0.0.1 SARIMA

The ARIMA model (Auto Regressive Integrated Moving Average) is a widely used statistical model for analyzing and forecasting time series data. It combines autoregressive terms (AR), differencing for stationarity (I), and moving average components (MA).
The Seasonal ARIMA (SARIMA) model extends the ARIMA framework by incorporating seasonal components, thereby enabling the capture of both non-seasonal and seasonal patterns in the data. This model will be employed in our analysis to produce the forecasts. We will begin by examining the key characteristics and preliminary steps of the model, including the assessment of stationarity, the analysis of the autocorrelation function, and other relevant diagnostics, before proceeding to a detailed explanation of the model itself.

Stationary Time Series
Stationarity is introduced to formalize the idea that time series often show consistent patterns or regularity over time. When we talk about stationarity, we refer to weak stationarity, which is sufficient for our model Shumway and Stoffer [2011]. Strict stationarity is a stronger condition that is seldom required in practical applications; therefore, it will not be addressed in this thesis.
Weak stationarity is defined by the fulfillment of three main conditions:

- The mean value μ_t is constant over time: $\mathbb{E}[x_t] = \mu \ \forall t$.
- The variance is finite and does not change with time: $\mathrm{Var}(x_t) = \sigma^2 < \infty \ \forall t$.
- The autocovariance function, $\gamma(s,t) = \mathrm{cov}(x_s, x_t) = \mathbb{E}[(x_s - \mu_s)(x_t - \mu_t)]$, depends only on the time lag $|s-t|$, and not on the specific time s and t points themselves. That is,
$$\gamma(s,t) = \gamma(h) \quad \text{with} \quad h = |s-t|.$$

Since the autocovariance does not depend on the specific time t, but only on the time lag, we can simplify the notation by omitting the second argument and writing it as $\gamma(h)$ instead of $\gamma(h,0)$ Shumway and Stoffer [2011]. Moreover, for a weakly stationary time series where the mean is constant and the autocovariance depends solely on the time lag $h = s - t$, the expression reduces to:

$$\gamma(h) = \mathrm{cov}(x_{t+h}, x_t) = \mathbb{E}[(x_{t+h} - \mu)(x_t - \mu)].$$

Similarly, the general form of the autocorrelation function (ACF) can be simplified in the case of a stationary time series:

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\,\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}.$$

[Shumway and Stoffer, 2011].

Autoregressive model of order p - AR(p)
Autoregressive models are based on the idea that the current value of a time series, x_t,

depends on its own past values. In particular, x_t is estimated using the previous p observations, namely $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$. The parameter p determines how many past time steps the model takes into account to forecast the current value. Consider an autoregressive model of order p, given by:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t,$$

where the time series $\{x_t\}$ is assumed to be stationary, and the coefficients $\phi_1, \phi_2, \ldots, \phi_p$ are constants, with at least one $\phi_p \neq 0$. We assume the error term w_t is a Gaussian white noise process with mean zero and variance σ_w^2, unless specified otherwise. If the series $\{x_t\}$ has a non-zero mean μ, then we can center the series by replacing x_t with $x_t - \mu$ in previous equation, resulting in:

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t,$$

That can alternatively be expressed in the form:

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t,$$

where the intercept α is defined as:

$$\alpha = \mu(1 - \phi_1 - \cdots - \phi_p).$$

The backshift operator formulation provides a compact and flexible way to express autoregressive models. Unlike traditional regression, where predictors are assumed fixed, the regressors in AR models are lagged values of the same stochastic process and hence are random. Using the backshift operator B, we can express the AR(p) model as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) x_t = w_t,$$

which simplifies algebraic manipulation, supports theoretical analysis (e.g., invertibility and stationarity), and seamlessly integrates with broader time series models like ARMA and ARIMA. More concisely as

$$\phi(B) x_t = w_t.$$

And $\phi(B)$ is called the autoregressive operator, defined as:

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

[Shumway and Stoffer, 2011]

Moving Average model of order q - MA(q)

A Moving Average (MA) process creates a linear combination of error terms, which—unless stated otherwise—are assumed to follow a Gaussian white noise process w_t with zero mean and variance σ_w^2. Although w_t is uncorrelated over time, the linear combination introduces autocorrelation in the resulting series x_t. As a result, the final time series can exhibit temporal dependencies and short-term predictable behavior, making it useful for modeling real-world data [Thistleton and Sadigov, 2023c]. Moving average is defined as:

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q},$$

There are q lag terms in the moving average model, and the coefficients $\theta_1, \theta_2, \ldots, \theta_q$ (with $\theta_q \neq 0$) represent the model parameters. Note that the MA model is sometimes written with negative coefficients [Shumway and Stoffer, 2011]. In the context of an MA(q) process, the parameter q defines the number of lag terms included in the model, indicating how far back we look along the white noise sequence when forming the linear combination.
Next, we examine the theoretical properties of the MA(q) process:

- Mean: $E[X_t] = \theta_0 E[w_t] + \cdots + \theta_q E[w_{t-q}] = 0$
- Variance: $V[X_t] = \sigma_w^2 \sum_{i=0}^{q} \theta_i^2$
- Autocovariance: the function is defined as $\gamma(k) = \text{Cov}(X_t, X_{t+k})$. For $k > q$, the autocovariance is zero $\gamma(k) = 0$, reflecting the absence of overlap in the white noise terms. When $k \leq q$, there is overlap, and the autocovariance is given by

$$\gamma(k) = \sigma_w^2 \sum_{i=0}^{q-k} \theta_i \theta_{i+k}.$$

  The parameter k represents the lag spacing and specifies the temporal distance between two observations for which we measure the relationship.

- Autocorrelation: the autocorrelation function of an MA(q) process is given by $\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\sum_{i=0}^{q-k} \theta_i \theta_{i+k}}{\sum_{i=0}^{q} \theta_i^2}$. By definition, $\rho(0) = 1$, and for $k > q$, the autocorrelation is zero, reflecting the fact that there is no linear dependence between observations separated by more than q time steps [Thistleton and Sadigov, 2023c].

As explained before, we can also simplify the MA process using the backshift operator:

$$x_t = \theta(B) w_t,$$

where $\theta(B)$ represents the moving average operator:

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

Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \theta_2, \ldots, \theta_q$ [Shumway and Stoffer, 2011].

Autocorrelation Function (ACF)
Autocorrelation Function (ACF) helps determine the order q of a Moving Average process, since an MA(q) model has an ACF that cuts off after q lags [Thistleton and Sadigov, 2023b]. We first present the key mathematical steps involved in deriving the autocorrelation function, followed by an examination of its interpretation.
An AR(1) process can be expressed in its $MA(\infty)$ representation under the causality condition ($|\phi| < 1$), allowing it to be written as a linear combination of current and past white noise terms. This representation is essential for determining the autocovariance

structure and, in turn, the ACF. This result generalizes to AR(p) processes, which, using the backshift operator B ($Bx_t = x_{t-1}$), take the form [Thistleton and Sadigov, 2023d]:

$$\phi(B)x_t = w_t,$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$. If the roots of $\phi(B) = 0$ lie outside the unit circle, the process is stationary and $\phi(B)$ is invertible:

$$x_t = \frac{1}{\phi(B)} w_t.$$

For the AR(1) case, $\phi(B) = 1 - \phi B$, so:

$$x_t = (1 - \phi B)^{-1} w_t.$$

Using the geometric series expansion, valid for $|\phi| < 1$,

$$(1 - \phi B)^{-1} = 1 + \phi B + \phi^2 B^2 + \ldots,$$

and applying $B^j w_t = w_{t-j}$, we obtain:

$$x_t = w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \ldots,$$

or equivalently,

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}.$$

Thus, a stationary AR(1) process can be written as an $MA(\infty)$ process [Thistleton and Sadigov, 2023d]. Representing AR(p) processes in $MA(\infty)$ form enables straightforward derivation of their moments and dependence structure:

- Mean: $E[x_t] = 0$
- Variance: assuming $w_t \sim$ i.i.d.$(0, \sigma_w^2)$, $\text{Var}(x_t) = \sigma_w^2 \sum_{i=0}^{\infty} \theta_i^2$
- Autocovariance: $\gamma(k) = \sigma_w^2 \sum_{i=0}^{\infty} \theta_i \theta_{i+k}$
- Stationarity condition: $\sum_{i=0}^{\infty} |\theta_i| < \infty$ [Thistleton and Sadigov, 2023d]

Finally, the autocorrelation function (ACF) is obtained by normalizing the autocovariance:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}.$$

This framework allows AR(p) processes to be analyzed with the same tools as MA(q) processes, provided the $MA(\infty)$ series converges [Shumway and Stoffer, 2011].
After estimating the parameters, we can proceed to visually assess the AR(1) model's behavior through graphical analysis.

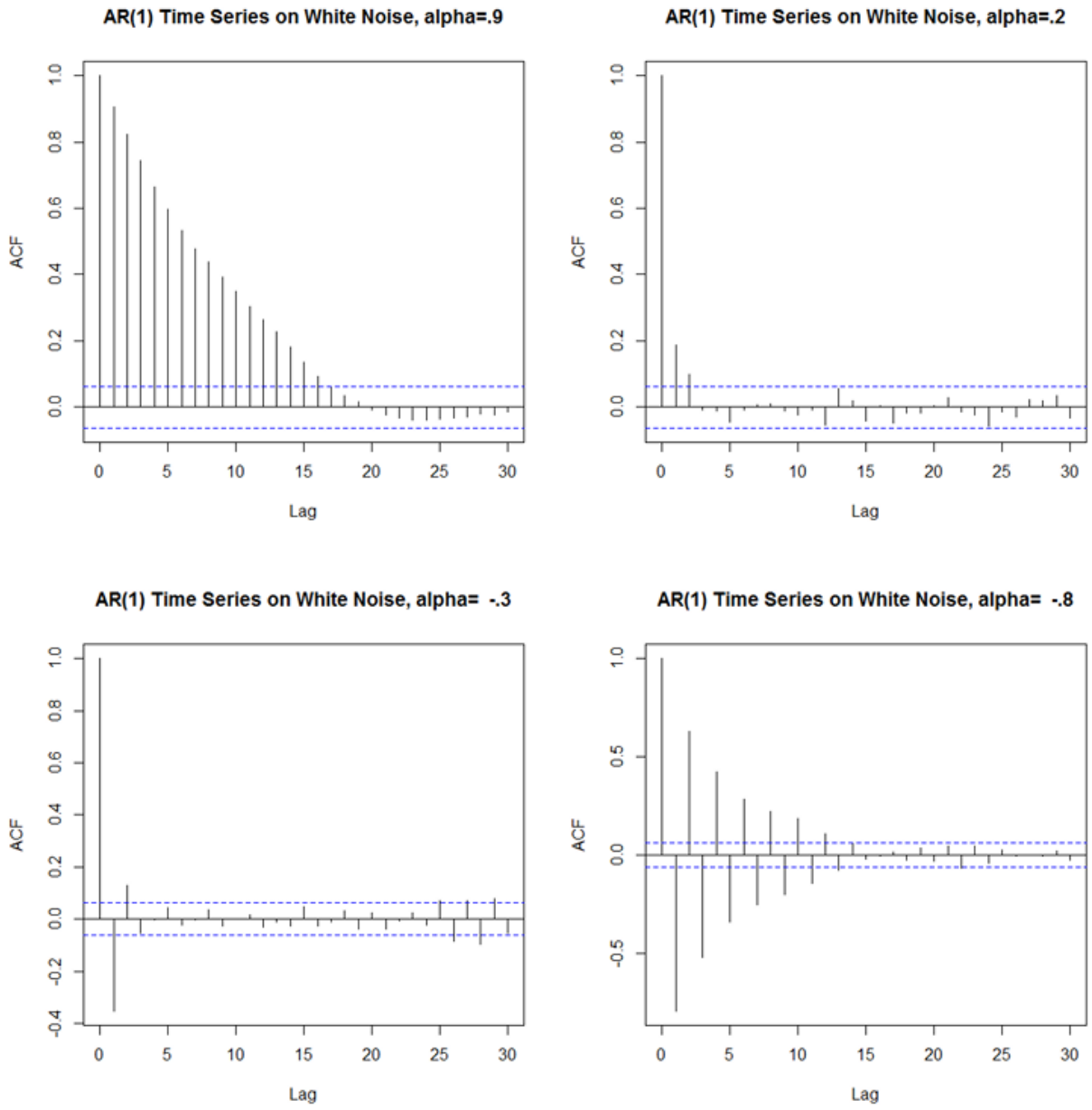

Figure 1: ACF plots of AR(1) processes with varying autoregressive coefficients. The parameter α used in this figure corresponds to the autoregressive coefficient ϕ adopted throughout the paper in the AR(1) model: $x_t = \phi x_{t-1} + w_t$.

Data source: [Thistleton and Sadigov, 2023d].

When the autoregressive coefficient ϕ_1 is close to 1, the AR(1) process behaves similarly to a random walk, which is nonstationary and exhibits increasing variance over time. As ϕ_1 decreases toward 0, the autocorrelations decay more rapidly, and when $\phi_1 = 0$, the process becomes pure white noise. For negative values of ϕ_1, the autocorrelations alternate in sign, causing the series to "flip back and forth" between positive and negative values [Thistleton and Sadigov, 2023d]. It is important to note that an AR process is not necessarily stationary. When $\phi = 1$, past data points remain significantly correlated with future observations over time. This behavior is characteristic of a random walk, where each point heavily depends on its immediate predecessor, leading to a nonstationary

process [Thistleton and Sadigov, 2023d].

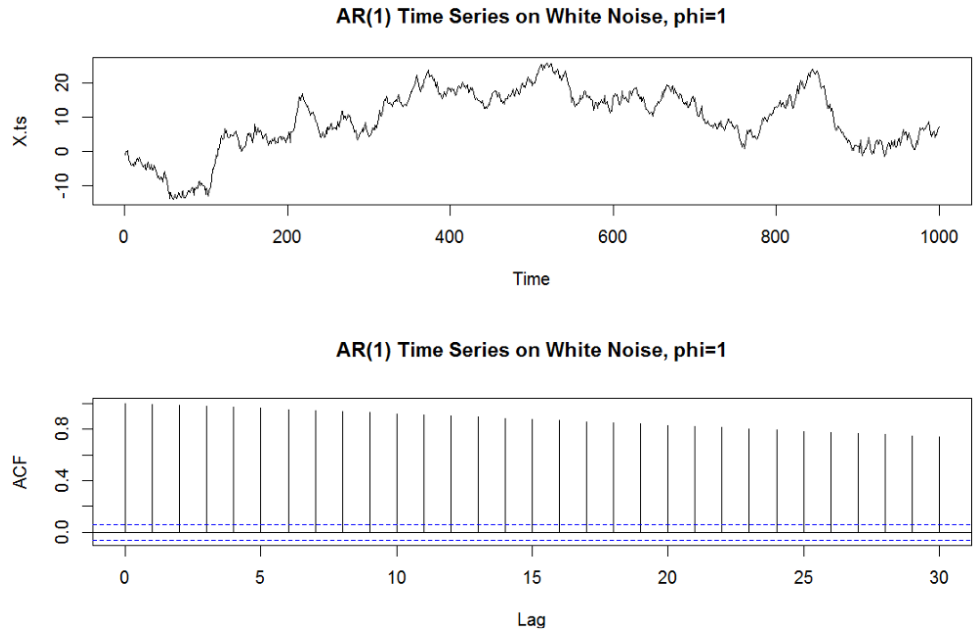

Figure 2: AR(1) process with $\phi = 1$, illustrating the behavior of a random walk and its persistent autocorrelation.

Data source: [Thistleton and Sadigov, 2023d].

If the data are stationary, the covariances drop off rapidly as the lag increases, such as:

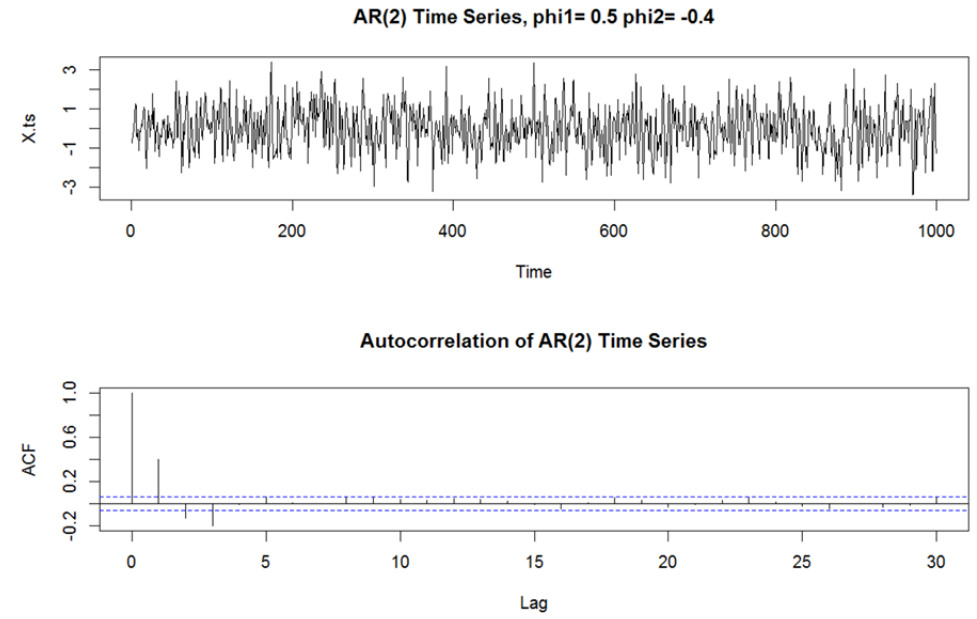

Figure 3: AR(2) process with $\phi_1 = 0.5$ and $\phi_2 = -0.4$, illustrating the behavior of a second-order autoregressive process and its decaying autocorrelation structure.

Data source: [Thistleton and Sadigov, 2023d].

Partial Autocorrelation Function (PACF)

To identify the order p of an Autoregressive process AR(p), we use the Partial Autocorrelation Function (PACF). In an AR(p) process, the partial autocorrelations are nonzero only up to lag p, and drop to zero thereafter. This means the PACF "cuts off" after p lags, making it a reliable tool for determining the order of the autoregressive model [Thistleton and Sadigov, 2023b]. Before examining the mathematical details of the PACF, it is important to understand the concept of redundancy in correlation. Two variables may appear correlated not because of a direct relationship, but due to the influence of a third variable affecting both. In time series, this occurs when the correlation between X_t and X_{t-k} is inflated by intermediate lags $(X_{t-1}, X_{t-2}, \ldots)$. The Partial Autocorrelation Function (PACF) removes the influence of these intermediate terms, isolating the direct impact of lag k and clarifying the structure of an autoregressive (AR) process. This makes the PACF a key tool for determining the correct lag order p in AR models, as it highlights only the non-redundant correlations between lags [Thistleton and Sadigov, 2023b].
In an AR(p) model, we assume that only the last p values of the series are relevant. The PACF helps determine p as follows:

- At lag k, the PACF measures the *direct relationship* between X_t and X_{t-k} after removing the effects of intervening lags.
- A significant PACF at lag k indicates that X_{t-k} provides unique information about X_t.

By isolating the true, non-redundant effect of each lag, the PACF identifies the appropriate order p for an AR model — specifically, by locating the point where the PACF "cuts off" (drops to zero) [Thistleton and Sadigov, 2023b].

We now examine two methods for computing the PACF at lag h. Although the theoretical definition based on conditional correlations is of conceptual interest, it is often neither the most efficient nor the most practical for empirical applications. Therefore, we focus on two alternative computational approaches [Shumway and Stoffer, 2011].
The first method introduced for computing the PACF at lag h is based on the use of residuals obtained from forward and backward regressions, and corresponds to the default implementation employed by R in our analysis. To estimate the Partial Autocorrelation Function (PACF) at lag h, we evaluate the relationship between x_{t+h} and x_t after removing the linear influence of the intermediate variables. This is done by performing two regressions:

- One that predicts x_{t+h} using the values $x_{t+1}, x_{t+2}, \ldots, x_{t+h-1}$

$$\hat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_2 x_{t+h-2} + \cdots + \beta_{h-1} x_{t+1}$$

  we estimate $\hat{x}_{t+h}$ by looking backward over the last several terms,

- Another that predicts x_t using the same set of intermediate values.

$$\hat{x}_t = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots + \beta_{h-1} x_{t+h-1}$$

  now we estimate $\hat{x}_t$ by looking forward over the next several terms.

Due to stationarity, the same regression coefficients $\beta_1, \ldots, \beta_{h-1}$ can be used in both directions. We then compute the correlation between the residuals of these two regressions — that is, the parts of x_t and x_{t+h} that are not explained by the intermediate values. This correlation is the partial correlation coefficient at lag h, denoted as ϕ_{hh}.

$$\phi_{hh} \equiv \text{corr}[x_{t+h} - \hat{x}_{t+h},\ x_t - \hat{x}_t]$$

In essence, the PACF isolates the direct linear dependence between two time points by removing the influence of the values that lie in between [Shumway and Stoffer, 2011].
The second method, known as the Yule–Walker approach, is one of the most commonly employed techniques for parameter estimation. Unlike the regression-based method described earlier, it relies on the values of the ACF to solve for the AR coefficients. Although this method is not employed in our analysis, it remains of interest. Due to space constraints, its detailed mathematical formulation is provided in the Appendix .1.

We now proceed to examine the interpretation of the Partial Autocorrelation Function (PACF). The Partial Autocorrelation Function (PACF) plot is another valuable tool for analyzing the underlying stochastic process that generated the observed data. It helps determine the number of autoregressive (AR) terms that should be included in a time series model.

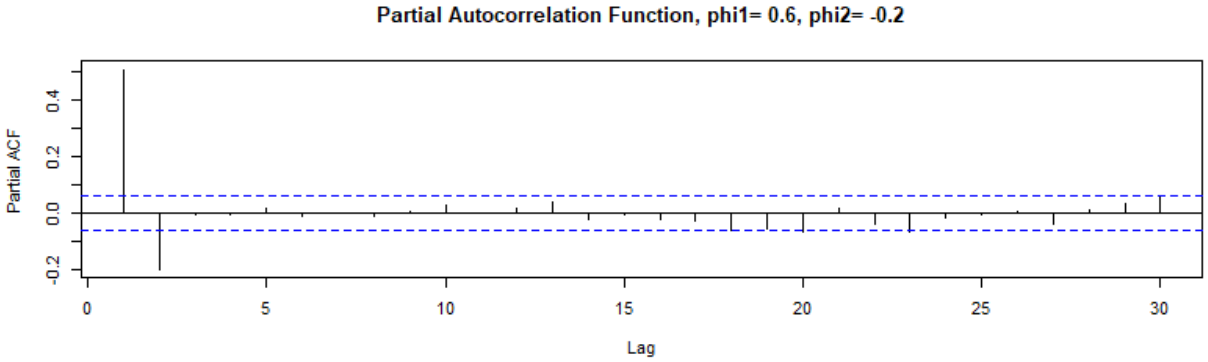

Figure 4: The plot displays the Partial Autocorrelation Function (PACF) for a time series generated by an AR(2) process with $\phi_1 = 0.6$ and $\phi_2 = -0.2$, illustrating the behavior of a second-order autoregressive process and its decaying autocorrelation structure. The x-axis represents the lag, while the y-axis shows the partial autocorrelation at each lag.

Data source: R simulation using `arima.sim()`.

The plot displays the Partial Autocorrelation Function (PACF) for a time series generated by an AR(2) process with parameters $\phi_1 = 0.6$ and $\phi_2 = -0.2$. The x-axis represents the lag, while the y-axis shows the partial autocorrelation at each lag. Key observations: significant spikes are observed at lag 1 and lag 2, indicating the presence of autocorrelation up to the second lag. Beyond lag 2, the partial autocorrelations fall within the 95% confidence interval bands (dashed blue lines), suggesting that additional lags do not significantly contribute to the model. This pattern is consistent with an AR(2) process, in which only the first two lags are expected to exhibit meaningful partial autocorrelations.

Summary of ACF and PACF
To aid in the identification of time series models, the table below outlines the characteristic

patterns of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) across AR(p), MA(q), and ARMA(p, q) processes. Recognizing these distinct behaviors is essential for selecting the appropriate model structure based on empirical autocorrelation analysis [Shumway and Stoffer, 2011].

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Figure 5: Behavior of the ACF and PACF for ARMA Models.

Data source:[Shumway and Stoffer, 2011] .

Autoregressive Moving Average Models - ARMA(p,q)

The ARMA model is used to describe stationary time series data by combining two components previously seen. The Autoregressive (AR) component of order p models the relationship between a value and its previous values, and the Moving Average (MA) component of order q models the relationship between a value and past error terms (shocks). The advantage of ARMA over separate AR or MA models is its ability to represent the autocorrelation structure more parsimoniously, requiring fewer parameters for comparable accuracy [Shumway and Stoffer, 2011].

A time series $\{x_t\}$ follows an ARMA(p, q) model if it satisfies:

$$x_t = \textit{AutoRegressive Part} + \textit{Noise} + \textit{Moving Average Part}$$

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q},$$

where:

- w_t is a white noise process with mean zero and constant variance σ_w^2,
- ϕ_i are the autoregressive parameters,
- θ_j are the moving average parameters,
- At least one of $\phi_p \neq 0$ and $\theta_q \neq 0$ must hold.

If the series has a non-zero mean μ (as in practical cases), it can be adjusted using a constant $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$, leading to the model:

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{-q}.$$

Take in consideration that if $q = 0$, the model becomes an autoregressive model of order p, denoted as AR(p), while if $p = 0$, the model becomes a moving average model of order q, denoted as MA(q) [Shumway and Stoffer, 2011].

The ARMA(p, q) model can also be compactly expressed using the backshift (lag) operator B:

$$\phi(B)x_t = \theta(B)w_t,$$

where $\phi(B)$ and $\theta(B)$ are polynomials in B corresponding to the AR and MA parts respectively [Shumway and Stoffer, 2011].

Common problems of ARMA

1. Models with too many parameters can lead to redundancy and overfitting, as they may capture noise rather than the underlying signal.
2. Autoregressive (AR) models that rely on future values violate the principle of causality; models should only depend on past and present information.
3. Moving Average (MA) models may not be uniquely defined unless we impose the condition of invertibility, which ensures that an ARMA model has a well-defined and unique representation. [Shumway and Stoffer, 2011].

To address the **first common problem**, an ARMA(p,q) model is considered valid only in its simplest form, that is, when its AR and MA polynomials have no common factors. Example – consider the following time series model:

$$x_t = 0.5x_{t-1} - 0.5w_{t-1} + w_t$$

At first glance, this appears to be an ARMA(1,1) process. However, both the AR and MA polynomials are:

$$\phi(B) = 1 - 0.5B \quad \text{and} \quad \theta(B) = 1 - 0.5B$$

Since they share the common factor $1 - 0.5B$, the model is not in its simplest form. To simplify, we rewrite the equation using the backshift operator B:

$$(1 - 0.5B)x_t = (1 - 0.5B)w_t$$

Dividing both sides by the common factor $1 - 0.5B$, we obtain:

$$x_t = w_t$$

Thus, the process reduces to white noise, and the original ARMA(1,1) specification was redundant. This example illustrates why we require ARMA(p, q) models to be in their simplest form, meaning the AR and MA polynomials must have no common factors. This ensures a unique representation and avoids unnecessary complexity or overfitting.
To address the **second common problem** we introduce the concept of causality: a time series model is considered causal when the current value x_t depends solely on present and past white noise terms, such as $w_t, w_{t-1}, w_{t-2}, \ldots$, and not on future values. This is called one-sided linear process and can be written as:

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)w_t$$

This is known as the *MA representation*, in which x_t is made up of all past shocks $w_t, w_{t-1}, w_{t-2}, \ldots$, each multiplied by its corresponding weight $\psi_0, \psi_1, \psi_2, \ldots$ [Shumway and Stoffer, 2011]. When an ARMA process is causal?
Example: Consider the AR(1) model:

$$x_t = \phi x_{t-1} + w_t$$

The process is causal if and only if $|\phi| < 1$. This condition ensures the model is stable and that x_t can be fully represented using only past and present information. Equivalenty, we can say that:

$$\begin{aligned} x_t &= \phi x_{t-1} + w_t \\ (1 - \phi B)x_t &= w_t \end{aligned}$$

Here, B is the backshift operator, defined by $Bx_t = x_{t-1}$. This allows us to express the model compactly using polynomial notation and to determine causality, we analyze its root by substituting B with a complex variable z, giving:

$$\phi(z) = 1 - \phi z$$

Solving $\phi(z) = 0$ yields the root $z = \frac{1}{\phi}$. The process is said to be causal and stationary if all the roots of the AR polynomial lie outside the unit circle — that is, $|z_i| > 1$ (or $z_i > 1$ if the root is real). To find the $\psi(z)$ we have to:

$$\begin{aligned} x_t &= \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \\ x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} &= w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \\ \phi(B)x_t &= \theta(B)w_t, \\ x_t &= \frac{\theta(B)}{\phi(B)} w_t = \psi(B)w_t, \end{aligned}$$

To analyze the convergence of the infinite series, we replace the backshift operator B with a complex variable z:

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1.$$

[Shumway and Stoffer, 2011]
In conclusion to address the **third common problem** we must explain invertibility of $ARMA(p, q)$ model. An $ARMA(p, q)$ model is said to be invertible if the white noise process w_t can be expressed as a linear function of the current and past values of the observed process x_t:

$$\pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t,$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ and $\sum_{j=0}^{\infty} |\pi_j| < \infty$.
Property: An ARMA(p, q) process is invertible if and only if

$$\theta(z) \neq 0 \quad \text{for all } |z| \leq 1,$$

where $\theta(z)$ is the moving average (MA) polynomial:

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

In other words, the process is invertible only if the roots of $\theta(z)$ lie outside the unit circle (i.e., $|z| > 1$). This condition ensures that the infinite sum used to represent w_t in terms of x_t converges. Under the invertibility condition, the coefficients π_j of the inverse MA representation can be determined using the generating function:

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1,$$

where $\phi(z)$ is the autoregressive (AR) polynomial. This representation is analogous to the causal MA(∞) form of an ARMA model, which expresses x_t in terms of past white noise shocks. In the invertible case, the inverse representation allows expressing the unobservable white noise process w_t in terms of observable past values of x_t, which is essential for estimation and filtering [Shumway and Stoffer, 2011].
In order to address the common issue of overfitting in practical implementation, we introduce the principle of parsimony: only SARIMA models satisfying the restriction $p + d + q + P + D + Q \leq 10$ are considered. This constraint reduces the risk of overparameterization, enhances model interpretability, and helps maintain a balance between model complexity and forecasting accuracy [Ledolter and Abraham, 2012]. While, the issues of causality and invertibility are automatically imposed by the `arima()` function during parameter estimation.

Autoregressive Integrated Moving Average Models - ARIMA(p,d,q)
Autoregressive Integrated Moving Average (ARIMA) models require the time series to be stationary. If it is not, differencing is applied until stationarity is achieved. The number of differences, denoted by d, defines the ARIMA(p, d, q) model [Shumway and Stoffer, 2011]. A process x_t follows an ARIMA(p, d, q) model if

$$\nabla^d x_t = (1 - B)^d x_t$$

where ∇^d is the d-th order differencing operator and B is the backshift operator, $Bx_t = x_{t-1}$. For example, if x_t follows a random walk, $x_t = x_{t-1} + w_t$, then first differencing ($d = 1$) gives $\nabla x_t = x_t - x_{t-1} = (1 - B)x_t = w_t$, which is stationary. While first differencing is often sufficient, higher-order differencing may be required in some cases.
In general, the ARIMA(p, d, q) model can be written as:

$$\phi(B)\nabla^d x_t = \theta(B)w_t$$

or, equivalently:

$$\phi(B)(1 - B)^d x_t = \theta(B)w_t$$

where $\phi(B)$ and $\theta(B)$ are polynomials in the backshift operator B representing the autoregressive and moving average components, respectively, and w_t is white noise [Shumway and Stoffer, 2011].

To analyze processes that evolve through small and stable percentage changes—such as financial returns—we will use the log transformation. In particular, we consider the difference of the logarithm of the series, $\nabla \log(x_t)$, which approximates the percentage change or growth rate when these changes are small. This transformation helps stabilize the variance and makes the process more suitable for modeling [Shumway and Stoffer, 2011].

A common formal test to determine whether a time series is stationary is the Augmented Dickey-Fuller (ADF) test [Dettling, 2016]. In addition to formal statistical tests, the Autocorrelation Function (ACF) plot is a valuable visual tool for assessing stationarity. For stationary series, the ACF typically declines rapidly toward zero, whereas non-stationary series display a slow, gradual decay. This contrast allows a visual distinction between stationary and non-stationary behavior [Shumway and Stoffer, 2011].

Box-Jenkins Seasonal ARIMA model (SARIMA)
Time series data often exhibit seasonal patterns, where observations are influenced not only by recent lags but also by periodic components repeating every s observations. To capture such patterns, the standard ARIMA framework is extended to account for both seasonality and nonstationarity. For example, in monthly data with a strong yearly component, current values X_t may depend on $X_{t-12}, X_{t-24}, \ldots$, with $s = 12$ denoting yearly periodicity. These seasonal effects often reflect natural or economic rhythms, such as temperature variations or fiscal quarters [Shumway and Stoffer, 2011].
To effectively model such behavior, Seasonal ARIMA (SARIMA) models incorporate seasonal lags into autoregressive and moving average structures, typically denoted as $\text{ARMA}(P, Q)_s$, where s is the seasonal period.

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t,$$

where the seasonal autoregressive operator $\Phi_P(B^s)$, with seasonal period s, is defined as $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$, while the seasonal moving average operator $\Theta_Q(B^s)$, also with seasonal period s, is defined as $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$. Here, P and Q denote the orders of the seasonal autoregressive and seasonal moving average components, respectively. A seasonal $\text{ARMA}(P, Q)_s$ model is causal if the roots of $\Phi_P(z^s)$ lie outside the unit circle, and invertible if the roots of $\Theta_Q(z^s)$ lie outside the unit circle—just like in non-seasonal ARMA models [Shumway and Stoffer, 2011], as established in the preceding section.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model consists of a non-seasonal component (p, d, q) and a seasonal component (P, D, Q), defined over a specified seasonal period s.

$$\Phi_P(B^s)\,\phi(B)\,\nabla_s^D \nabla^d x_t = \delta + \Theta_Q(B^s)\,\theta(B)\,w_t,$$

where:

- $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ is the non-seasonal MA.
- $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$ is the seasonal MA.

- $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the non-seasonal AR.
- $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$ is the seasonal AR.

The ordinary and seasonal differencing components are defined as $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$, respectively. The constant term δ is eliminated by the differencing operations, as these remove the mean level of the series. After differencing, the data is centered around zero, making δ redundant. However, if the model is applied to undifferenced data, δ remains part of the equation [Shumway and Stoffer, 2011].

Regarding the interpretation of the plot, consider a SARIMA$(0, 0, 1)(0, 0, 1)_{12}$ model with $\theta_1 = 0.7$ and $\Theta_1 = 0.6$, which combines short-term and seasonal MA components.

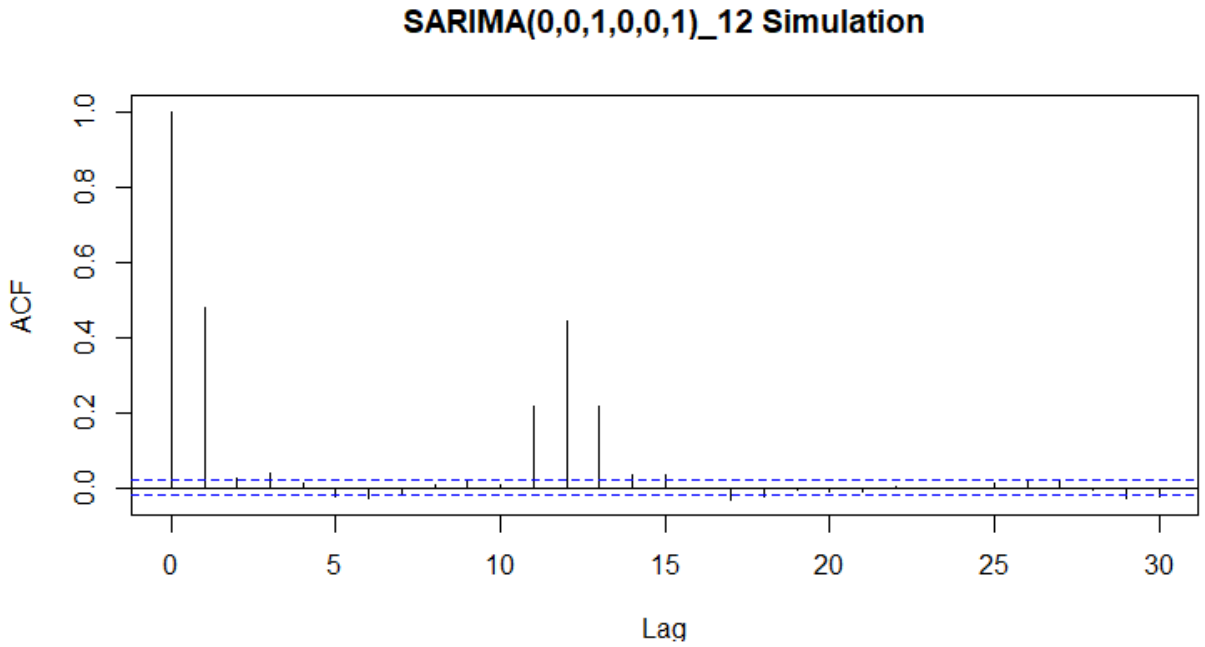

Figure 6: The ACF of the SARIMA(0,0,1)(0,0,1)$_{12}$ model with parameters $\theta_1 = 0.7$ and $\Theta_1 = 0.6$.

Data Data source: R simulation using `arima.sim()`

In the ACF of the simulation, we observe significant spikes at lags $ks = 1, 11, 12$, and 13, corresponding to both non-seasonal and seasonal moving average effects. This pattern is consistent with the expected behavior of a seasonal MA($Q = 1$) model, where the ACF cuts off after lag Qs. Specifically, the spike at lag 1 reflects the presence of the non-seasonal MA(1) component, while the spike at lag 12 that is $h = 1$, since $h = \frac{\text{lag}}{12}$, confirms the influence of the seasonal MA(1) term.

	AR$(P)_s$	MA$(Q)_s$	ARMA$(P, Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots,$ are zero.

Figure 7: Behavior of the ACF and PACF for SARMA Models.

Data source:[Shumway and Stoffer, 2011] .

According to table above, for a pure seasonal moving average process, the ACF is expected to cut off at lag Qs. The observed ACF behavior in the simulation align with this theoretical expectation, confirming the suitability of the SARIMA$(0,0,1)(0,0,1)_{12}$ specification for capturing the seasonal structure in the data [Shumway and Stoffer, 2011].

Akaike Information Criterion

After estimating the values of p using the Partial Autocorrelation Function (PACF) and q using the Autocorrelation Function (ACF), we must determine the optimal combination of these parameters.

To evaluate the quality of fit for different ARIMA models, we can use measures such as the *Sum of Squared Errors (SSE)* and the *Akaike Information Criterion (AIC)*. The SSE quantifies the total squared difference between observed and predicted values, helping identify models with better in-sample fit. However, minimizing SSE alone can lead to *overfitting*, especially with more complex models. The AIC addresses this by evaluating the relative performance of competing models: it rewards a better fit while penalizing model complexity, thus encouraging a balance between accuracy and parsimony. The standard procedure involves estimating model coefficients for each (p, q) specification, computing the *Residual Sum of Squares (RSS)*, and selecting the model with the lowest overall AIC [Thistleton and Sadigov, 2024].

The formula for the Akaike Information Criterion (AIC) in terms of the residual variance estimate and model complexity is given by:

$$AIC = \log(\hat{\sigma}^2) + \frac{n + 2 \cdot p}{n}, \quad \text{where} \quad \hat{\sigma}^2 = \frac{SSE}{n} \tag{1}$$

The value of p that minimizes the AIC identifies the most suitable model. Conceptually, minimizing $\hat{\sigma}$ would seem like a reasonable objective, but this quantity tends to decrease monotonically as p increases. Therefore, it is necessary to penalize the error variance by a term that is proportional to the number of parameters in the model [Shumway and Stoffer, 2011]. Ultimately, the goal is to compare the AIC values across a range of candidate models. As long as the comparisons are made consistently, we can draw meaningful conclusions. In general, the preferred model is the one that yields the lowest AIC. If the difference in AIC values between models is small, this indicates that the competing models have comparable performance, and the choice between them may be guided by considerations such as parsimony or interpretability [Thistleton and Sadigov, 2024].

Analysis of the residuals

After fitting the SARIMA model, it is essential to conduct a residual analysis to verify whether the model adequately captures the data's dynamics.

Residual analysis is crucial for evaluating the adequacy of time series models such as SARIMA. In a well-calibrated model, the standardized innovations e_t should behave like white noise—independent, identically distributed with zero mean and unit variance. If this holds, the model has captured the underlying data structure, and the error term contains

no systematic information. Post-estimation diagnostics focus on verifying that residuals meet key properties, such as absence of autocorrelation and approximate normality. These checks are essential for confirming both the reliability of the forecasts and the validity of the model's assumptions [Shumway and Stoffer, 2011]. While there are various methods for residual analysis, in this project I chose to evaluate them using a series of diagnostic plots:

1. Time Plot of Residuals: This allows a visual inspection of any patterns or trends that may remain in the residuals. Ideally, the plot should appear structureless and randomly scattered around zero.

2. Autocorrelation and Partial Autocorrelation Functions (ACF/PACF): These plots help detect serial correlation in residuals. In a properly specified model, the autocorrelations at all lags should lie within the 95% confidence bands, indicating no significant autocorrelation.

3. Normal Q-Q Plot: A quantile-quantile plot helps assess marginal normality. If the residuals are normally distributed, the points should lie approximately on the reference line. This visual test complements the histogram and provides additional insight into departures from normality.

4. Computation of residual variance to evaluate the model's predictive uncertainty.

5. Ljung–Box Test: A Q-statistic used to assess whether the residuals exhibit autocorrelation, which will be discussed in detail later in this work.

The combination of these diagnostic checks helps determine whether the residuals resemble white noise, which in turn indicates the adequacy of the model. In this case, the residuals exhibit no significant autocorrelation, follow an approximately normal distribution, and display no discernible patterns—suggesting that the ARIMA(p, d, q) model effectively captures the underlying dynamics of the data [Shumway and Stoffer, 2011].

The Ljung-Box Q-statistic is a test used in time series analysis to evaluate whether a series of residuals is independently distributed—that is, whether there is significant autocorrelation remaining in the data after fitting a model. The test builds on the Portmanteau statistic, defined as:

$$Q^*(m) = T \sum_{l=1}^{m} \rho_l^2$$

- T is the sample size

- ρ_l is the sample autocorrelation at lag l

- m is the number of lags being tested, $m \approx \ln(T)$

- l: Lag index (from 1 to m)

Under the assumption i.i.d., the Portmanteau statistic asymptotically follows a chi-squared distribution with m degrees of freedom $Q^*(m) \sim \chi^2(m)$. This means that if the residuals are truly uncorrelated, the test statistic should follow a chi-squared distribution with m

degrees of freedom. For finite samples, an improved version of the test was proposed by Ljung and Box to enhance its power. The modified test statistic is defined as [Thistleton and Sadigov, 2024]

$$Q(m) = T(T+2)\sum_{l=1}^{m}\frac{r_l^2}{T-l}$$

Following the guidance in Shumway and Stoffer [2011], the value of l can be chosen arbitrarily, but we fix it at 20, as suggested in the book, for the implementation of the test.
We now turn our attention to the hypothesis testing procedure. We now turn to the hypothesis testing procedure:

- Null hypothesis (H_0): No autocorrelation up to lag m,

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_m = 0$$

- Alternative hypothesis (H_a): At least one autocorrelation is non-zero,

$$H_a : \rho_i \neq 0 \quad \text{for some } i \in \{1, \ldots, m\}$$

We reject the null hypothesis of no autocorrelation if the test statistic $Q(m)$ is sufficiently large $Q(m) > \chi^2_\alpha$. Alternatively, in practice, software packages compute the p-value associated with the test. The null hypothesis is rejected if $p < \alpha$ where α is the chosen significance level (commonly 0.05). This test is a crucial step in residual diagnostics, ensuring that the fitted time series model has adequately captured the autocorrelation structure of the data [Shumway and Stoffer, 2011].

Practical Implementation of SARIMA model
In the practical implementation, the key distinction between the SARIMA model and the previously applied regression-based approaches lies in the data structure: SARIMA is estimated on a univariate time series of AAPL prices, whereas the other models relied on multivariate predictors. Another crucial aspect concerns the model assumptions that must be considered with respect to the data:

- Homogeneity of variance (homoscedasticity): the variability (variance) of a dependent variable is consistent [Nelson, 1998]. During the implementation of the model, I initially employed the full sample from 2010 to 2024. However, the variance in the seasonal differenced series exhibited signs of heteroskedasticity, likely attributable to structural breaks, regime shifts, and volatility clustering that naturally arise over long horizons in financial data. For this reason, the training sample had to be restricted to shorter time windows.

- Continuous variables should be measured at equal time intervals [Nelson, 1998]. This assumption is satisfied, as the daily data are recorded at uniform time intervals.

- Time periods for occurrences of an event should be of equal durations [Nelson, 1998]. This assumption is satisfied, as each observation in the dataset corresponds to a single trading day, ensuring equal time durations between consecutive observations in trading time.

- Sequential independence of outcomes is not required [Nelson, 1998]. The method doesn't assume that consecutive data points are independent; patterns or correlations between them are allowed.

- at least 50–100 data points needed (data-intensive) [Nelson, 1998].

As previously discussed, an important consideration during model fitting is the principle of parsimony: we will select only SARIMA models where $p + d + q + P + D + Q \leq 10$ to avoid overfitting [Ledolter and Abraham, 2012].

Several attempts were undertaken to identify an appropriate balance between training sample length and model performance. Different SARIMA specifications were estimated, incorporating alternative seasonal periods (weekly, monthly, quarterly, and yearly).
The final specification that achieved the best compromise between model fit and computational efficiency was

$$\text{SARIMA}(1, 1, 0)(0, 1, 2)_5,$$

obtained through a systematic grid search with the parsimony restriction $p+d+q+P+D+Q \leq 10$. The model was selected based on the minimization of the Akaike Information Criterion (AIC), which reached one of the lowest observed values (AIC $= -3050.313$), indicating a good trade-off between fit and complexity. The sum of squared errors (SSE) was also among the lowest observed (SSE $= 0.2957$), and the model returned a p-value of 0.4379 (greater than 0.05), further confirming the adequacy of the fit. The estimated coefficients highlight the relevance of the seasonal moving average component: *sma1* is highly significant ($p < 0.01$), while *sma2* is not significant at the 5% level. Although the latter does not contribute substantially in statistical terms, it was retained in order to preserve the seasonal dynamics of the process and to ensure model stability.
Model adequacy was further assessed through residual diagnostics. The residuals did not exhibit significant autocorrelation, as confirmed by the Ljung–Box test (p-value $\approx$ 0.38), suggesting that the SARIMA model has captured the main temporal dependencies and that the residuals can be regarded as white noise. Overall, the chosen specification balances parsimony, predictive accuracy, and residual adequacy, making it a suitable representation of the underlying process. Nevertheless, the presence of a non-significant coefficient may suggest possible simplifications in future refinements of the model.
Regarding the forecasting procedure, and following the approach described in [Parmezan et al., 2019], we adopt a multi-step-ahead iterative strategy with updates. Specifically, a rolling window is employed, advancing through the series and updating the model at each iteration with the newly observed real value.
The complete implementation of the SARIMA model, including the parameter search and diagnostic checks, is available on my GitHub repository: GitHub Repository (SARIMA model)

Bibliography

Dr. Marcel Dettling. Applied time series analysis, 2016. Lecture script, Zurich University of Applied Sciences.

Gil Eshel. Yule-walker equations, n.d. URL `https://people.duke.edu/~rnau/411arim.htm`. Lecture notes on time series analysis.

Johannes Ledolter and Bovas Abraham. Parsimony and its importance in time series forecasting. *Technometrics*, 23(4):411–414, 2012. doi: 10.1080/00401706.1981.10487687. URL `https://www.researchgate.net/publication/254332340_Parsimony_and_Its_Importance_in_Time_Series_Forecasting`.

1998 Nelson. *Statistical Methodology: V. Time Series Analysis Using Autoregressive Integrated Moving Average (ARIMA) Models*. Academic Emergency Medicine, 1998.

Antonio Rafael Sabino Parmezan, Vinicius M. A. Souza, and Gustavo E. A. P. A. Batista. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*, 484:302–337, 2019. doi: 10.1016/j.ins.2019.01.076.

Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer, New York, third edition, 2011. ISBN 978-1-4419-7864-6. doi: 10.1007/978-1-4419-7865-3. URL `https://github.com/nickpoison/tsa5/blob/main/textRcode.md`.

W. Thistleton and T. Sadigov. Practical time series analysis: Yule-walker equations and ar model estimation, 2023a. Lecture slides, Week 3.

William Thistleton and Elchin Sadigov. Partial autocorrelation and the pacf: First examples, 2023b. Practical Time Series Analysis, Week 3.

William Thistleton and Elchin Sadigov. Stationarity: Properties and examples, 2023c. Practical Time Series Analysis, Week 2.

William Thistleton and Orkhan Sadigov. Akaike information criterion and model quality, 2024. Week 4 lecture material, Practical Time Series Analysis.

William Thistleton and Rovshan Sadigov. Backshift operator and the acf: Week 2, 2023d. Lecture notes for Practical Time Series Analysis.

.1 APPENDIX: Yule–Walker Method

I believe it is still valuable to discuss one of the methods commonly used for parameter estimation.
Earlier in this document, we estimated the autocovariance values $\gamma(h)$, and subsequently derived the autocorrelation coefficients using $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$. We now proceed to utilize these results to estimate the model parameters of an auto-regressive (AR) model [Eshel, n.d.]. Consider that we have:

$$\rho_k = \sum_{j=1}^{p} \phi_j \rho_{k-j}, \quad \text{for } k = 1, 2, \ldots, p$$

that can be extended as:

$$\begin{aligned}
\rho_1 &= \phi_1\rho_0 + \phi_2\rho_1 + \phi_3\rho_2 + \cdots + \phi_p\rho_{p-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2\rho_0 + \phi_3\rho_1 + \cdots + \phi_p\rho_{p-2} \\
&\vdots \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p\rho_0
\end{aligned}$$

which can also be written in matrix form as:

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix} = \begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & \rho_0 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}$$

considering that $\rho_0 = 1$, the system becomes:

$$\underbrace{\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix}}_{\boldsymbol{\rho}} = \underbrace{\begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{pmatrix}}_{\mathbf{R}} \underbrace{\begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}}_{\boldsymbol{\Phi}}$$

This system can be compactly written as:

$$\mathbf{R}\boldsymbol{\Phi} = \boldsymbol{\rho}.$$

Since Φ represents the vector of AR coefficients and R is symmetric and full-rank, its inverse exists. Therefore, the Yule-Walker equations can be solved explicitly as:

$$\hat{\boldsymbol{\Phi}} = \mathbf{R}^{-1}\boldsymbol{\rho}.$$

[Eshel, n.d.]
In conclusion, it is also possible to estimate the innovation variance σ_Z^2. As an illustration, consider the following AR(2) process:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

with a variance of:

$$\text{Var}(X_t) = \phi_1^2 \text{Var}(X_{t-1}) + \phi_2^2 \text{Var}(X_{t-2}) + 2\phi_1\phi_2 \text{Cov}(X_{t-1}, X_{t-2}) + \sigma_Z^2$$

Since the process is stationary, we have:

$$\text{Var}(X_t) = \gamma(0), \quad \text{Cov}(X_{t-1}, X_{t-2}) = \gamma(1)$$

Substituting in:

$$\gamma(0) = \phi_1^2\gamma(0) + \phi_2^2\gamma(0) + 2\phi_1\phi_2\gamma(1) + \sigma_Z^2$$

Solving for σ_Z^2:

$$\sigma_Z^2 = \gamma(0)\left[1 - \phi_1^2 - \phi_2^2 - 2\phi_1\phi_2\frac{\gamma(1)}{\gamma(0)}\right] = \gamma(0)\left[1 - \phi_1^2 - \phi_2^2 - 2\phi_1\phi_2\rho_1\right]$$

Using the Yule-Walker equations:

$$\rho_1 = \phi_1 + \rho_1\phi_2 \quad \Rightarrow \quad \rho_1 = \frac{\phi_1}{1-\phi_2}, \quad \rho_2 = \phi_1\rho_1 + \phi_2$$

We can rewrite the expression as:

$$1 - \phi_1^2 - \phi_2^2 - 2\phi_1\phi_2\rho_1 = 1 - \phi_1\rho_1 - \phi_2\rho_2$$

Thus,

$$\sigma_Z^2 = \gamma(0)[1 - \phi_1\rho_1 - \phi_2\rho_2]$$

Than, using the Yule–Walker approach, and given the sample estimates $\hat{\phi}_1$, $\hat{\phi}_2$, along with the sample autocorrelations r_1 and r_2, we estimate:

$$\hat{\sigma}_Z^2 = c_0[1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2]$$

where c_0 is the sample variance of the process (i.e., the sample autocovariance at lag 0) [Thistleton and Sadigov, 2023a].