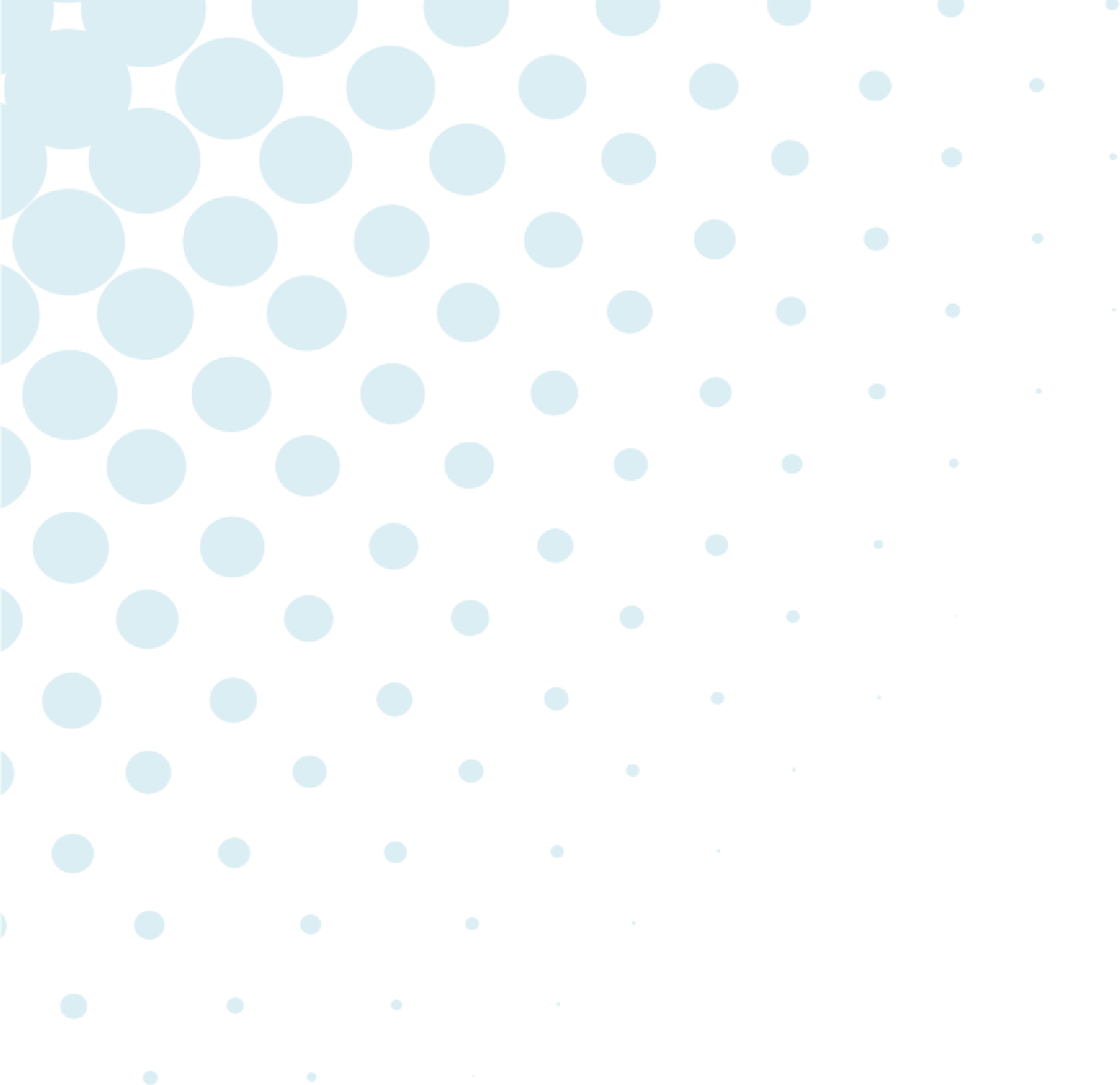
**Time Series Report**



(Electric Production)

Hadi Aldhaywi/Hussein Kshour

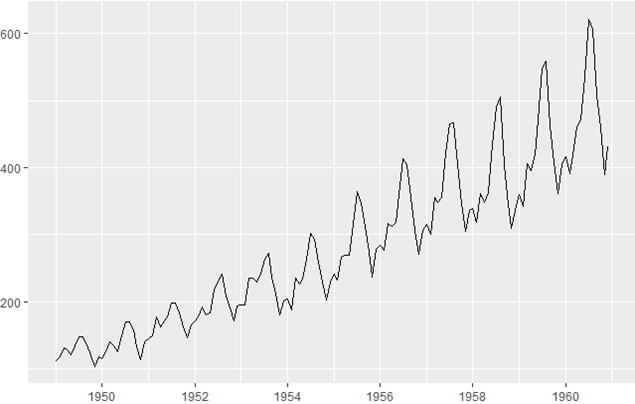
**Chapter 1: Technology watch Time series models**

# **Introduction:**

The objective of this monitoring is to introduce different approaches to develop a reliable forecasting model while explaining the associated methods in detail. An alternative approach to numerical models and a promoter in the construction of forecasting models is the time series data approach. We will explore several time series models by highlighting their characteristics, applications, and advantages, including SARIMA and ARIMA.

# **Data time series approach:**

Time series data involves the systematic collection of information at regular intervals. Its widespread application across diverse fields such as climate science, economics, engineering, environmental studies, medicine, and other scientific disciplines underscores its pivotal role in forecasting. Referred to as series, this data is collected at various time intervals, ranging from hourly to yearly observations.



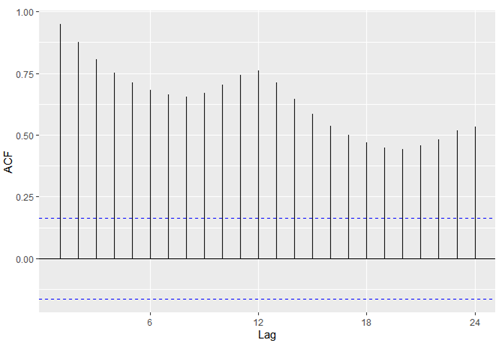
Analyzing a time series constitutes a crucial preparatory step preceding the development of forecasts. This analysis aims to unravel the underlying structures, trends, patterns, and phenomena evolving over time. By gaining insights into these aspects, decision-makers can make informed choices, predictions, and gather pertinent information to drive decision-making processes.

# **Characteristics of time series:**

## Autocorrelation

Autocorrelation is a prominent characteristic of time series data, indicating the correlation between successive observations. It signifies that the value at time 𝑡*t* partially depends on the value at the preceding time 𝑡−1*t*−1, implying a resemblance between consecutive observations. This resemblance is quantified by a correlation coefficient.

In practical analysis, the correlation coefficient is computed for various time lags, denoted as 𝑘*k*. For instance, correlations are calculated between 𝑦𝑡*yt*​ and 𝑦𝑡−1*yt*−1​ (lag 1), 𝑦𝑡*yt*​ and 𝑦𝑡−2*yt*−2​ (lag 2), and so forth. These correlation coefficients can be visually depicted on a correlogram, providing a graphical representation of the autocorrelation function (ACF).



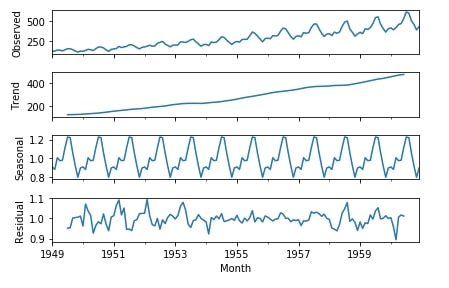
The dashed blue lines indicate whether the correlations are significantly different from zero.

# Decomposition of time series

A time series can be decomposed into three primary components:

* **Trend :** A trend represents the underlying function of the time series that changes gradually over time, capturing its long-term orientation.
* **Seasonality :** Seasonality refers to the periodic behavior observed in the time series, characterized by recurring patterns such as peaks and troughs that occur at regular intervals.
* **Residual or Error :** The residual, also known as the error term, comprises the part of the observation that cannot be explained by the trend or the seasonal component. It is essentially random noise in the data and is independent of the trend and seasonality.

To better understand, here is a breakdown of the time series.



The decomposition of a time series can be either additive (Series = Trend + Seasonality + Residual) or multiplicative (Series = Trend \* Seasonality \* Residual). Choose additive decomposition when the magnitude of seasonality remains constant over time, and opt for multiplicative decomposition when the seasonality's magnitude changes with the level of the series.

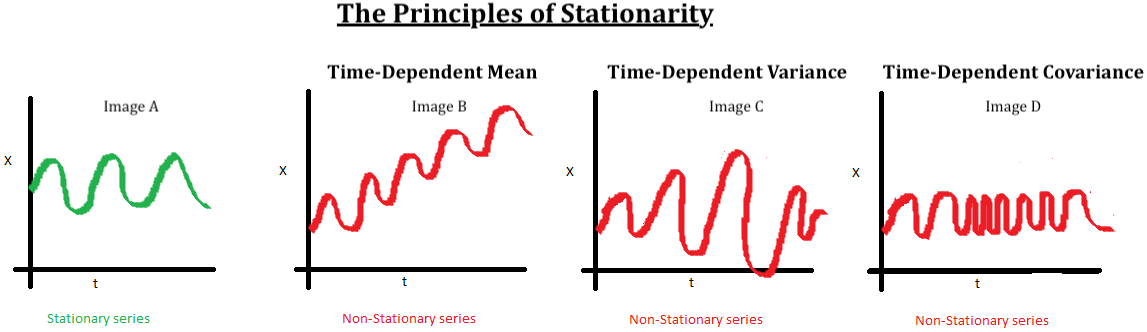
# Interest in time series

Time series analysis serves two main purposes:

1. Describing a Phenomenon: Time series analysis allows us to understand the behavior of a phenomenon over time. We can determine if the trend is increasing, decreasing, or stable, identify seasonal patterns and their frequency, and detect any abrupt changes or anomalies.
2. Making Forecasts: Time series analysis enables us to predict future observations based on historical data. By identifying patterns and trends in the data, we can develop models to forecast future outcomes.

# Stationarity of a time series

A time series is considered stationary if its statistical properties, such as mean, variance, and autocorrelation, remain constant over time. Stationarity implies the absence of trends or seasonality within the series.



# Why is stationarity so important ?

Stationarity is vital in time series analysis because it simplifies the prediction process. When a series is stationary, it allows us to make assumptions that future statistical properties will resemble those currently observed. This consistency facilitates more accurate and reliable predictions, as the underlying patterns in the data remain stable over time.

# Why make a non-stationary series stationary before making predictions?

An important reason for transforming a non-stationary series into a stationary one before making predictions is that many forecasting models, such as autoregressive models, rely on linear regression techniques. These models use the series' own lagged values as predictors. Linear regression performs optimally when the predictors are nearly independent variables. By stationarizing the series, we mitigate the issue of non-stationarity, ensuring that the predictors are more independent and improving the accuracy of the forecasting model.

# **Time series models :**

**ARMA (Autoregressive Moving Average)**

ARMA is a time series model that combines two components: autoregressive (AR) and moving average (MA). It models data where values depend on both past values (lags) and past errors to predict future values.

A purely autoregressive (AR) model of order 𝑝*p* is represented by:

𝑌𝑡=𝛼+𝛽1𝑌𝑡−1+𝛽2𝑌𝑡−2+…+𝛽𝑝𝑌𝑡−𝑝+𝜖𝑡

Similarly, a Moving Average (MA) model of order 𝑞*q* is represented by:

𝑌𝑡=𝜖𝑡+𝜃1𝜖𝑡−1+𝜃2𝜖𝑡−2+…+𝜃𝑞𝜖𝑡−𝑞

Applications: Financial forecasting, demand forecasting, economic data analysis.

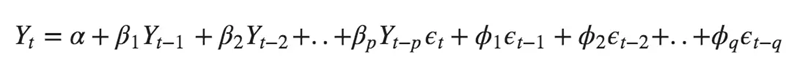
Advantages: Simplicity, efficiency for modeling stationary data with linear trends.

To assess stationarity, the series undergoes the Dickey-Fuller statistical test. If the series is found to be non-stationary, differencing is applied to achieve stationarity.

**ARIMA (Autoregressive Integrated Moving Average)**

ARIMA extends ARMA by adding an integration (I) component to handle non-stationary data. It is suitable for time series requiring differentiation of the current observation.

𝑌𝑡=Constant+Linear combination of lags of 𝑌+Linear combination of lagged forecast errors



Applications: Sales forecasting, climate modeling, economic analysis.

Benefits: Management of non-stationary data with patterns, short and long-term forecasting.

ARIMA is characterized by three parameters: *p*, *d*, *q*, where 𝑝 is the order of the AR model, 𝑞 is the order of the MA model, and 𝑑 is the number of differentiations necessary to achieve stationarity.

**SARIMA (Seasonal ARIMA)**

SARIMA extends ARIMA to account for periodic seasons in the data. It is characterized by seven parameters denoted as SARIMA(p,d,q)(P,D,Q)s, where 𝑠 is the period of seasonality and 𝑃, 𝐷, 𝑄 are the orders of the seasonal part.

Applications: Seasonal sales forecast, weather analysis, energy consumption forecast.

Benefits: Adapted to seasonal data, offers better precision**.**

# **Choice of model and Conclusion**

|  |  |  |
| --- | --- | --- |
| **Model** | **ARIMA** | **SARIMA** |
| **Type of problem** | without temperature | without temperature |
| **Data with seasonality** | Yes but remove seasonality | Yes |
| **Suitable for non-stationary data** | May require making stationary | Yes |
| **Using a popular library** | Yes | Yes |
| **Configuration complexity** | Moderate | Moderate |
| **Ability to manage large time series** | Limited | Limited |
| **It behaves like a black box** | No | No |

**Choice of Model:**

The selection of an appropriate time series model is paramount in accurately analyzing and forecasting data. Each model offers distinct advantages based on the characteristics of the dataset and the objectives of the analysis.

* **ARMA (Autoregressive Moving Average)** models are ideal for data exhibiting linear trends and stationary behavior. They offer simplicity and efficiency in capturing underlying patterns by considering both autoregressive and moving average components.
* **ARIMA (Autoregressive Integrated Moving Average)** models extend the capabilities of ARMA by incorporating an integration component to handle non-stationary data. Through differencing techniques, ARIMA models are adept at capturing trends and making forecasts for datasets with evolving statistical properties.
* **SARIMA (Seasonal ARIMA)** models further refine forecasting accuracy by accounting for seasonal patterns in the data. By integrating seasonal differencing and additional parameters, SARIMA models provide enhanced precision in capturing seasonal variations and making predictions over time.

**Advantages and Disadvantages of models**:

|  |  |  |
| --- | --- | --- |
| Model | Advantages | Disadvantages |
| **ARIMA (Autoregressive Integrated Moving Average)** | * Effective for a wide range of time series data that exhibit non-seasonal patterns. * Flexible in modeling various types of autocorrelation structures with its AR, I, and MA components. * Provides clear insights into the underlying patterns of the data. * Widely used and well-supported in many statistical software packages. | * Assumes stationarity after differencing, which may not hold true for all real-world data. * Does not inherently handle seasonal variations, which can be a limitation for data with strong seasonal effects. * Model identification (selecting appropriate p, d, q values) can be complex and subjective. * Sensitive to outliers, which can significantly affect the model’s accuracy. |
| **SARIMA (Seasonal ARIMA)** | * Extends ARIMA by explicitly modeling seasonal patterns, making it more effective for seasonal data. * Can handle both non-seasonal and seasonal fluctuations in the data. * Allows for more comprehensive modeling with additional seasonal parameters (P, D, Q, m). * Better suited for datasets with clear and strong seasonal trends, improving forecast accuracy. | * More parameters to estimate than ARIMA, which can increase model complexity and computational cost. * Risk of overfitting if too many parameters are used, especially in small datasets. * Requires careful tuning of both non-seasonal and seasonal components. * The selection of seasonal period (m) can be crucial and sometimes not straightforward. |

**Conclusion:**

In conclusion, time series analysis serves as a cornerstone in various fields, including finance, economics, and climate science. The utilization of ARMA, ARIMA, and SARIMA models offers valuable insights into temporal trends and facilitates informed decision-making processes.

By carefully evaluating the nature of the data and selecting the appropriate model, practitioners can derive meaningful interpretations, make accurate forecasts, and ultimately drive impactful outcomes. These models empower analysts to navigate the complexities of time-dependent data, thereby unlocking new opportunities for innovation and growth across diverse domains.

**Chapter 2: Forecasting short-term Electric production**

1. Introduction:

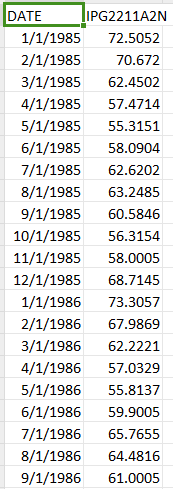
In the ever-evolving landscape of energy production, accurate forecasting of electric production transcends strategic importance to become an essential tool. The ability to reliably predict future electricity production is crucial, impacting a range of critical sectors from policy formulation to utility management. This influence extends across both economic and environmental aspects of society, underscoring the need for precise and adaptive forecasting methods as electricity demand fluctuates with technological innovations and seasonal variations.

This report presents a detailed analysis of the U.S. Electric Production Index, a pivotal measure of the nation's electric output, spanning from January 1985 to December 2017. In this study, we advance from the traditional AutoRegressive Integrated Moving Average (ARIMA) to the more complex Seasonal ARIMA (SARIMA) methodology. SARIMA extends the capabilities of ARIMA by effectively incorporating seasonal fluctuations, which are particularly pronounced in electric production data due to periodic changes in consumer behavior and climatic conditions.

The employment of SARIMA reflects a strategic choice motivated by its enhanced flexibility in modeling data that exhibit seasonal non-stationarities—a frequent trait of time series in economic and production indices. By harnessing SARIMA's comprehensive modeling framework, this analysis aims to decipher the intricate patterns of electric production and project future trends with heightened accuracy.

Through meticulous statistical testing and model refinement, this report strives to furnish stakeholders with more precise and dependable electric production forecasts. The ultimate goal is to equip decision-makers with robust analytical tools that support strategic planning aligned with anticipated energy demands and sustainability objectives, thereby fostering informed decision-making and proactive management in the energy sector.

1. Dataset:



The above figure shows the dataset of U.S. Electric Production that we’ll use in our models to forecast the U.S. Electric Production.

1. ARIMA study:

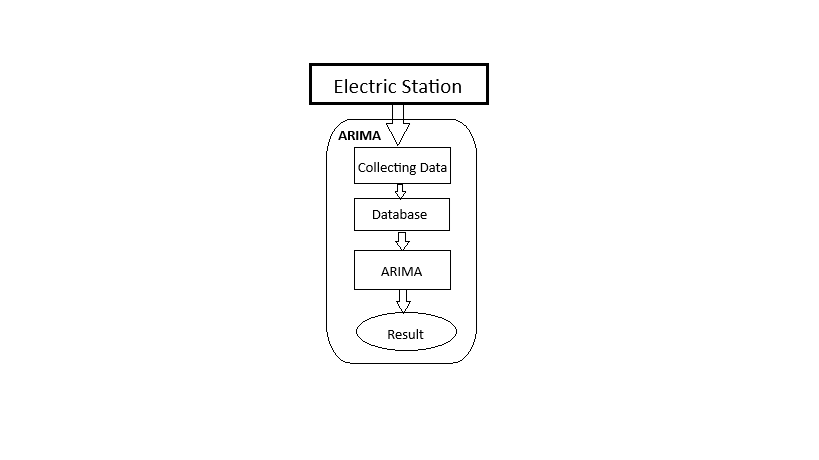


Figure1: predictive model to predict short-term electric production: ARIMA.

Forecasting with ARIMA:

Electric production data for a period of 32 years from January 1985 to December 2017 was collected from the U.S. Energy Information Administration (EIA). The ARIMA model uses this electric production time series data as input. The three mandatory parameters p, d, and q which must be selected, represent the order of the ARIMA model: p is the number of autoregressive terms, d is the number of differences required for stationarity, and q is the number of moving average terms. The ARIMA model is a robust statistical technique well-suited for modeling time series data in electric production forecasting.

Data collection procedure:

The data was sourced from the publicly available databases of the U.S. Energy Information Administration, which compiles comprehensive electricity production data across the country. This dataset includes monthly measurements of electric power generated across various sectors, which reflect the overall electric productivity.

Data processing:

The data is stored in structured query language (SQL) databases and is organized in CSV format for ease of access and manipulation. As part of the data processing steps, a Python script is used daily to extract, clean, and prepare the data for analysis. This script handles tasks such as parsing dates, managing missing values, and ensuring data quality. Additionally, the script is designed to perform initial transformations such as differencing the series to achieve stationarity, a prerequisite for ARIMA modeling.

Construction of the sequential ARIMA model:

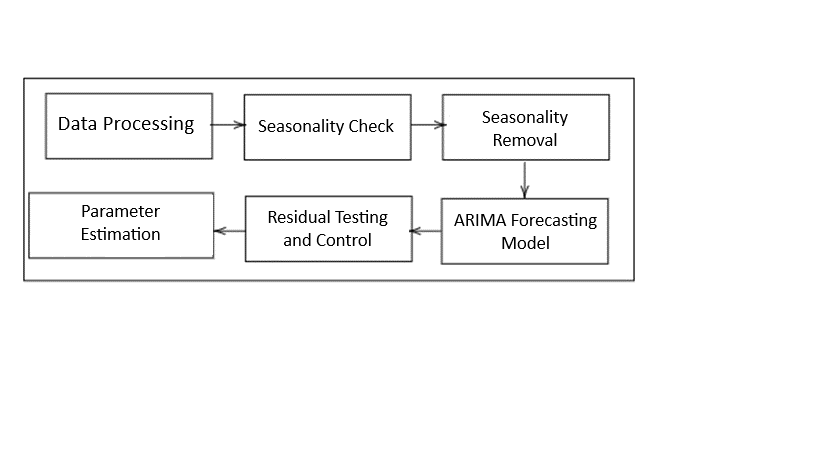


Figure2: Functional diagram of the ARIMA Model methodology

1. **Data Processing**

During this initial phase, the dataset is meticulously scrutinized for any anomalies, such as missing or zero data. Such entries, if found, are handled appropriately—typically, they are replaced with proximate average values calculated from the same time frame within the dataset. This ensures that the input data for the ARIMA model is both complete and robust, providing a solid foundation for accurate forecasting.

1. **Checking Seasonality**

To confirm the presence of seasonality in the electric production data, a thorough analysis is conducted on the historical data spanning from January 1985 to December 2017. The findings from this analysis reveal distinct patterns corresponding to seasonal fluctuations, which are crucial for the ARIMA model's accuracy and effectiveness.

This inherent seasonality, once identified, is addressed through differencing—the seasonal differencing (Xt - Xt−12) is typically employed, considering the monthly data collection frequency, to mitigate any seasonal effects before further analysis.

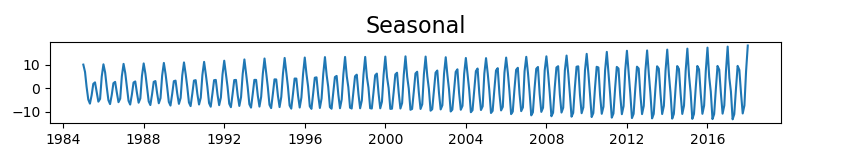


Figure3: Representation of seasonality for 9 years in the electric production series.

1. **Augmented Dickey-Fuller (ADF) test (before stationarity)**

The ADF (Augmented Dickey Fuller) statistical test is necessary to check whether the series obtained is stationary in trend or not.

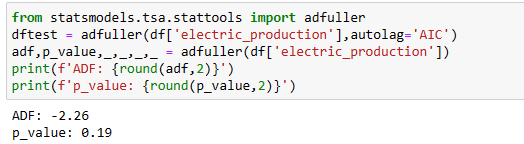


Figure4: ADF test

This figure (Figure 4) shows that after using the Augmented Dickey-Fuller (ADF) test, the p-value is 0.19, which is greater than 0.05 (0.19 > 0.05). This indicates that there is not enough evidence to reject the null hypothesis of a unit root, suggesting that the time series may not be stationary.

1. **Achieving stationarity**

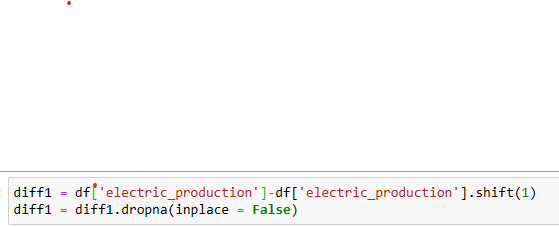


Figure5

The code in the above figure (Figure5) calculates the first difference of the 'electric\_production' time series to remove trends and help stabilize the mean to achieve stationarity. It then removes the resulting NA (not available) values to prepare the data for further time series analysis, ensuring the dataset is clean and consistent for accurate modeling.

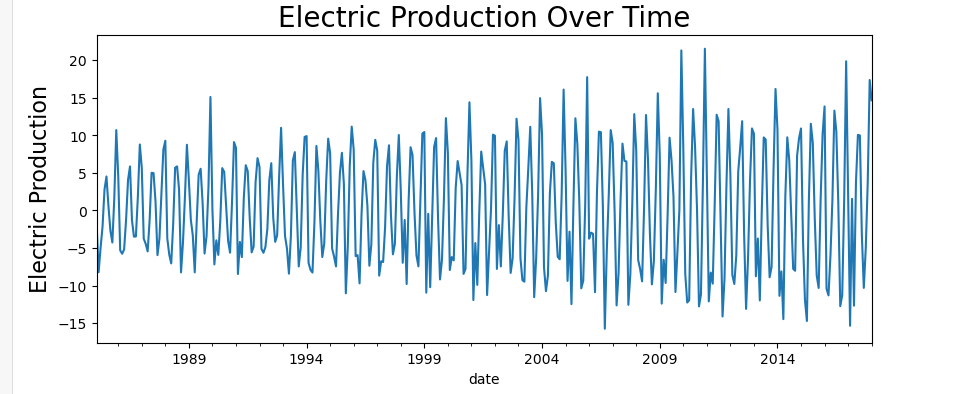
****

Figure6: Model after achieving stationarity

1. **Augmented Dickey Fuller (ADF) test (After stationarity)**

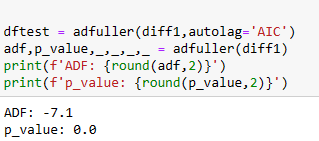


Figure7: ADF test

This figure (Figure 7) shows that after using the Augmented Dickey-Fuller (ADF) test, the p-value is 0.0, which is less than 0.05 (0.0 < 0.05). This indicates that there is a very strong statistical evidence against the null hypothesis of a unit root, suggesting that the time series is stationary.

1. **Choice of parameters p, d, q which satisfy the order of the ARIMA**

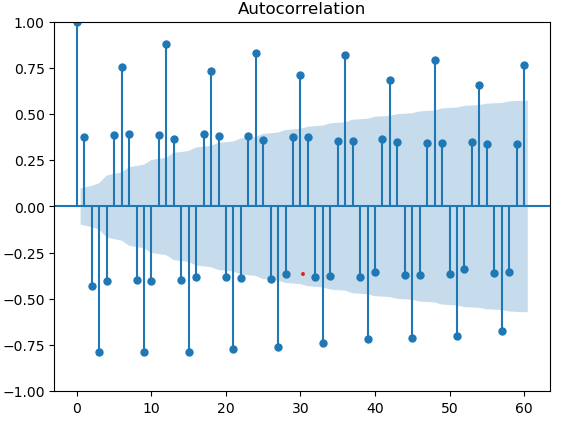
****

Figure8: Autocorrelation

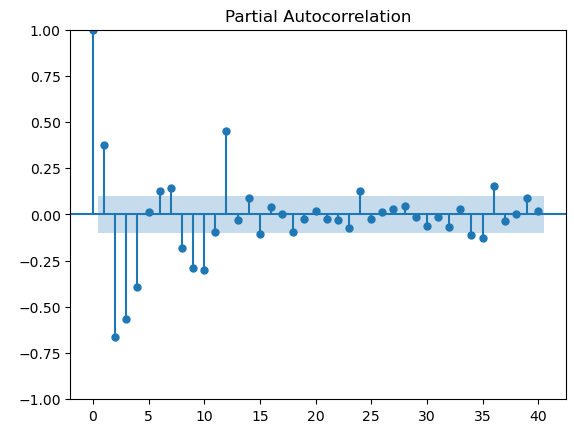
****

Figure9: Partial Autocorrelation

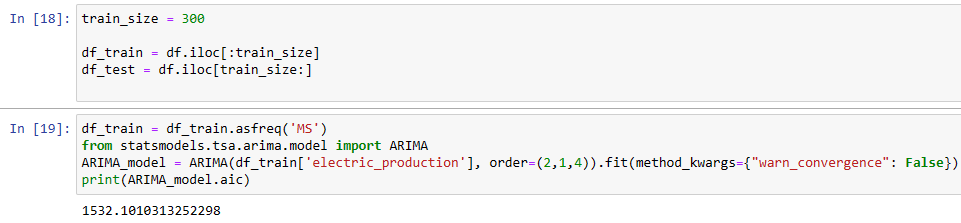


Figure10:ARIMA

The parameters p, d, q chosen to satisfy the order of the ARIMA model are informed by the characteristics observed in the actual ACF and PACF plots:

1. **p**: Referring to the PACF plot (now designated as Figure 9), the decision to use two autoregressive terms (p=2) is justified by the PACF showing significant correlations at the first two lags, after which it tails off. This indicates that the model needs to account for the impact of the last two observations in predicting the future values, implying the necessity of two AR terms.
2. **q**: As observed in the ACF plot (now designated as Figure 8), the selection of four moving average terms (q=4) is supported by the autocorrelations that extend prominently up to the fourth lag. This suggests that the residuals from up to four periods ago continue to affect the current value, which is a clear indication for the inclusion of four MA terms.
3. **Residue Monitoring and Testing**

After selecting the ARIMA model parameters p = 2, d = 1, and q = 4, validation of these parameters involves comprehensive checks:

1. **Residual Analysis**: The ACF and PACF of the residuals from the ARIMA(2, 1, 4) model should not show significant autocorrelation, ensuring that the residuals are effectively random. This check verifies that the model has successfully captured the essential information in the series, and the residuals resemble white noise, confirming an effective model fit.
2. **Model Summary and Statistics**

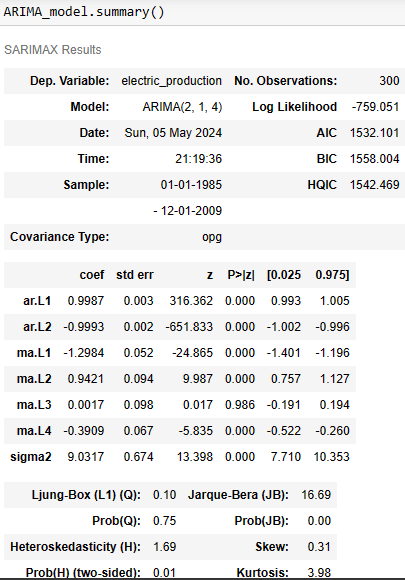


Figure11:ARIMA summary

* **Model Type**: ARIMA(2, 1, 4), implying an autoregressive component of order 2, one differencing step, and a moving average component of order 4.
* **Log Likelihood**: -759.051, which measures the model's goodness of fit; higher (less negative) values generally indicate a better fit.
* **AIC (Akaike Information Criterion)**: 1532.101, used to compare models; the lower the better. It penalizes complexity to avoid overfitting.
* **BIC (Bayesian Information Criterion)**: 1558.004, similar to AIC but with a stronger penalty for adding parameters, useful in model selection.
* **HQIC (Hannan-Quinn Information Criterion)**: 1542.469, another criterion to evaluate model fit with a penalty factor between AIC and BIC.

**Coefficient Estimates**

* **AR Coefficients**:
  + **ar.L1**: 0.9987 (p < 0.000), extremely significant with nearly complete positive autocorrelation.
  + **ar.L2**: -0.9993 (p < 0.000), extremely significant and suggests a strong negative autocorrelation immediately following the first lag.
* **MA Coefficients**:
  + **ma.L1**: -1.2984 (p < 0.000), shows a strong negative correlation with the immediate past value.
  + **ma.L2**: 0.9421 (p < 0.000), significant positive correlation two lags ago.
  + **ma.L3**: 0.0017 (p = 0.986), not significant, indicating no substantial effect at this lag.
  + **ma.L4**: -0.3909 (p < 0.000), significant negative effect four lags ago.

**Diagnostic Tests**

* **Ljung-Box Test**: Prob(Q) = 0.75, indicating no significant autocorrelation in residuals, suggesting that the model captures the autocorrelation structure effectively.
* **Jarque-Bera Test**: Prob(JB) = 0.00, indicates that the residuals do not follow a normal distribution, which might affect inference based on the model.
* **Heteroskedasticity Test**: Prob(H) = 0.01, significant heteroskedasticity in the residuals, suggesting varying residual variances over time, which could imply that volatility changes are not adequately captured by the model.

**Model Implications**

* The **coefficients** are highly significant, indicating a strong model with well-defined parameters.
* Despite a good fit (low AIC, BIC, effective Ljung-Box result), the **non-normal residuals** and **heteroskedasticity** indicate potential areas for model improvement, possibly through transformations or by incorporating GARCH models for variance modeling.
* The insignificant MA term at lag 3 suggests you could potentially simplify the model without losing explanatory power, which might also improve other diagnostics like normality and heteroskedasticity.

1. **Prediction model**

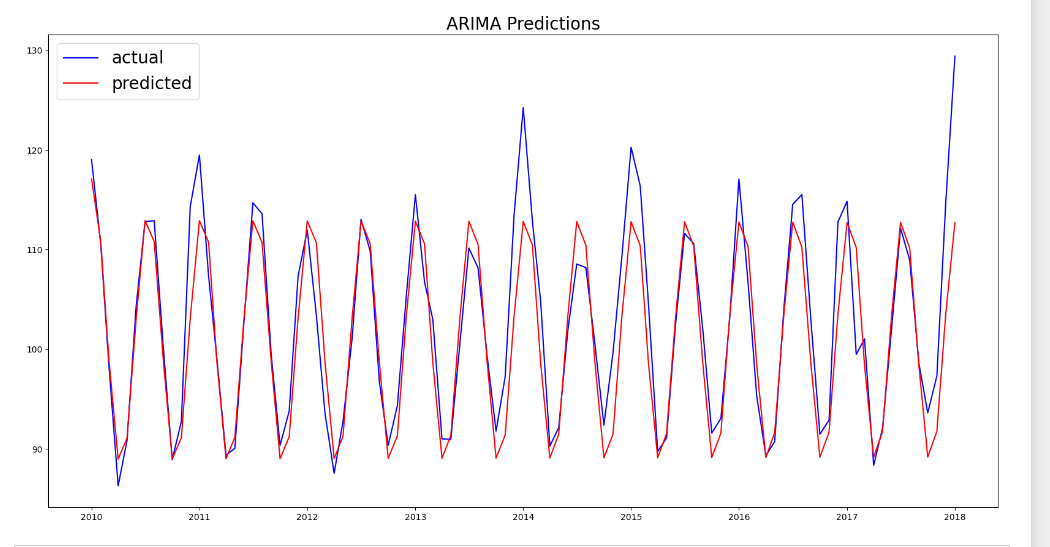


Figure12: actual and predicted values of ARIMA model

The graph (Figure12) shows actual (blue) versus predicted (red) values of electric production using an ARIMA(2, 1, 4) model from 2010 to 2018. Overall, the model performs well, closely tracking the seasonal patterns in the data. Notably, in 2014, the predicted values significantly underestimate the actual production, and in 2018 the predicted values deviate more noticeably from the actual values, with the predicted trend showing a sharper increase. This divergence suggests that the model might not fully account for some recent underlying changes or new trends in the data.

1. SARIMA study:

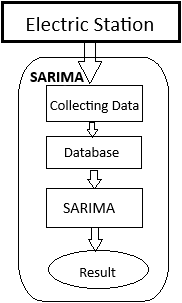


Figure1: predictive model to predict short-term electric production: SARIMA.

Forecasting with SARIMA:

Electric production data spanning 32 years, from January 1985 to December 2017, was collected from the U.S. Energy Information Administration (EIA). The Seasonal ARIMA (SARIMA) model is employed for analyzing and forecasting this electric production time series data. SARIMA extends the ARIMA model by adding seasonal elements, making it particularly adept at handling data with seasonal variations. The parameters of the SARIMA model include p, d, q for the non-seasonal component, and P, D, Q, m for the seasonal component: p is the number of autoregressive terms, d is the number of non-seasonal differences, q is the number of moving average terms, P is the number of seasonal autoregressive terms, D is the number of seasonal differences, Q is the number of seasonal moving average terms, and m represents the number of periods in each season. This model is highly effective for forecasting in contexts like electric production where data exhibit clear seasonal patterns.

Data collection procedure:

The data was sourced from the publicly available databases of the U.S. Energy Information Administration, which compiles comprehensive electricity production data across the country. This dataset includes monthly measurements of electric power generated across various sectors, which reflect the overall electric productivity.

Data processing:

The data is stored in structured query language (SQL) databases and is organized in CSV format for ease of access and manipulation. As part of the data processing steps, a Python script is used daily to extract, clean, and prepare the data for analysis. This script handles tasks such as parsing dates, managing missing values, and ensuring data quality. Additionally, the script is designed to perform initial transformations such as differencing the series to achieve stationarity, which is essential not only for ARIMA but also for SARIMA modeling, where seasonal differencing might also be included to account for periodic fluctuations. This preparation facilitates the effective application of SARIMA models to the time series data in electric production forecasting.

Construction of the sequential ARIMA model:

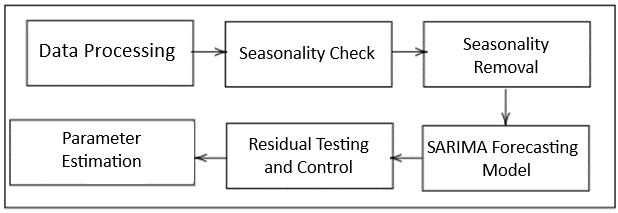


Figure2: Functional diagram of the SARIMA Model methodology

**a) Data Processing**

During this initial phase, the dataset is meticulously scrutinized for any anomalies, such as missing or zero data. Such entries, if found, are handled appropriately—typically, they are replaced with proximate average values calculated from the same time frame within the dataset. This ensures that the input data for the SARIMA model is both complete and robust, providing a solid foundation for accurate forecasting.

1. **Checking Seasonality**

To confirm the presence of seasonality in the electric production data, a thorough analysis is conducted on the historical data spanning from January 1985 to December 2017. The findings from this analysis reveal distinct patterns corresponding to seasonal fluctuations, which are crucial for the ARIMA model's accuracy and effectiveness.

This inherent seasonality, once identified, is addressed through differencing—the seasonal differencing (Xt - Xt−12) is typically employed, considering the monthly data collection frequency, to mitigate any seasonal effects before further analysis.

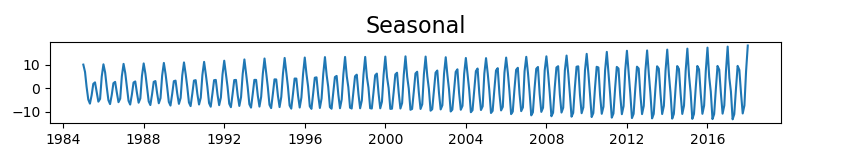


Figure3: Representation of seasonality for 9 years in the electric production series.

1. **Achieving stationarity (Removing Seasonality)**



Figure4

The code in the above figure (Figure4) demonstrates seasonal differencing applied to ‘electric production’ data. Seasonal differencing helps remove repeating patterns that occur at regular intervals, such as those observed in electric production data, typically on a yearly basis. This preprocessing step aims to make the data more stationary, which is essential for effective modeling with SARIMA.

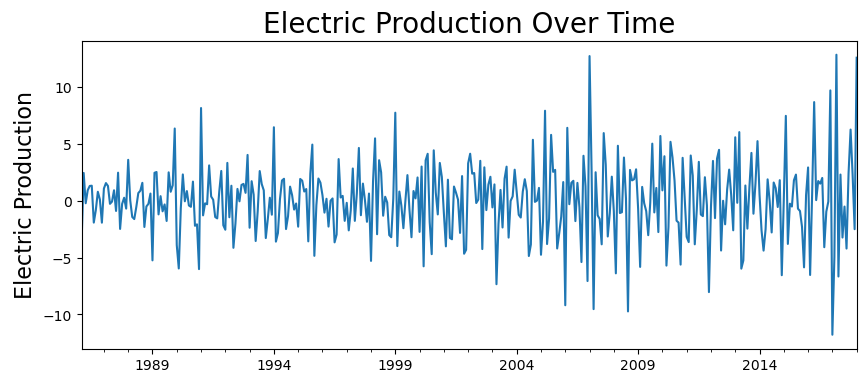


Figure5: Model after achieving stationarity (Removing Seasonality)

1. **Choice of parameters p, d, q, m, P, D, Q which satisfy the order of the SARIMA**

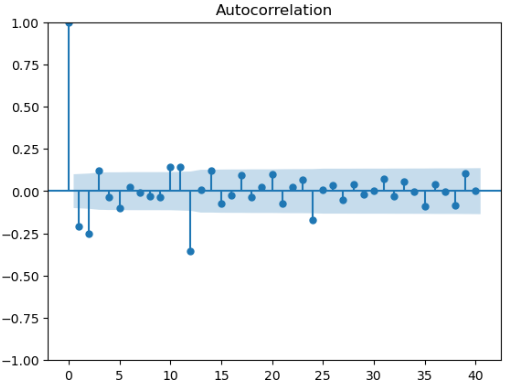
****

Figure6: Autocorrelation

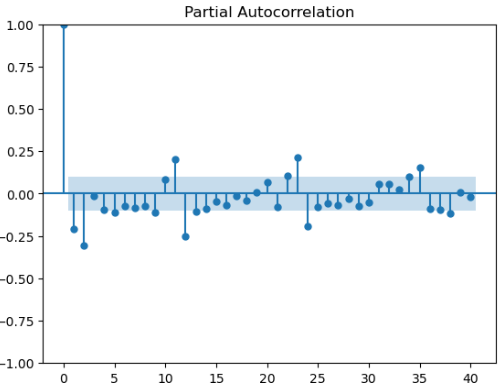
****

Figure7: Partial Autocorrelation



Figure8: SARIMA Model

The choice of parameters *p*, *d*, *q*, *m*, *P*, *D*, *Q* for the SARIMA model order is carefully guided by the characteristics observed in the actual ACF and PACF plots:

* **p**: The decision to use four autoregressive terms (p=4) is based on the PACF plot, where significant correlations are observed at the first four lags. This pattern indicates the necessity to account for the impact of the last four observations in predicting future values, hence the inclusion of four AR terms.
* **q**: As seen in the ACF plot, the selection of six moving average terms (q=6) is supported by autocorrelations that extend prominently up to the sixth lag. This suggests that the residuals from up to six periods ago continue to influence the current value significantly, warranting the inclusion of six MA terms.

The seasonal parameters *P*, *D*, *Q*, *m* are incorporated to explicitly model the seasonal patterns evident in the data. These values are chosen empirically based on the observed seasonal behavior:

* **P and Q**: The seasonal autoregressive and moving average terms are selected based on the seasonal spikes in the PACF and ACF plots, respectively, adjusted to capture the primary seasonal effects evident at annual intervals (12 months).
* **D**: Seasonal differencing is applied once (D=1) to remove the broad seasonal trends, as confirmed by a reduction in seasonal patterns in the differenced data.
* **m**: The seasonal cycle is set to 12 months, reflecting the annual pattern typical in electric production data.

1. **Root Mean Square Error (RMSE)**

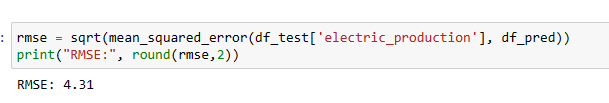


Figure9:RMSE of ARIMA

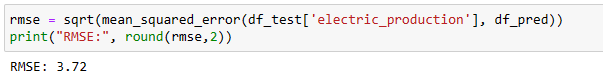


Figure10:RMSE of SARIMA

The Root Mean Square Error (RMSE) provides a clear metric for comparing the performance of the SARIMA and ARIMA models in forecasting electric production. For the SARIMA model, the RMSE is 3.72, indicating a closer fit to the actual data compared to the ARIMA model, which has an RMSE of 4.31. This improvement in RMSE with the SARIMA model suggests that incorporating seasonal components significantly enhances the model's accuracy, making it more effective for predicting patterns that are influenced by seasonal trends in electric production. The lower RMSE for the SARIMA model reflects its superior ability to model and predict the complex seasonal dynamics present in the electric production time series.

1. **Residue Monitoring and Testing**

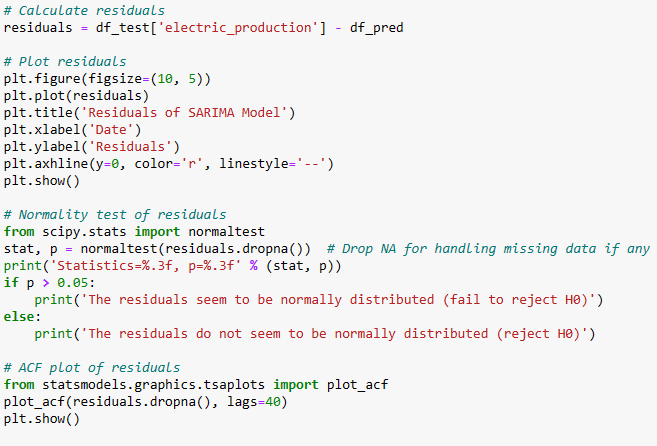


Figure11

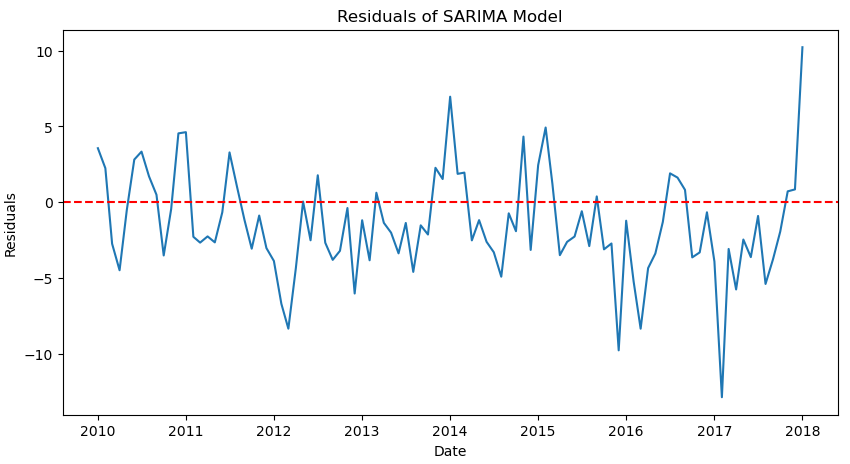


Figure12: Residuals Plot



Figure13: Normality Test

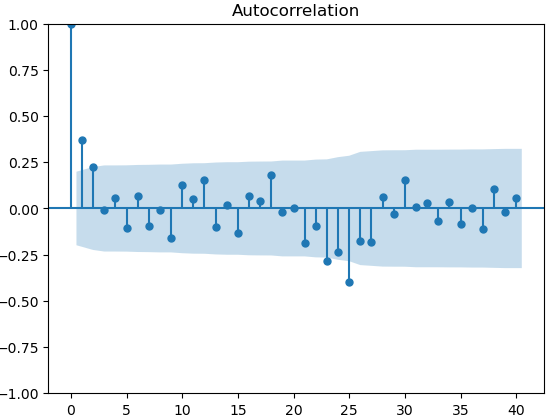


Figure14: Autocorrelation of Residuals

* **Residuals Plot (Figure12)**

The plot of the residuals from the SARIMA model shows fluctuations around zero, which is a good indication that the model does not have any obvious biases in predicting electric production. However, the variability of the residuals appears to increase over time, which might suggest the presence of heteroscedasticity. This could be an area for further investigation, possibly addressing it with transformations or variance-stabilizing techniques in the model.

* **Normality Test (Figure13)**

A normality test on the residuals (Shapiro-Wilk test) resulted in a p-value of 0.038, leading us to reject the null hypothesis that the residuals are normally distributed. This lack of normality could impact the reliability of some statistical inferences made from the model, and might suggest that the model could be improved by considering non-linear transformations or the inclusion of additional predictors to capture all the variability in the data.

* **Autocorrelation of Residuals (Figure14)**

The autocorrelation plot of the residuals shows several spikes outside the confidence bounds for different lags, suggesting that there are still autocorrelations present in the residuals. This indicates that the SARIMA model might not be fully capturing the dependent structure in the data. Further adjustments to the model's parameters, possibly exploring higher order differencing or adding more MA and AR terms, could help in capturing these dependencies more effectively.

* **Ljung-Box test**

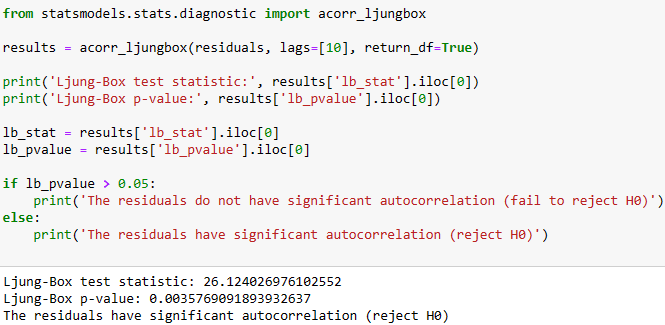
****

Figure15: Ljung-Box test

The Ljung-Box test results for the residuals of our SARIMA model, with a test statistic of 26.12 and a p-value of 0.0036, significantly indicate the presence of autocorrelation at lag 10. This suggests that the model fails to capture all the underlying temporal dependencies within the electric production data. The presence of autocorrelation in the residuals implies that the model may be underspecified or missing key predictors, which could affect the reliability and accuracy of our forecasts.

* Breusch-Pagan Test

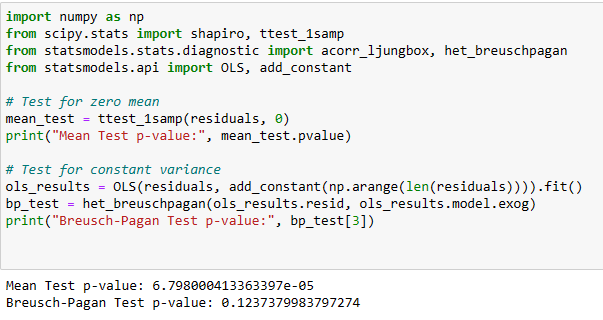


Figure16: Breusch-Pagan Test

The diagnostic tests on the SARIMA model's residuals reveal a statistically significant non-zero mean (p-value ≈ 6.80×10−56.80×10−5), indicating potential model bias. Conversely, the Breusch-Pagan test shows no evidence of heteroscedasticity (p-value = 0.1238), suggesting constant variance across residuals. These findings imply that while the model's predictions are stable across different values, adjustments are necessary to correct for the systematic bias to improve the model's accuracy

1. **Prediction model**

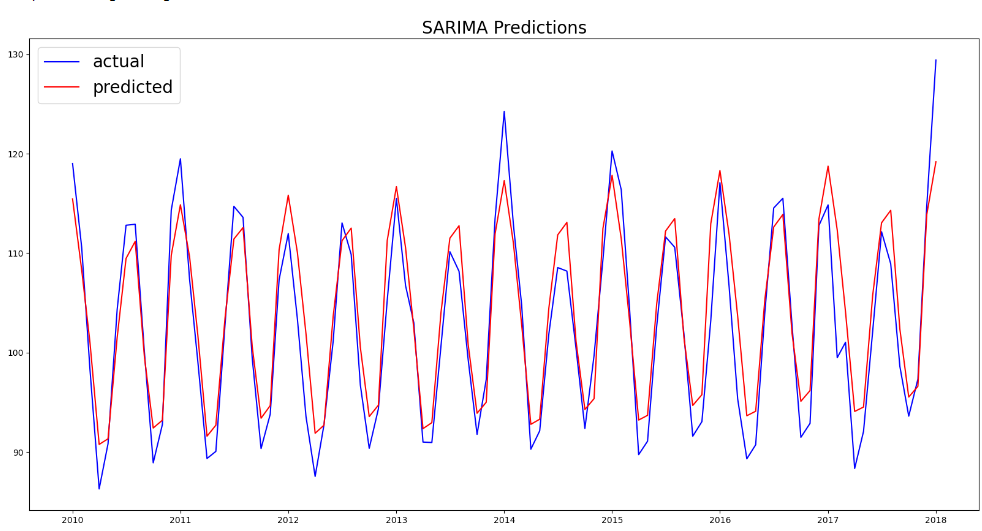


Figure17: actual and predicted values of SARIMA model

The graph (Figure17) shows actual (blue) and predicted (red) electric production values using the SARIMA model from 2010 to 2018. The model effectively captures the seasonal patterns, closely aligning with the actual data. However, there are noticeable discrepancies, such as in 2014 where it underestimates peaks, and in 2018 where it overpredicts, suggesting potential oversensitivity to recent trends or unaccounted external factors.

Conclusion:

This report has demonstrated the application of SARIMA and ARIMA models to forecast the U.S. Electric Production Index. The SARIMA model, in particular, has shown a superior ability to account for seasonal variations, as evidenced by a lower RMSE compared to the ARIMA model. However, the presence of autocorrelation in the residuals, as indicated by the Ljung-Box test, suggests the need for further refinement of the model.

Looking forward, we aim to explore multivariate time series analysis, which will allow us to incorporate multiple influencing factors simultaneously, enhancing the robustness and accuracy of our forecasting models. This advancement will enable us to adapt more dynamically to the complexities of electric production forecasting, ensuring our models remain effective in supporting energy management and policy development.

**Chapter 3: Time series perspective**

Introduction

In the pursuit of enhancing our forecasting accuracy for electric production, it becomes imperative to explore methodologies that extend beyond the traditional boundaries of univariate analysis represented by ARIMA and SARIMA models. This chapter introduces a strategic shift towards incorporating multivariate time series models, which allow for the simultaneous analysis of multiple interdependent variables. This transition is motivated by the need to understand and model the complex interactions between various factors that influence electric production, such as economic conditions, regulatory changes, and environmental factors. By embracing multivariate models, specifically Vector Autoregression (VAR), we aim to capture the dynamic interplays at work and improve the robustness and predictive power of our forecasts. This chapter will outline the rationale for this methodological advancement, discuss the potential models to be employed, and explore the implications of this shift for our forecasting capabilities.

Future Directions in Forecasting Models

As we continue to refine our forecasting techniques for electric production, it is crucial to expand our analytical framework beyond the current SARIMA and ARIMA models. While these models have proven effective in capturing the temporal dynamics of the series, they are inherently limited to univariate data. To address the complex interdependencies between multiple influencing factors, such as economic indicators, weather conditions, and policy changes, we plan to explore multivariate time series models.

**Introduction to Multivariate Models**

Multivariate models allow for the simultaneous analysis of several data series, which can provide a more holistic understanding of the factors influencing electric production. By integrating multiple time series, these models can capture the interactions and causal relationships between variables, offering richer insights and more accurate predictions.

**Focus on Vector Autoregression (VAR)**

One promising multivariate model is Vector Autoregression (VAR). The VAR model is particularly suited for forecasting systems where the variables influence each other interdependently. It extends the concept of univariate autoregression by creating a system of equations where each variable is a linear function of past lags of itself and past lags of other variables. This approach is powerful for modeling the dynamic relationships within the data, making it ideal for scenarios where variables are expected to interact with each other over time.

**Advantages of VAR**

* **Interdependency Modeling**: VAR can model the joint behavior of several time series, making it invaluable for understanding complex systems where variables interact with each other.
* **Impulse Response Analysis**: VAR allows analysts to investigate the impact of shocks to one variable on the entire system over time, which is crucial for scenario analysis and planning.
* **Forecasting Accuracy**: By incorporating multiple data streams, VAR can potentially offer more accurate forecasts than univariate models, particularly when variables are closely interlinked.

**Implementation Considerations**

Implementing VAR requires careful consideration of the variables included, as each additional variable increases the model complexity and computational demand. It is also critical to ensure stationarity of all series involved, possibly necessitating differencing or transformation of the data. Model selection, in terms of lag length and variable inclusion, must be guided by both theoretical considerations and empirical data analysis, such as using information criteria like AIC or BIC for optimal lag selection.

Conclusion

Expanding our analytical capabilities to include VAR and other multivariate models represents a significant step forward in our forecasting methodology. These models will allow us to incorporate broader data sets and more accurately predict electric production under a variety of conditions, ultimately supporting more informed decision-making and strategic planning in the energy sector.

**Chapter 4: Conclusion**

Reflecting on Our Journey with Time Series Forecasting:

As we conclude our exploration into the application of ARIMA and SARIMA models for forecasting the U.S. Electric Production Index, we revisit our journey through the intricate world of time series analysis. This report aimed to advance electric production forecasting by employing robust statistical methods that could handle the complexities of time-dependent data characterized by trends, seasonality, and other cyclical changes.

Encountered Challenges:

Throughout this project, several challenges were encountered:

* **Modeling Complexities**: ARIMA and SARIMA models, while powerful, presented complexities in parameter selection and model fitting, especially when addressing data non-stationarity and seasonal adjustments.
* **Data Issues**: We faced challenges related to data quality, including missing values and anomalies, which required meticulous preprocessing to ensure reliable model inputs.
* **Computational Demands**: The intensive computational requirements for model training and validation posed limitations, affecting our ability to rapidly iterate and refine our models.

Insights and Knowledge Gained:

Despite these challenges, our efforts were not without substantial gains:

* **Enhanced Forecasting Techniques**: We have significantly enhanced our forecasting techniques, gaining a deeper understanding of how to model and predict electric production dynamics effectively.
* **Statistical Proficiency**: Working with ARIMA and SARIMA models has improved our statistical proficiency, enabling us to better understand and implement these models for complex forecasting tasks.
* **Practical Implications**: The insights gained from this study have practical implications for energy management and policy formulation, helping stakeholders make informed decisions based on robust forecasts.

Future Directions:

Looking ahead, the experience gained has paved the way for future research:

* **Multivariate Time Series Models**: Encouraged by the limitations of univariate models, we plan to explore multivariate time series models. These models will allow us to incorporate multiple influencing factors simultaneously, potentially enhancing the accuracy and robustness of our forecasts.
* **Continued Methodological Improvement**: We will continue to refine our forecasting methodologies, incorporating advances in computational techniques and new statistical methods to stay at the forefront of time series forecasting.

Closing Thoughts:

This project has been a profound learning experience, providing us with valuable insights into both the potential and limitations of time series forecasting models. As we move forward, we are inspired to continue our work in this field, aiming to develop more sophisticated models that can better serve the needs of the energy sector. The challenges we encountered have only strengthened our resolve to push the boundaries of what these forecasting models can achieve.