaller than that observed in $\{x_t\}$. However, the SACF is still decaying really slowly. In fact, looking at the plot with a tight bound, the SACF of $\{\nabla x_t\}$ at lag $k = 45$, the SACF is very close to -0.196. This means, with a slightly higher level of significance, the SACF at lag $k = 45$ would result in a rejection of the hypothesis that ACF of $\{\nabla X_t\}$ is negligible. Even though the plot corroborates that the SACF cuts off after lag 26, the fluctuation in SACF and the ‘touching’ of the tight boundary would be strong enough reasons to support the assumption that SACF is decaying slowly instead of cutting off. A similar issue of ‘touching’ the loose boundary can be seen with the plot using the loose bound, but this time it happens at $k = 19$. Hence, together with the non-stationary-looking plot of $\{\nabla x_t\}_{t=0}^{98}$ seen in Figure 3, we will assume that $\{\nabla x_t\}_{t=0}^{98}$ is non-stationary.

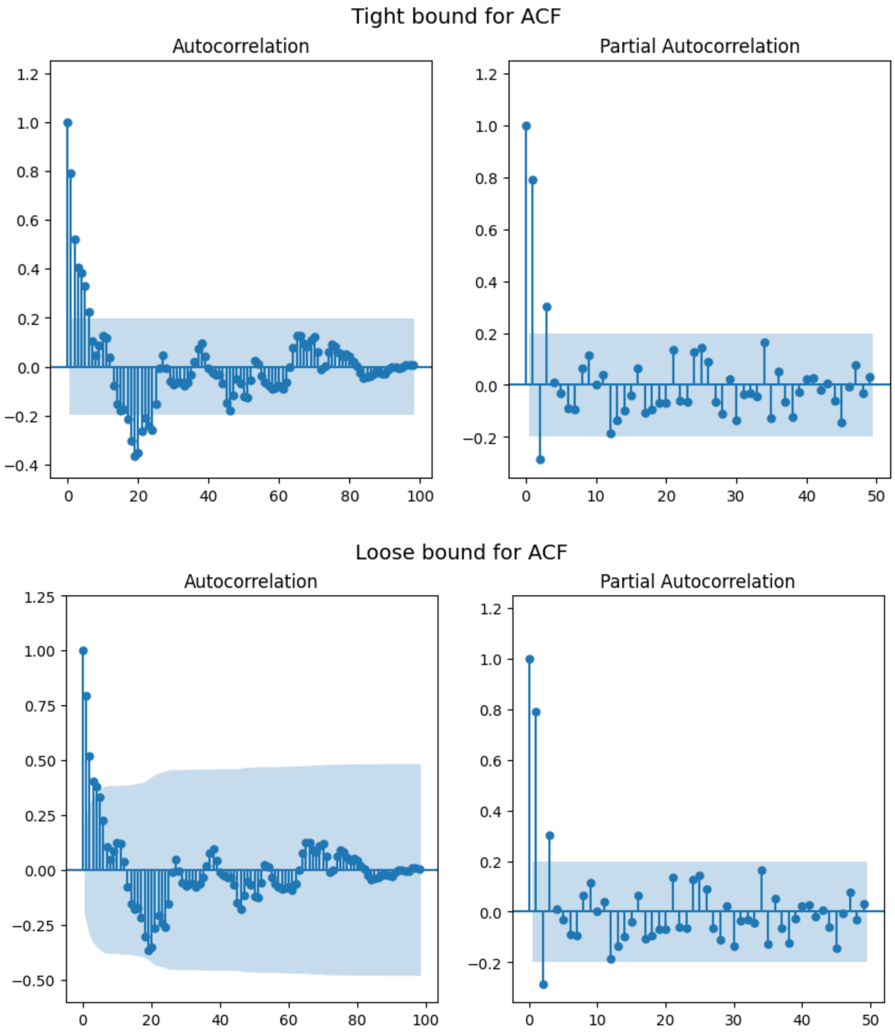


Figure 4: SACF and SPACF plots of $\{\nabla x_t\}$, with tight bounds (top) and loose bounds (bottom).

3.2Differencing operation: $d = 2$

For $d = 2$, we get the following plot of $\{\nabla^2 x_t\}_{t=0}^{97}$ (Figure 5).

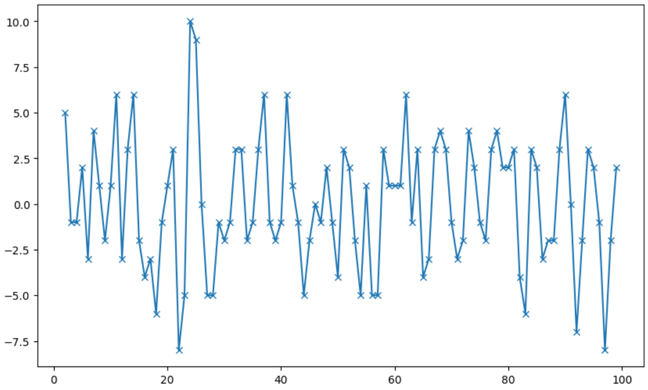


Figure 5: Visualization of the observations $\{\nabla^2 x_t\}_{t=1}^{98}$

The plot now looks more like a stationary time series, where the mean and variance of the observations is roughly constant at any given time period. We move on to look at the SACF and SPACF in Figure 6.

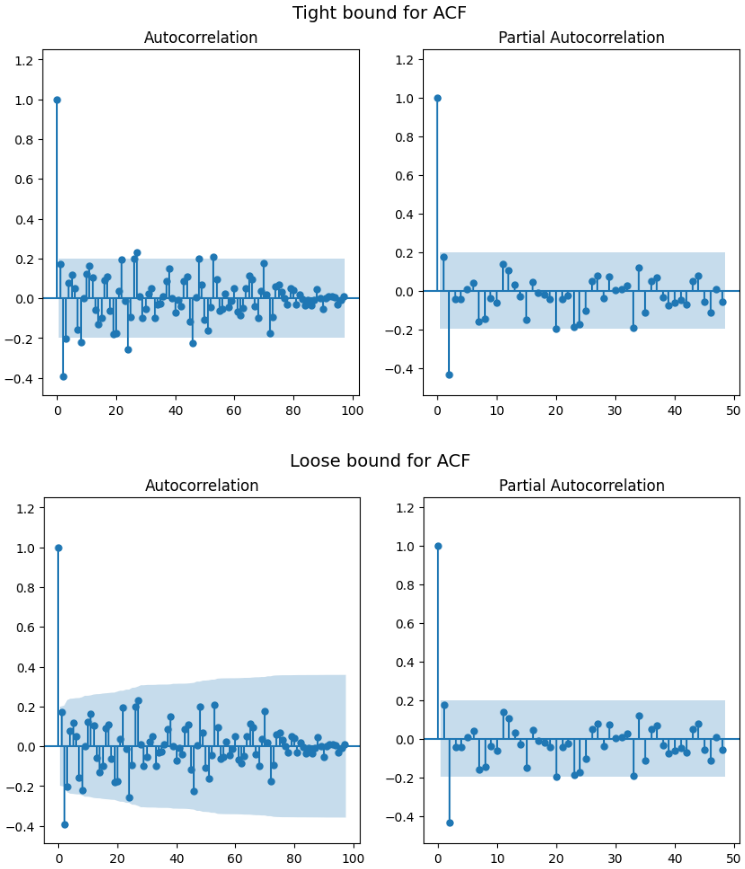


Figure 6: SACF and SPACF plots of $\{\nabla^2 x_t\}$, with tight bounds (top) and loose bounds (bottom).

Looking at the tight bounds, we would conclude that the SACF does not cut off early and decay slowly. However, the extent in which they exceed the boundary is minimal. If we adopt the loose bound, the SACF conclusively stays within the bound after lag 2. Together with the stationary-looking plot of $\{\nabla^2 x_t\}$, we will assume that we have sufficient evidence to conclude that the ACF of $\{\nabla^2 X_t\}$ cuts off after lag 2, while its PACF cuts off after 2, and propose the ARIMA models ARIMA(2, 2, 0), ARIMA(0, 2, 2) and ARIMA(2, 2, 2). However, we will continue to apply differencing once more, to see if the results would be better.

3.3Differencing operation: $d = 3$

For $d = 3$, we get the following plot for $\{\nabla^3 x_t\}_{t=0}^{96}$ (Figure 7).

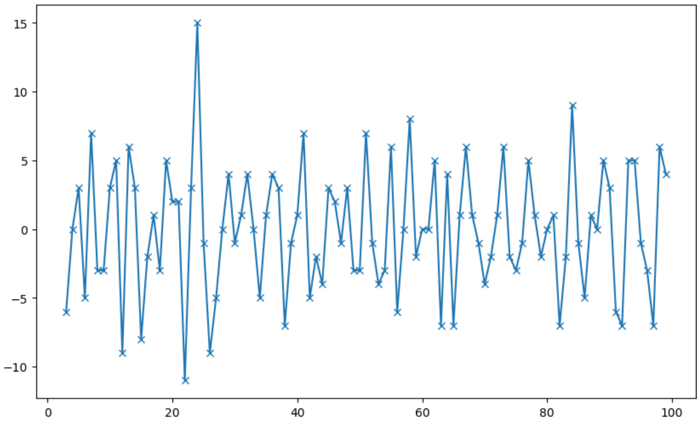


Figure 7: Visualization of the observations $\{\nabla^3 x_t\}_{t=0}^{96}$.

Similarly, the plot of $\{\nabla^3 x_t\}$ suggests that the time series is likely to be stationary, with its mean and variance looking roughly constant. We will continue plotting the SACF and SPACF (Figure 8).

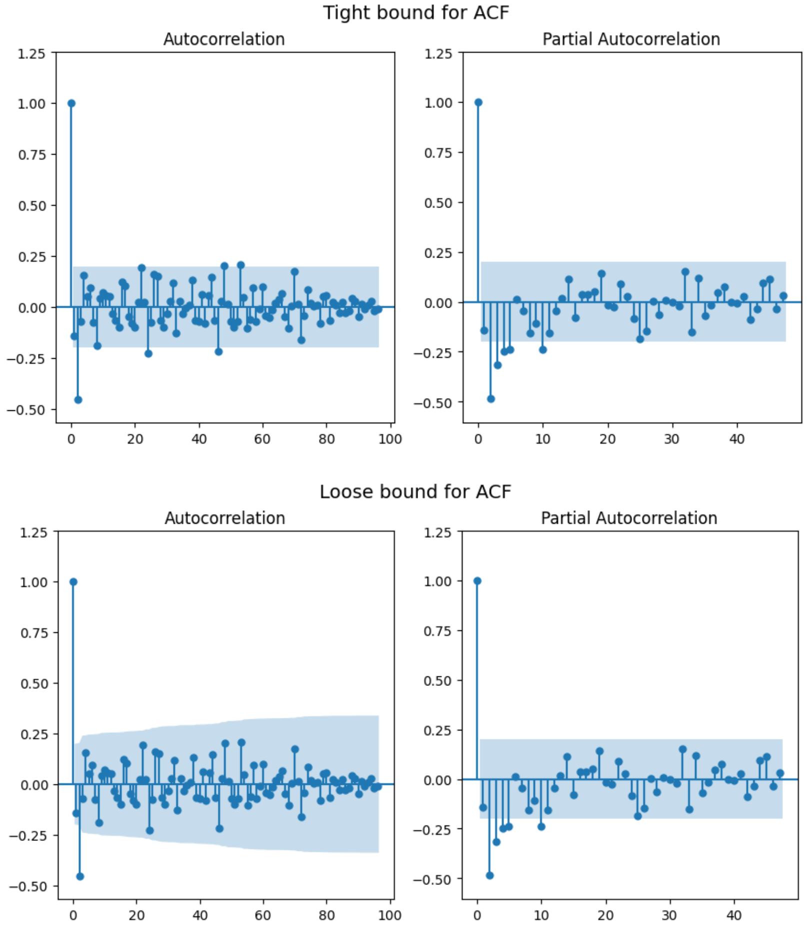


Figure 8: SACF and SPACF plots of $\{\nabla^3 x_t\}$, with tight bounds (top) and loose bounds (bottom).

Generally, the pattern exhibited by the SACF of $\{\nabla^3 x_t\}$ across the time lags seem to be similar to the SACF of $\{\nabla^2 x_t\}$ (Figure 6). Specifically, some of the SACF for larger time lags (e.g., $k \geq 10$) still cross the tight boundary, but not the loose one. However, there is a difference between the two time series' SPACF – the SPACF of $\{\nabla^3 x_t\}$ have more observations with small time lag (e.g., $k \leq 5$) crossing the boundary, compared to $\{\nabla^2 x_t\}$. This could be a result of overperforming the difference operator, which reduces the number of datapoint by d, hence leading to higher variability in the calculation of SPACF. Nonetheless, we will assume that the PACF cuts off after lag 5, treating the SPACF at lag 10 as insignificant (i.e., 5% chance of this happening under significance level of 5%), and assume that ACF cuts off at 2, following the loose bound of ACF. Therefore, we propose the ARIMA models ARIMA(0, 3, 2), ARIMA(5, 3, 0) and ARIMA(5, 3, 2).

3.4Differencing operation: $d = 4$

Following the tight bound of the ACF of $\{\nabla^3 x_t\}$, we cannot conclude that the ACF cuts off at some early lag k. This prompted us to experiment with $d = 4$, even though this makes the model more complicated – we will handle this in the later section. The plot of the observations $\{\nabla^4 x_t\}_{t=0}^{95}$can be seen in Figure 9. Similar to $\{\nabla^2 x_t\}$ and $\{\nabla^3 x_t\}$, the plot suggests that the time series is indeed stationary.

Again, we visualize the SACF and SPACF in Figure 10 below. This time round, the SACF stays within even the tight bound, for lag $k > 2$. This suggests that the 'trend' component of $\{X_t\}$ can be well approximated by a fourth-order polynomial. However, the phenomenon of multiple crossing of SPACF is still present going from $\{\nabla^3 x_t\}$ to $\{\nabla^4 x_t\}$. Nonetheless, we assume that PACF cuts off after lag 5, treating the crossing at lag 32 as insignificant, and assume that ACF cuts off after lag 2. Hence, we could propose the ARIMA models ARIMA(5, 4, 0), ARIMA(0, 4, 2), and ARIMA(5, 4, 2). We will stop differencing from here, since the SACF has died down after some small lag k even for the tight bound.

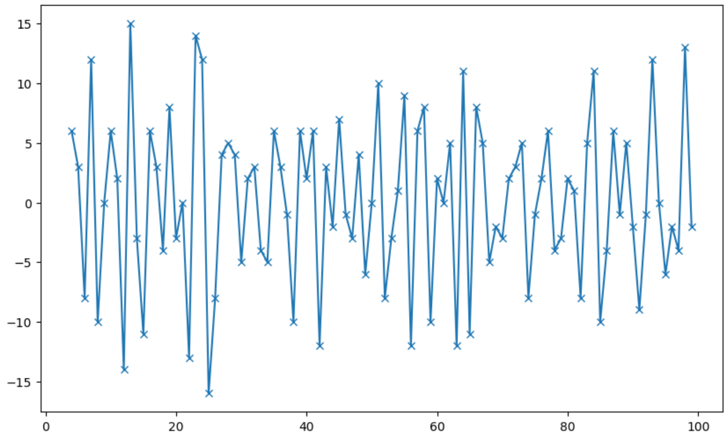


Figure 9: Visualization of the observations $\{\nabla^4 x_t\}_{t=0}^{95}$.

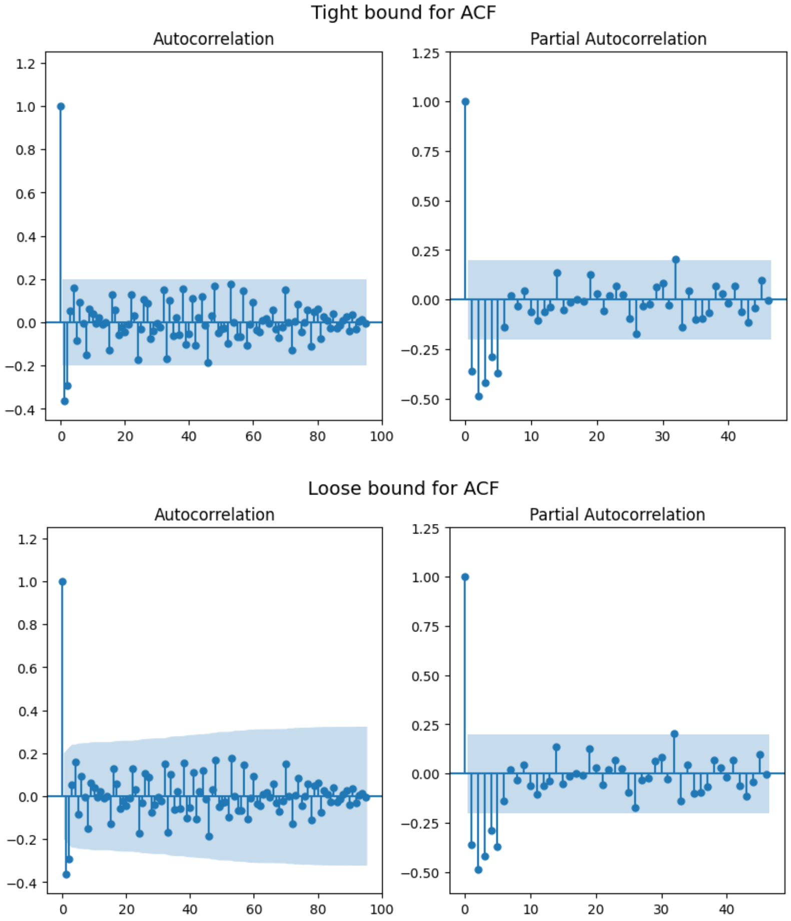


Figure 10: SACF and SPACF plots of $\{\nabla^4 x_t\}$, with tight bounds (top) and loose bounds (bottom).

4. Model Diagnostic

From the previous sections, by looking at the stationarity of the time series, together with their SACF and SPACF plots, we were able to come up with the following ARIMA models:

1. For $d = 2$: ARIMA(2, 2, 0), ARIMA(0, 2, 2), and ARIMA(2, 2, 2)
2. For $d = 3$: ARIMA(5, 3, 0), ARIMA(0, 3, 2), and ARIMA(5, 3, 2)
3. For $d = 4$: ARIMA(5, 4, 0), ARIMA(0, 4, 2), and ARIMA(5, 4, 2)

In general, the ARIMA(p, d, q) model used to fit a time series $\{X_t\}$ assumes that $\{\nabla^d X_t\}$ is stationary and $\{X_t\}$ can be modelled as:

$$\nabla^d X_t - \phi_1 \nabla^d X_{t-1} - \cdots - \phi_p \nabla^d X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

for some $\phi_i \in \mathbb{R}, \phi_p \neq 0$, and $Z_t \sim WN(0, \sigma^2)$ follows the white-noise distribution. The LHS is reminiscent of an autoregressive model (but in terms of $\nabla^d X_i$), while the RHS is similar to the moving average model.

Although the models were proposed from analyzing the SACF and SPACF, they may not be adequate in the sense that they may not fit describe the dependence structure of $\{X_t\}$ adequately. If a model is adequate, then its residuals should follow the white-noise distribution and their autocorrelations should be 0. Therefore, to diagnose our model's adequacy, we can make use of Ljung-Box statistic. The Ljung-Box statistic, $Q(m)$, of an ARIMA(p, d, q) is given to be,

$$Q(m) = N(N+2) \sum_{k=1}^{m} \frac{r_k^2}{N-k},$$

where r_k^2 is the residual autocorrelation at lag k and $0 \ll m \ll N$, typically $m \in \left\{0, \ldots, \frac{N}{5}\right\}$. If our model is adequate, then $Q(m) \sim \chi^2_{m-n_p}$, where n_p is the number of parameters (excluding the mean-offset term) in the ARIMA model. Hence, we will perform a hypothesis test on the null hypothesis that our ARIMA(p, d, q) model is adequate. Therefore, we could calculate and check how likely $Q(m)$ could be sampled from $\chi^2_{m-n_p}$, under 5% level of significance. If the null hypothesis is being rejected (i.e., p-value of $Q(m)$ is less than 0.05 for some m), then the model is not adequate.

To perform the Ljung-Box statistic model test, we will first fit each of the proposed ARIMA model to our time series, then we will check the p-value of its $Q(m)$ for each $m \in \left\{0, \ldots, \frac{N}{5}\right\}$. The former is handled by the function `ARIMA` in `statsmodels.tsa.arima.model`, while the latter is handled by `accor_ljungbox` in `statsmodels.stats.diagnostic`. We plot the p-value of $Q(m)$ for each of the ARIMA models against each $m \in \{0, 1, \ldots 10\}$ in Figures 11, 12, and 13. For each of the plot, there will be a red-dashed line, which indicates the p-value of 0.05.

From Figure 11, we have insufficient evidence at 5% level of significance to reject the hypotheses that ARIMA(2, 2, 0), ARIMA(0, 2, 2) and ARIMA(2, 2, 2) are adequate, the p-values of $Q(m)$, for each m, are above the threshold 0.05. The same can be said for ARIMA(5, 3, 0), ARIMA(5, 3, 2), ARIMA(5, 4, 0), and ARIMA(5, 4, 2), as can be observed in Figure 12 and 13. However, we have sufficient evidence at 5% level of significance to reject the hypotheses that ARIMA(0, 3, 2) and ARIMA(0, 4, 2) are adequate, as some of the p-values of their $Q(m)$ fall below 0.05. Hence, from this diagnostic test, we are left with the following models:

1. For $d = 2$: ARIMA(2, 2, 0), ARIMA(0, 2, 2), and ARIMA(2, 2, 2)
2. For $d = 3$: ARIMA(5, 3, 0) and ARIMA(5, 3, 2)
3. For $d = 4$: ARIMA(5, 4, 0) and ARIMA(5, 4, 2)

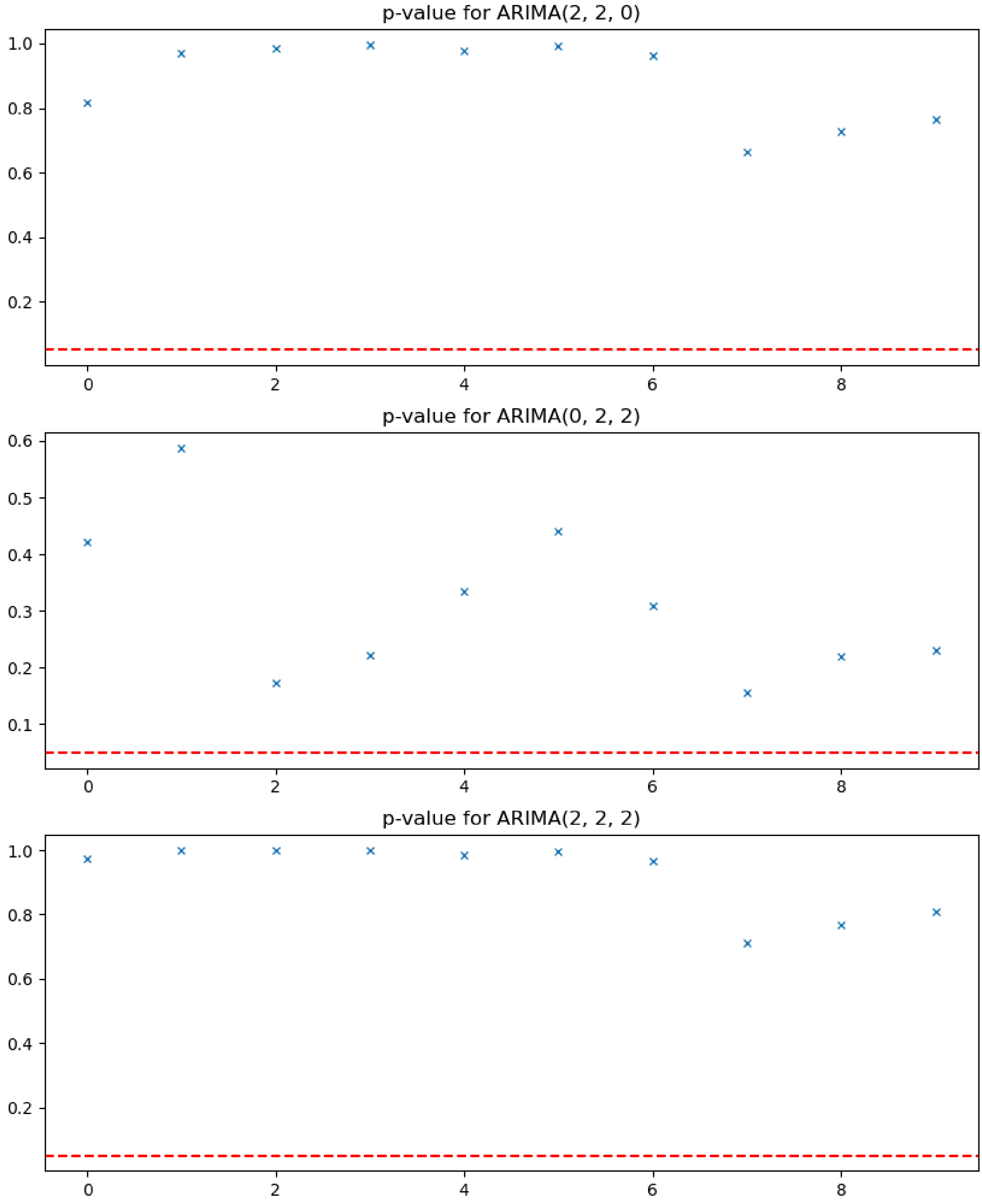


Figure 11: p-value of Ljung-Box statistic for ARIMA models with difference $d = 2$.

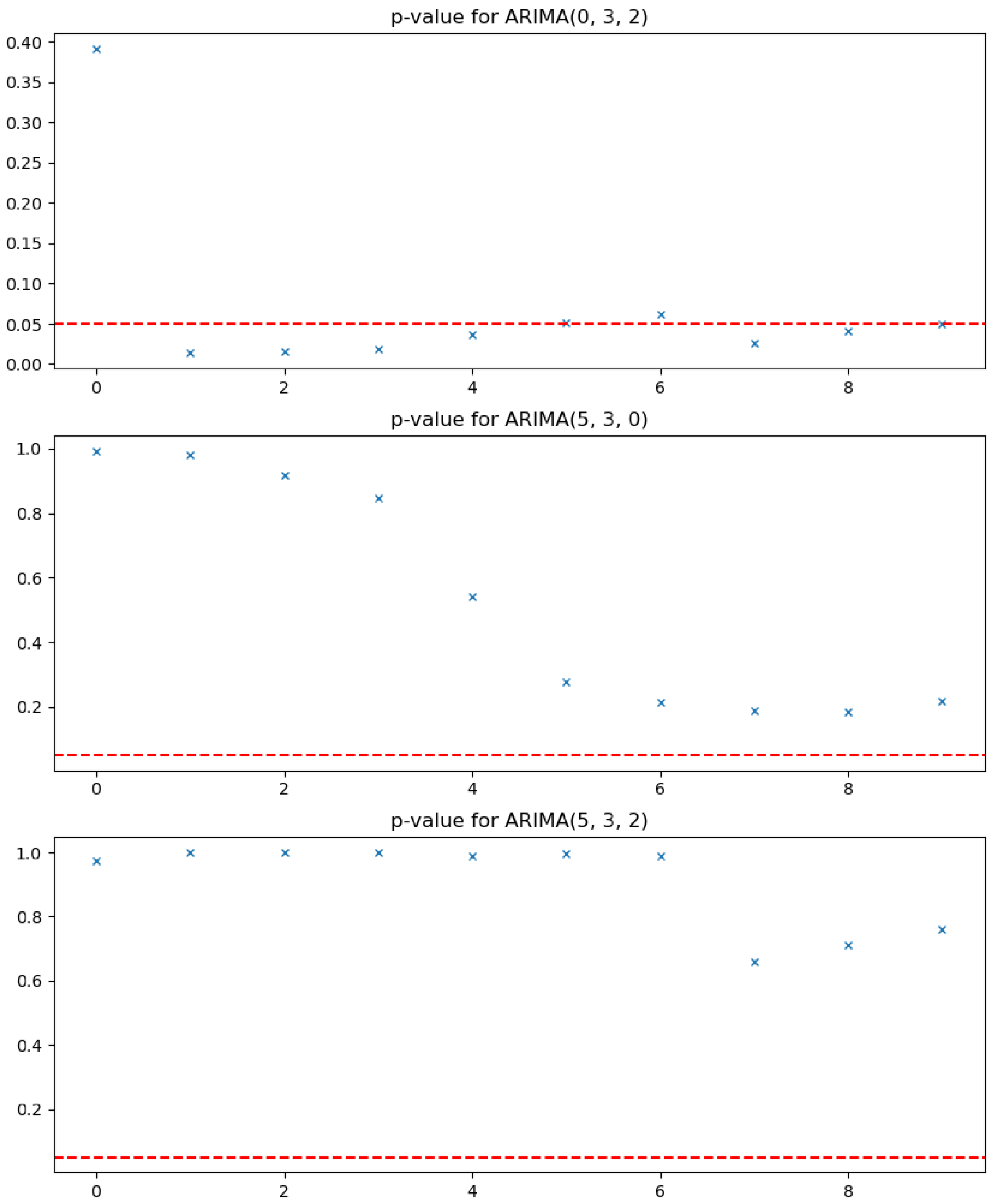
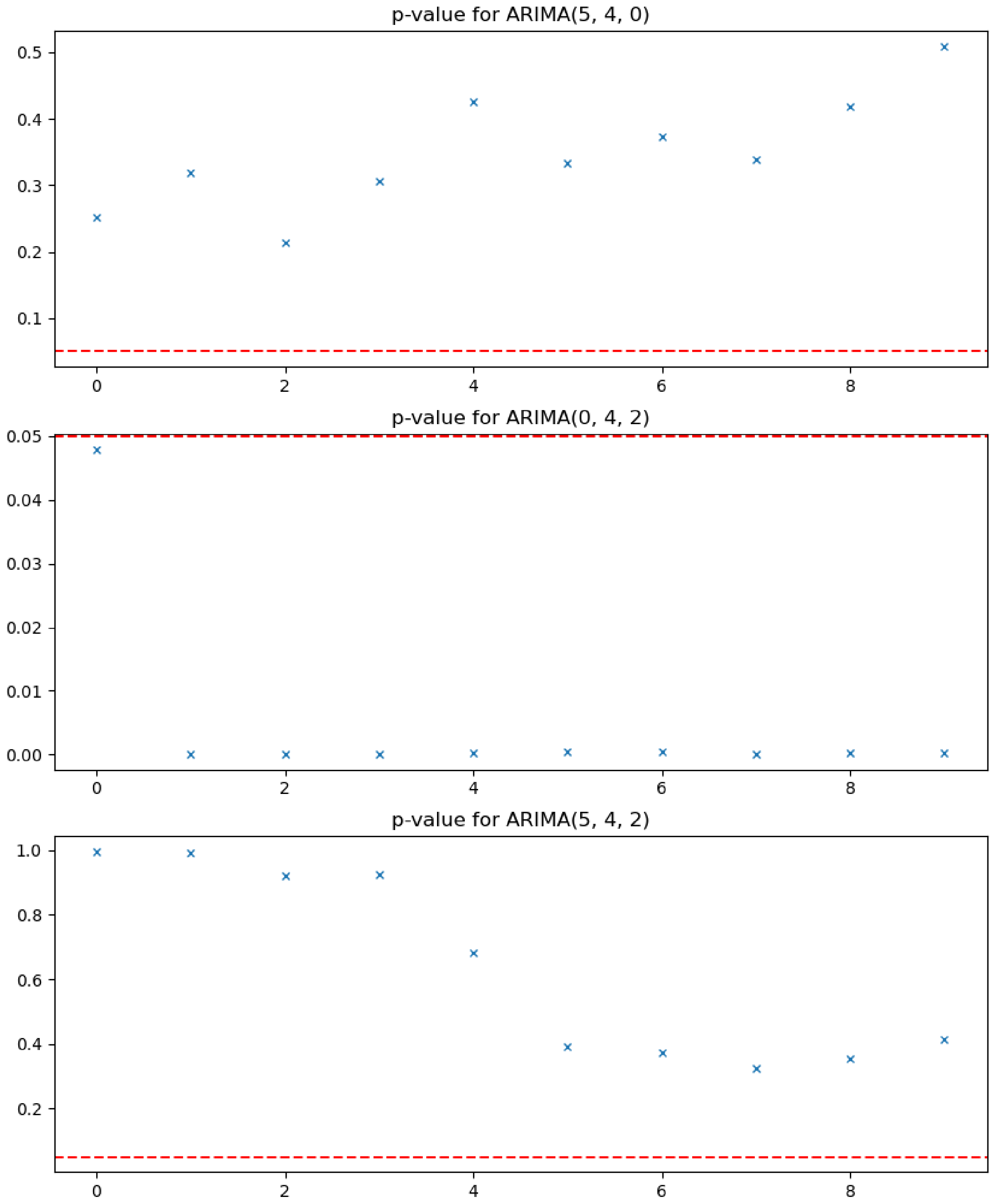


Figure 12: p-value of Ljung-Box statistic for ARIMA models with difference $d = 3$.



p-value for ARIMA(5, 4, 0)

p-value for ARIMA(0, 4, 2)

p-value for ARIMA(5, 4, 2)

Figure 13: p-value of Ljung-Box statistic for ARIMA models with difference $d = 4$.

5. Final Model Selection using Information Criteria

Although the remaining 7 models cannot be rejected under Ljung-Box statistic test, they may not perform equally well, in the sense that the model fits the observations well. On the other hand, we do not want a model that is overly complex that fits the observations too well, as it may not be able to generalize and forecast future predictions. To reach a compromise, we will evaluate our models using the Akaike's Information Criterion (AIC) statistic. The AIC for an ARIMA(p, q, d) model is defined as:

$$\text{AIC} = N \ln \hat{\sigma}^2 + 2(p + q),$$

where $\hat{\sigma}^2 = \frac{1}{N} \sum_{j=1}^{N} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}}$ is the estimated error variance. Here, r_{j-1} is a constant independent of $\hat{\sigma}^2$. Intuitively, with a lower $\hat{\sigma}^2$, we have a better fit of the data as the error variance is lower, all else holds constant. Indeed, when $\hat{\sigma}^2$ is decrease, so will $N \ln \hat{\sigma}^2$, as $\ln(x)$ is a monotonically increasing function. On the other hand, the second summand of $2(p + q)$ acts as a penalty term for a complex model, as p and q determines the order and hence complexity of our ARIMA models. As such, the most ideal ARIMA model would be one that has the least AIC score. In our codes, the AIC score of each of the fitted ARIMA model can be obtained by using the `aic` method on `ARIMA.fit()`. We tabulate the AIC scores (to two decimal places) for each of our remaining ARIMA models in Table 1.

Order of ARIMA model	AIC score
$\mathbf{(2, 2, 0)}$	$\mathbf{511.46}$
$(0, 2, 2)$	517.21
$(2, 2, 2)$	515.13
$(5, 3, 0)$	529.94
$(5, 4, 0)$	562.82
$(5, 3, 2)$	521.67
$(5, 4, 2)$	536.63

Table 1: AIC scores of ARIMA models, ordered in increasing complexity of the ARIMA models.

Notice that as the order of the ARIMA model increases (i.e., increasing $p + q$), the AIC score generally increases, as the models are being penalized more for its complexity. On the other hand, we can see that ARIMA(0, 2, 2) has a slightly higher AIC as compared to ARIMA(2, 2, 2), even though the latter has a higher complexity. This same situation is seen between ARIMA(5, 3, 0) and ARIMA(5, 3, 2). This suggests that the moving average component is a critical aspect of the time series. Besides this, notice that going from ARIMA(5, 3, 0) to ARIMA(5, 4, 0), the AIC has increased, even though the order of the autoregressive component remains the same. This suggests that we may overdone the differencing and a third order differencing is sufficient to model $\{\nabla^3 X_t\}$ as stationary. Finally, ARIMA(2, 2, 0) has the lowest AIC score, suggesting that it is a relatively simple model that could fit the observations sufficiently well. Therefore, we will choose this model for predictions, as it is the most likely model to have accurately modelled the underlying time series.

To visualize our model, we plot out the fitted ARIMA(2, 2, 0) together with the actual time series observations in Figure 14 below. We also include future predictions, for 5 steps, in red-dashed lines. From the figure, we can see that ARIMA(2, 2, 0) seems to follow the actual observations well. Finally, we will compare our model with ARIMA(3, 2, 1), which we found exhaustively by minimizing AIC (see the code snippet in Appendix). The AIC of ARIMA(3, 2, 1) stands at 510.71, only slightly better than our model ARIMA(2, 2, 0). The fitted models seem to coincide in many regions as well (see Figure 15). This therefore suggest that our ARIMA(2, 2, 0) model is decent in modelling the time series.

6. Conclusion

In this project, we analyzed a set of time series observations. From the preliminary analysis, we realized that the time series is non-stationary, which prompted us to perform differencing to eventually obtain a stationary time series. To determine whether the time series is stationary, we made use of its plot as well as its SACF. Once we have determined that the resulting time series is to some degree stationary, we proposed several ARIMA models using its SACF and SPACF plots. Then, with the collection of possible candidates, we determined their adequacy by using Ljung-Box statistic test, before evaluating their performance by using AIC which accounts for the complexity of the model. Through all these, we ended up with the ARIMA(2, 2, 0) model, which seemed to fit the observations very well. Furthermore, it is similar in performance as compared to the best model ARIMA(3, 2, 1), if we evaluate them using AIC. However, we could further analyze whether this model is useful for prediction by performing forecast values analysis using a holdout test set in the future.

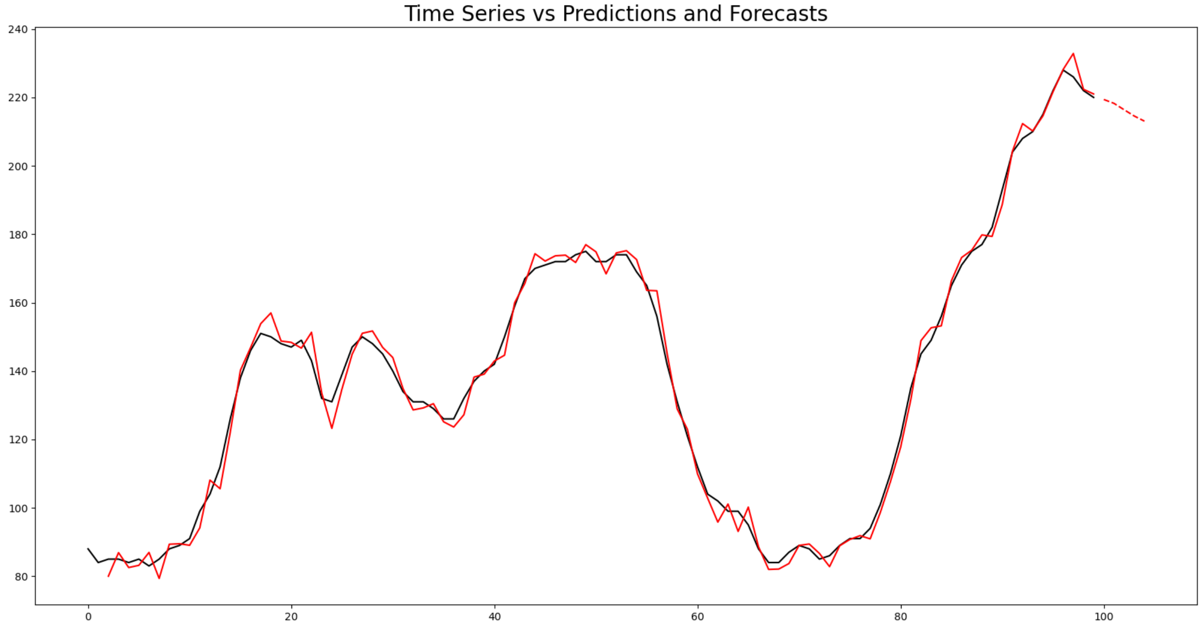


Figure 14: Original observations $\{x_t\}_{t=0}^{99}$ (solid black) with fitted ARIMA(2, 2, 0) model (solid red) and forecasts (dashed red).

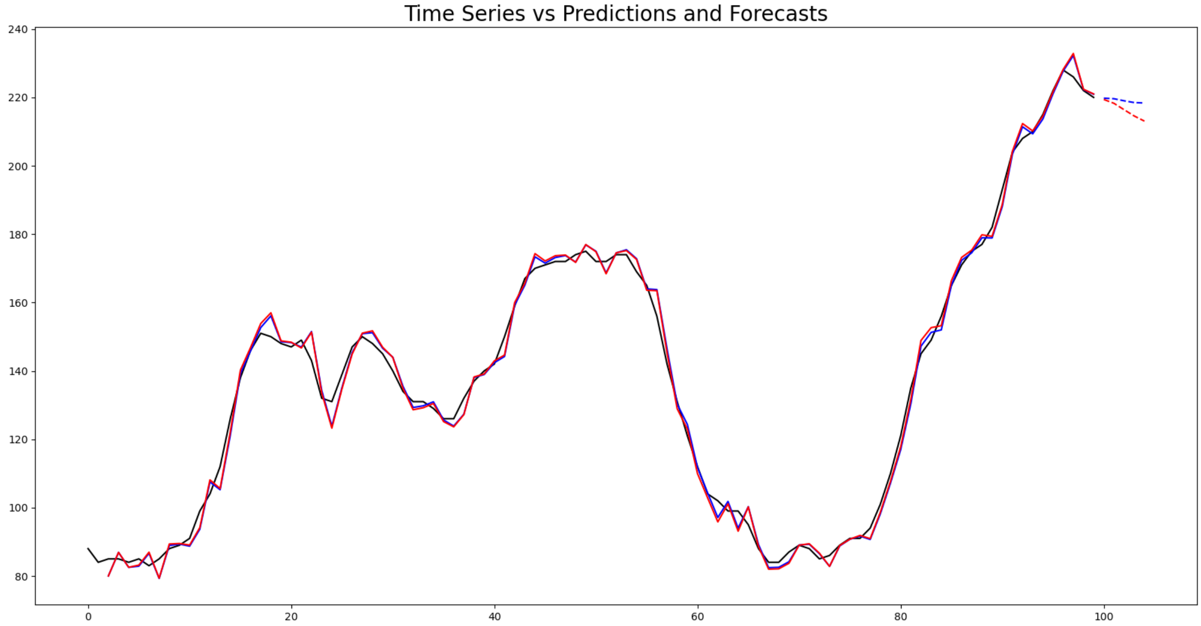
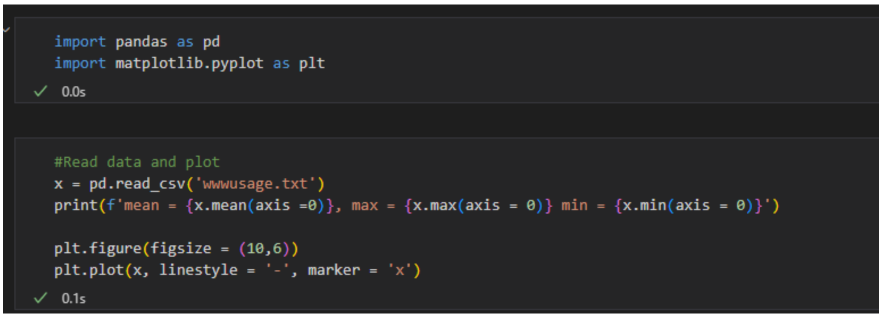


Figure 15: Original observations $\{x_t\}_{t=0}^{99}$ (solid black) with fitted ARIMA(2, 2, 0) model (solid red) and fitted ARIMA(3, 2, 1) model. Their forecasts are in dashed red and blue respectively.

Appendix: Code Snippets

We include important, non-repeating snippets of codes.

1. Read data as a dataframe



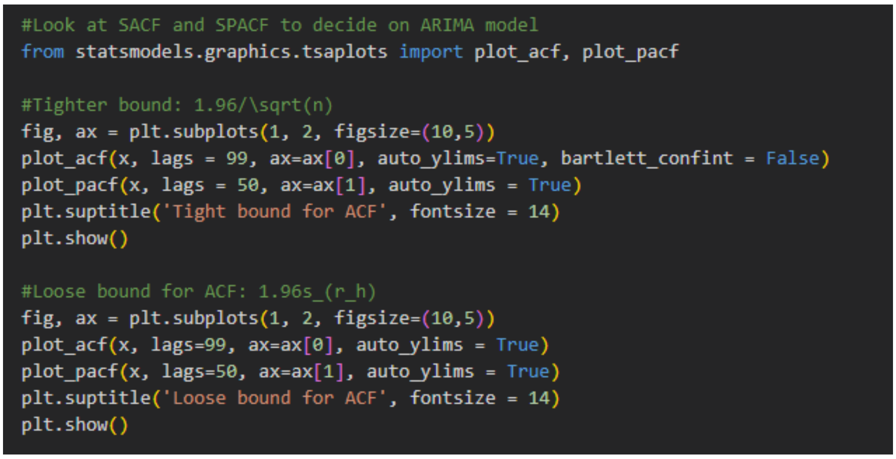
```
import pandas as pd
import matplotlib.pyplot as plt
```

✓ 0.0s

```
#Read data and plot
x = pd.read_csv('wwwusage.txt')
print(f'mean = {x.mean(axis =0)}, max = {x.max(axis = 0)} min = {x.min(axis = 0)}')

plt.figure(figsize = (10,6))
plt.plot(x, linestyle = '-', marker = 'x')
```

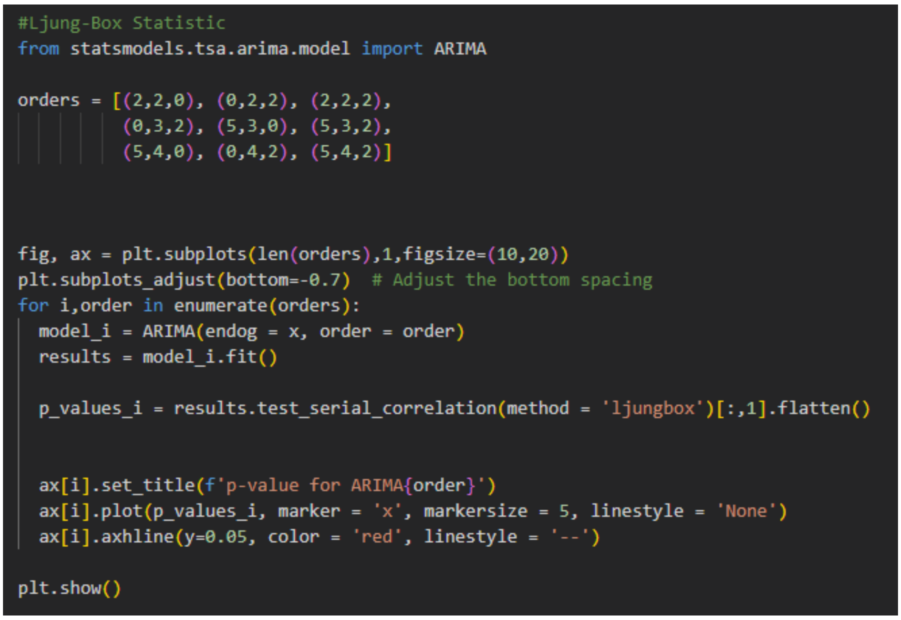
✓ 0.1s

2. Plotting of SACF and SPACF for



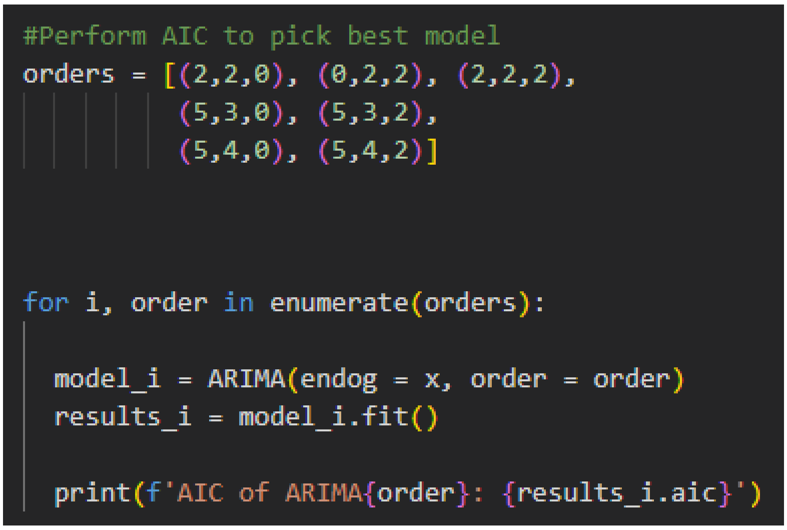
```
#Look at SACF and SPACF to decide on ARIMA model
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

#Tighter bound: 1.96/\sqrt(n)
fig, ax = plt.subplots(1, 2, figsize=(10,5))
plot_acf(x, lags = 99, ax=ax[0], auto_ylims=True, bartlett_confint = False)
plot_pacf(x, lags = 50, ax=ax[1], auto_ylims = True)
plt.suptitle('Tight bound for ACF', fontsize = 14)
plt.show()

#Loose bound for ACF: 1.96s_(r_h)
fig, ax = plt.subplots(1, 2, figsize=(10,5))
plot_acf(x, lags=99, ax=ax[0], auto_ylims = True)
plot_pacf(x, lags=50, ax=ax[1], auto_ylims = True)
plt.suptitle('Loose bound for ACF', fontsize = 14)
plt.show()
```

3. Ljung-Box statistic test for model diagnostic.



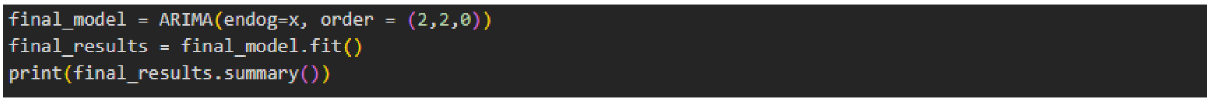
```
#Ljung-Box Statistic
from statsmodels.tsa.arima.model import ARIMA

orders = [(2,2,0), (0,2,2), (2,2,2),
          (0,3,2), (5,3,0), (5,3,2),
          (5,4,0), (0,4,2), (5,4,2)]



fig, ax = plt.subplots(len(orders),1,figsize=(10,20))
plt.subplots_adjust(bottom=-0.7)  # Adjust the bottom spacing
for i,order in enumerate(orders):
  model_i = ARIMA(endog = x, order = order)
  results = model_i.fit()

  p_values_i = results.test_serial_correlation(method = 'ljungbox')[:,1].flatten()


  ax[i].set_title(f'p-value for ARIMA{order}')
  ax[i].plot(p_values_i, marker = 'x', markersize = 5, linestyle = 'None')
  ax[i].axhline(y=0.05, color = 'red', linestyle = '--')

plt.show()
```

4. AIC calculation for all proposed models

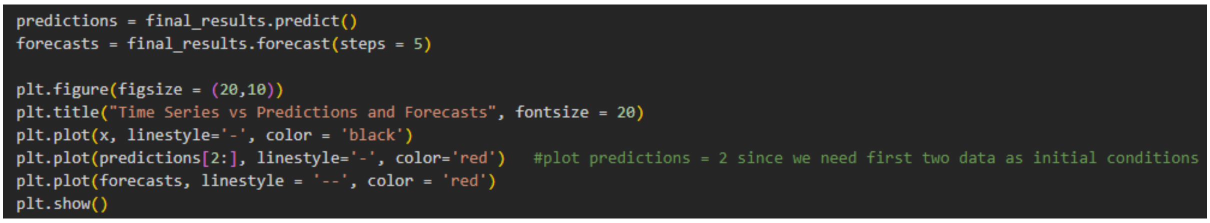


```
#Perform AIC to pick best model
orders = [(2,2,0), (0,2,2), (2,2,2),
          (5,3,0), (5,3,2),
          (5,4,0), (5,4,2)]



for i, order in enumerate(orders):

    model_i = ARIMA(endog = x, order = order)
    results_i = model_i.fit()

    print(f'AIC of ARIMA{order}: {results_i.aic}')
```

5. Final model and predictions.

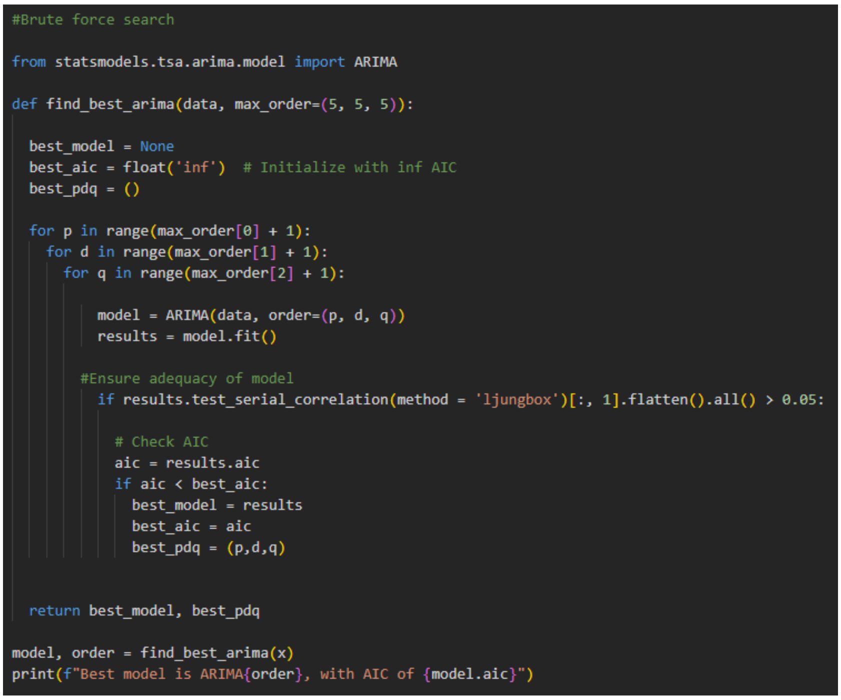


```
final_model = ARIMA(endog=x, order = (2,2,0))
final_results = final_model.fit()
print(final_results.summary())
```



```
predictions = final_results.predict()
forecasts = final_results.forecast(steps = 5)

plt.figure(figsize = (20,10))
plt.title("Time Series vs Predictions and Forecasts", fontsize = 20)
plt.plot(x, linestyle='-', color = 'black')
plt.plot(predictions[2:], linestyle='-', color='red')   #plot predictions = 2 since we need first two data as initial conditions
plt.plot(forecasts, linestyle = '--', color = 'red')
plt.show()
```

6. Exhaustive search for best adequate model that minimizes AIC.



```
#Brute force search

from statsmodels.tsa.arima.model import ARIMA

def find_best_arima(data, max_order=(5, 5, 5)):

  best_model = None
  best_aic = float('inf')  # Initialize with inf AIC
  best_pdq = ()

  for p in range(max_order[0] + 1):
    for d in range(max_order[1] + 1):
      for q in range(max_order[2] + 1):

          model = ARIMA(data, order=(p, d, q))
          results = model.fit()

        #Ensure adequacy of model
          if results.test_serial_correlation(method = 'ljungbox')[:, 1].flatten().all() > 0.05:

            # Check AIC
            aic = results.aic
            if aic < best_aic:
              best_model = results
              best_aic = aic
              best_pdq = (p,d,q)


  return best_model, best_pdq

model, order = find_best_arima(x)
print(f"Best model is ARIMA{order}, with AIC of {model.aic}")
```