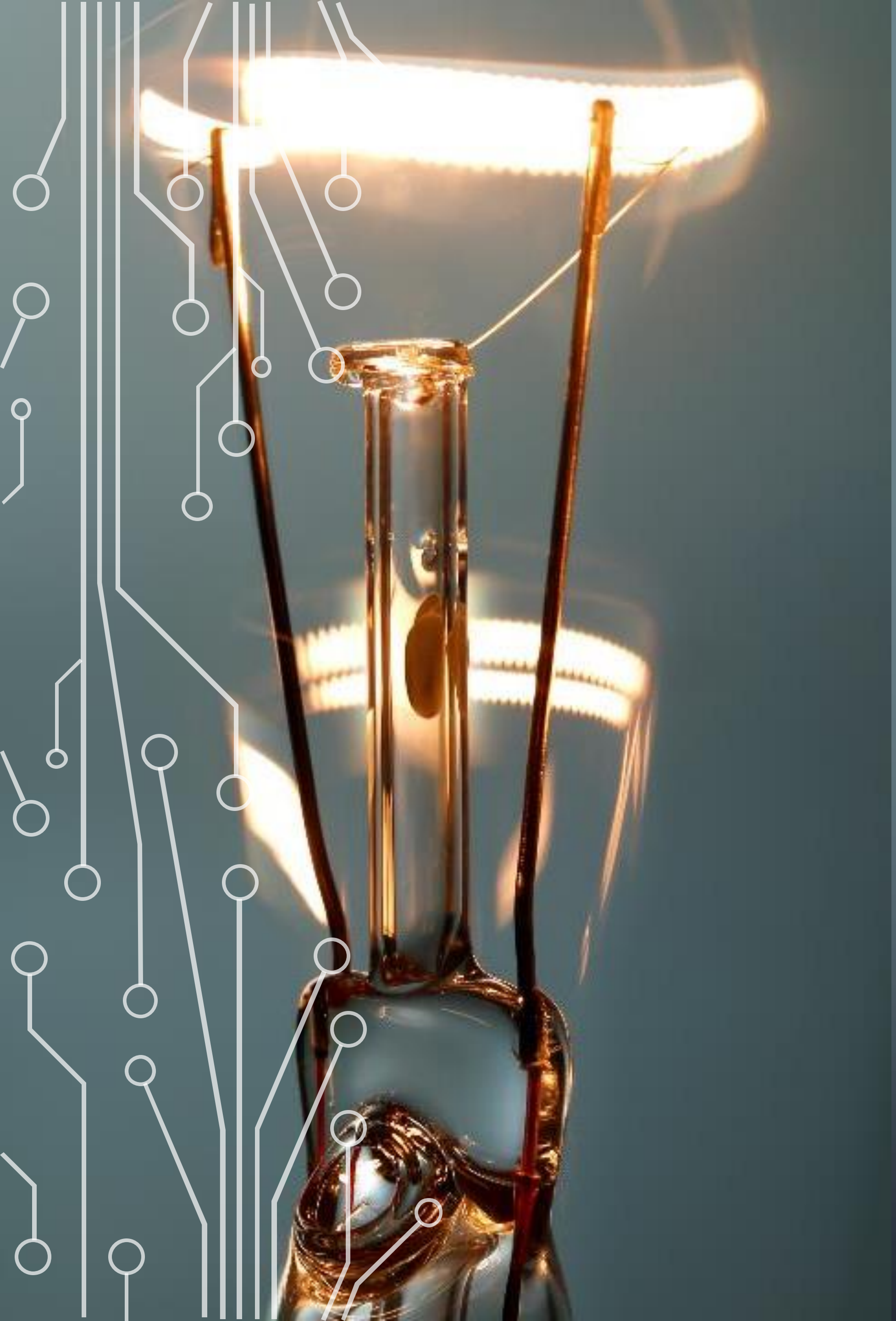


TIME SERIES ANALYSIS OF HOUSEHOLD ELECTRIC CONSUMPTION

HAIM ELBAZ

ASAF LEVI

AGENDA

- Article Introduction
- Data Description
- Article Workflow
- Methods
- Preprocessing & Data Preparation
- ARMA/ARIMA Modeling
- Model & Prediction Assessment

ARTICLE INTRODUCTION

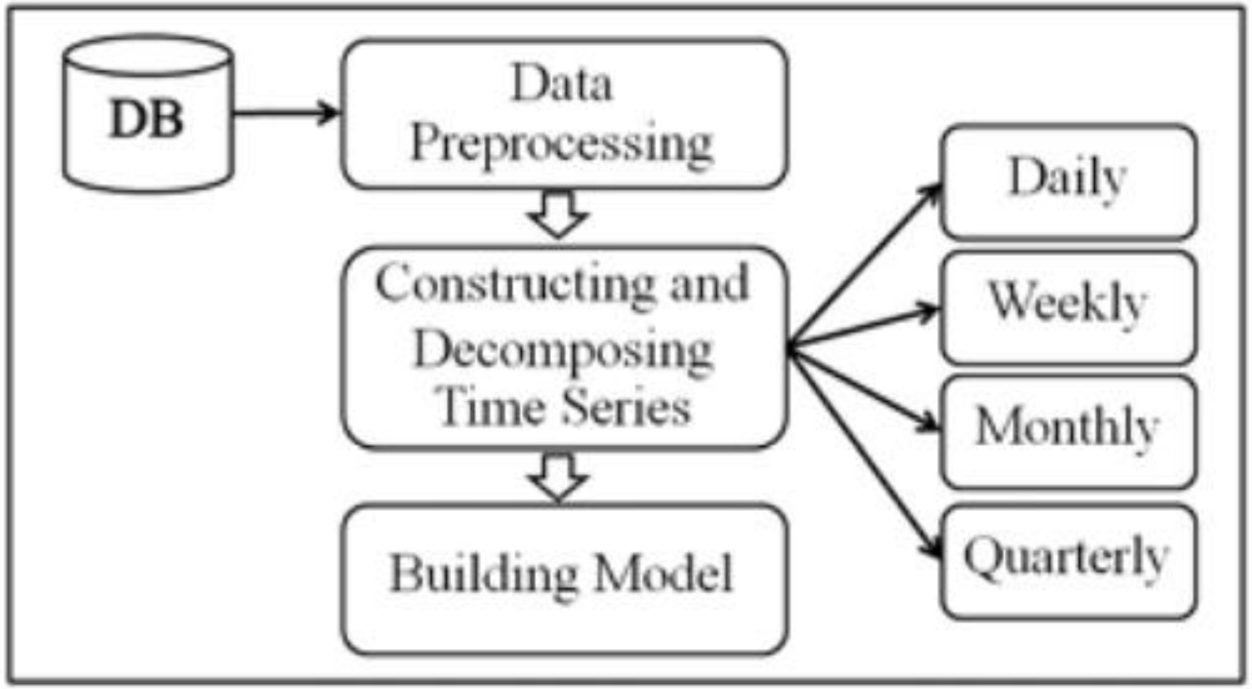
- *"Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models"*, Chujai et al. 2013
- Presented and published in "*Proceedings of the international multiconference of engineers and computer scientists*" Vol. 1. Hong Kong: IAENG, 2013
- The main target of the research was to fit the best ARIMA models for the daily, weekly, monthly and quarterly time scopes and assess their predictive power
- The author's prediction and calculation were originally conducted in R

DATA DESCRIPTION

- The dataset comprised of minute-averaged resolution data of a specific household power consumption over 47 months
- Data collected from a house in Sceaux, France between December 2006 and November 2010
- Contains several metrics, we will focus on the global active power, the minute-average of kW consumed over the entire household

ARTICLE WORKFLOW

- The authors decomposed the timeseries in 4 different frequencies
- Built a different ARIMA model for each of them
- Used R auto_arima package with default parameters to conduct a prediction
- Evaluate their models using RMSE criteria

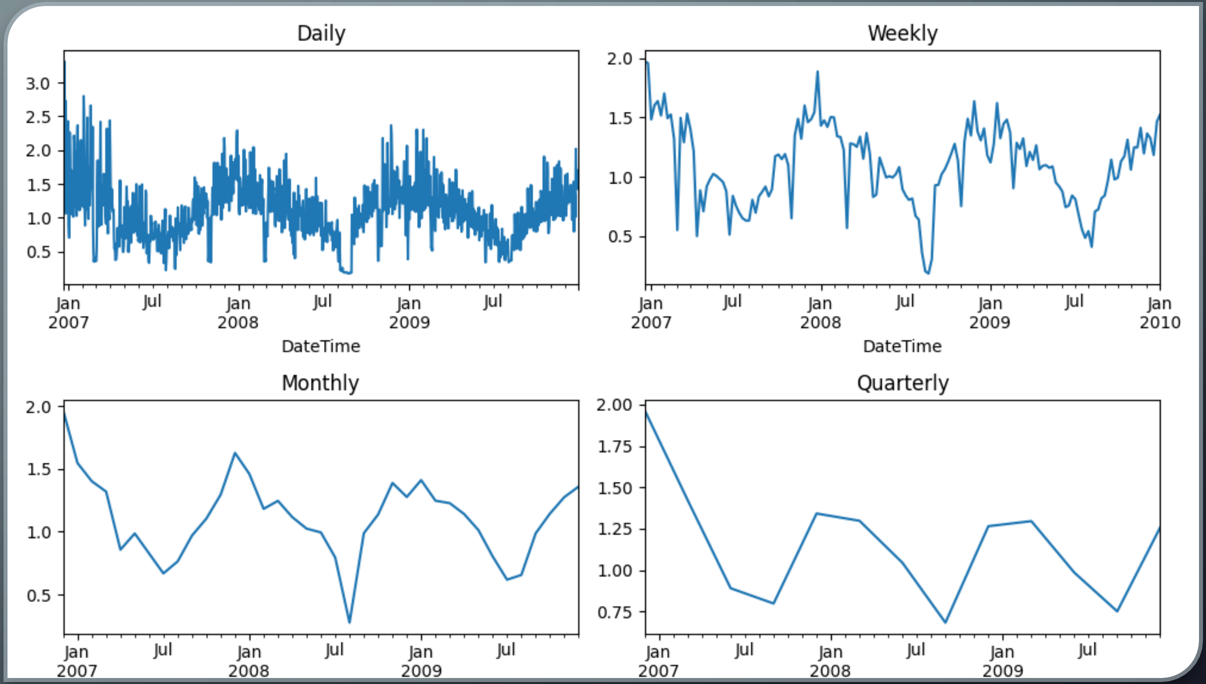


METHODS

- Aggregated the data over daily, weekly, monthly and quarterly periods to create 4 separate timeseries sets
- Used python's statsmodels to decompose the datasets
- Fitted ARIMA models
- Assessed our models fit using Ljunx-Box & Shapiro tests, QQ & histograms
- Evaluate the prediction by calculating RMSE against the last year of the timeseries

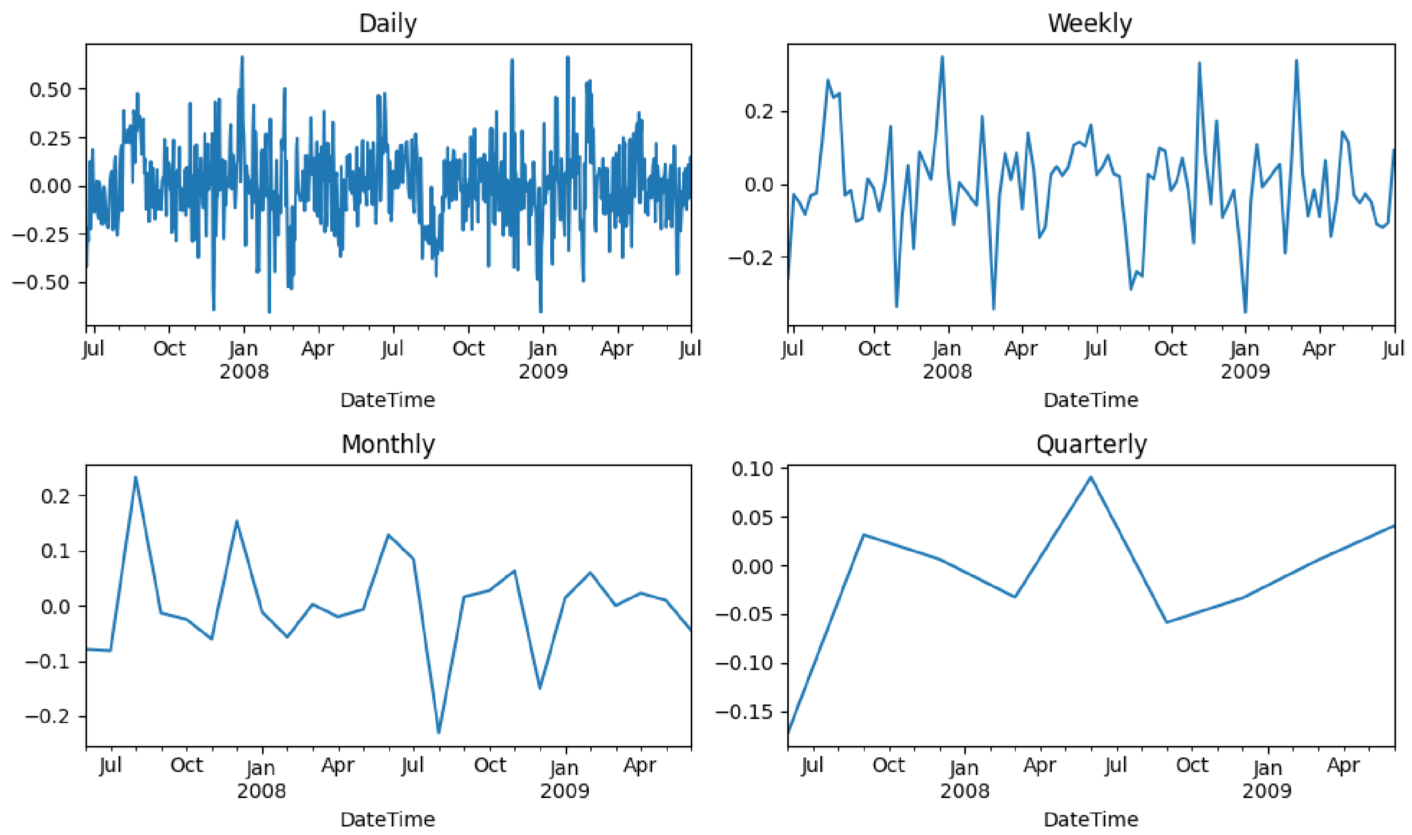
PREPROCESSING & DATA PREPARATION

- Raw data is a .txt file with semicolon separated values.
- "Date", "Time" strings evaluated and united into a single "datetime" index.
- "?" & "nan" values are forward-filled.
- Four time series were created – daily, weekly, monthly & quarterly by aggregating over these time scopes by mean, and changing timeseries frequencies to 365, 53, 12, 4 respectively which represented the changes in the "global_active_power" values across these different time periods.
- It is assumed that there is a ~year cycle and that forward-filling is a better prediction than mean or median.



DECOMPOSED DATA RESULTS

- Each series has been decomposed to remove any trend and seasonality that were in the data
- We then receive the following residuals series which you can see on the right
- The next step was to define if each series was indeed stationarity or not
- We use the ADF test to test this question
- The test results show that each series is stationarity



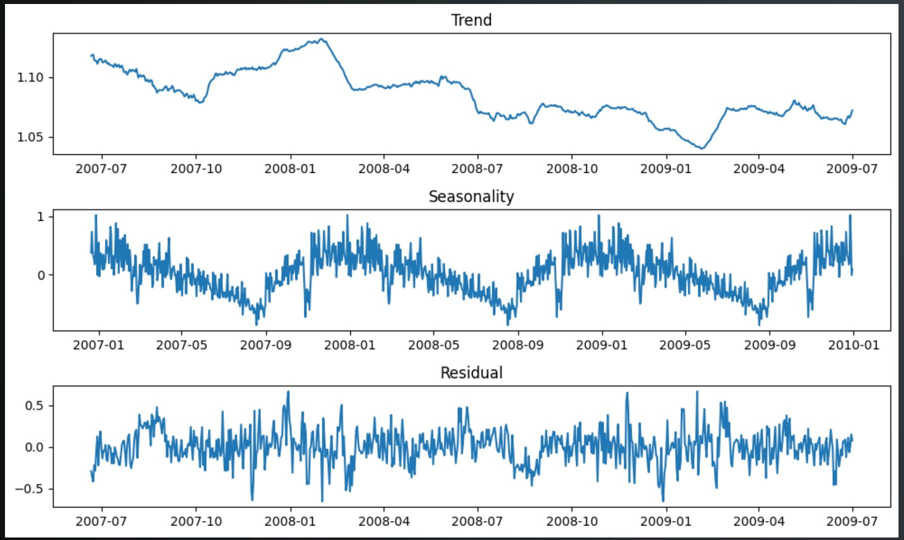

```
   Augmented Dickey-Fuller Results
=====================================
Test Statistic                 -5.177
P-value                         0.000
Lags                               18
-------------------------------------
```

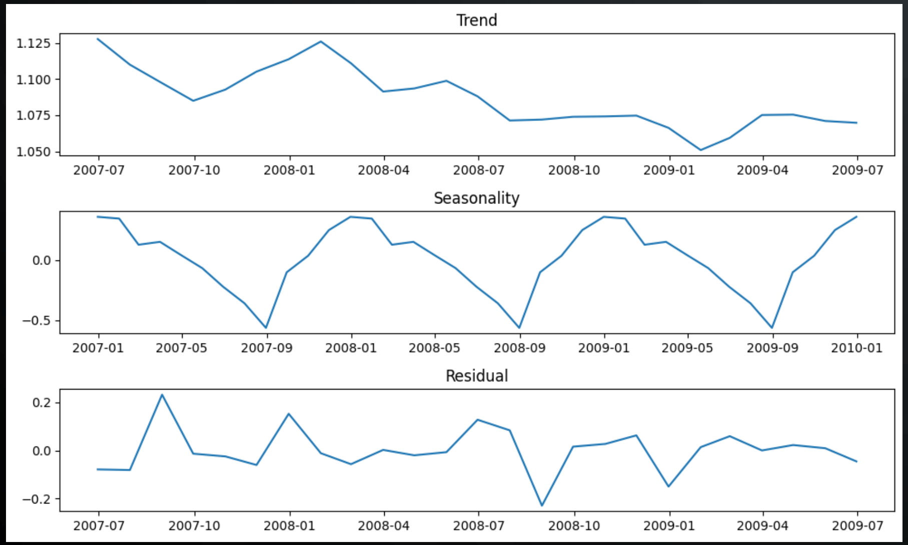
```
   Augmented Dickey-Fuller Results
=====================================
Test Statistic                 -5.095
P-value                         0.000
Lags                               11
-------------------------------------
```

```
   Augmented Dickey-Fuller Results
=====================================
Test Statistic                 -5.572
P-value                         0.000
Lags                                2
-------------------------------------
```

```
   Augmented Dickey-Fuller Results
=====================================
Test Statistic                 -5.572
P-value                         0.000
Lags                                2
-------------------------------------
```

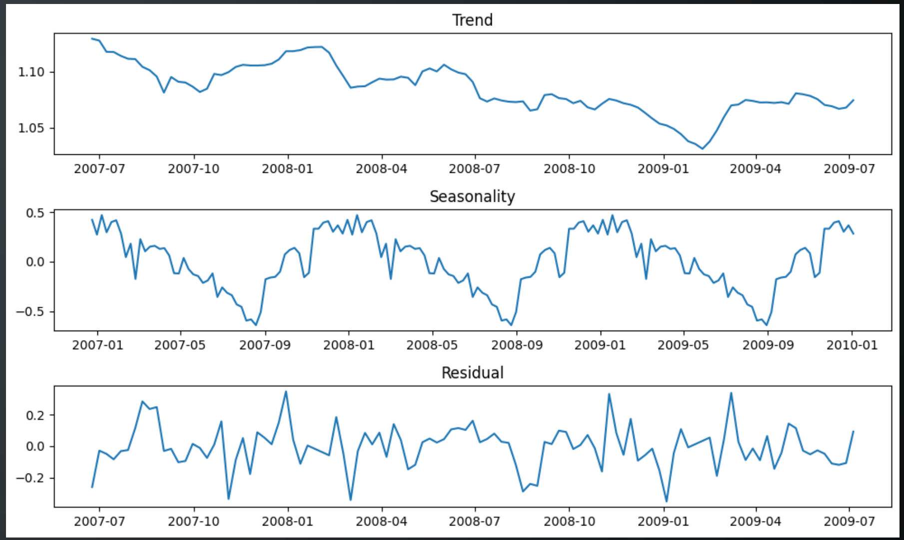
Daily Series Decomposition

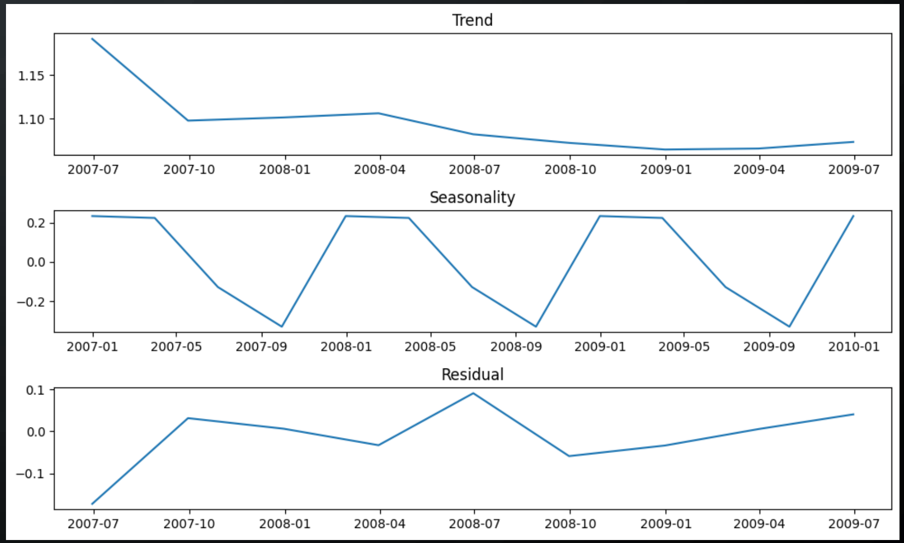




Monthly Series Decomposition

Weekly Series Decomposition





Quarterly Series Decomposition

FINDING THE BEST MODELS PARAMETERS

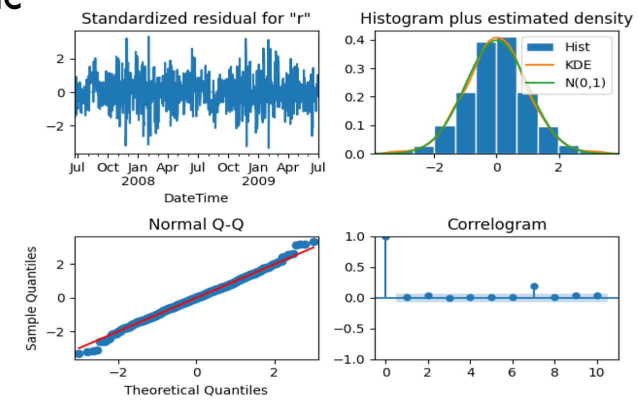
- We conducted a search for the set of p, d, q values that fit best to our training set
- We've limited the search of the values to be up until 3 for simplicity and running time efficiency.
- We used the AIC criteria to find the most suitable values
- We chose this criteria since it was also in use on the original research by the researchers.

```
Best daily ARIMA parameters & AIC: ((2, 0, 3), -386.12798117913235)
Best weekly ARIMA parameters & AIC: ((3, 0, 2), -143.4850205607858)
Best monthly ARIMA parameters & AIC: ((3, 0, 0), -51.477503096536445)
Best quarterly ARIMA parameters & AIC: ((0, 0, 2), -20.718878303845866)
```

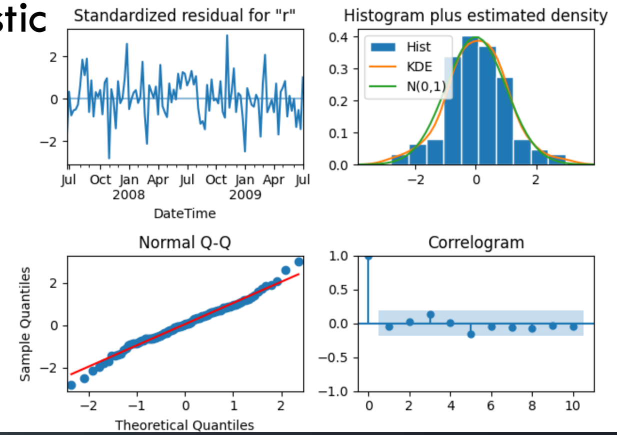
MODEL DIAGNOSTICS

- We ran a diagnostic check for each of the models to identify if each one of them is resulting from a normal distribution
- We ran the diagnostics both visually and statistically (Shapiro test)
- Both ways showed that the data's models are a result of normal distribution

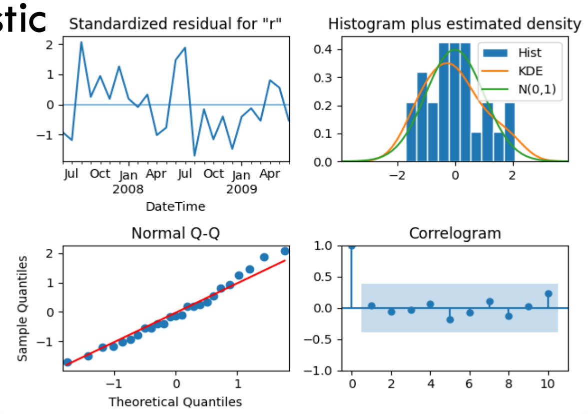
Daily Diagnostic



Weekly Diagnostic



Monthly Diagnostic



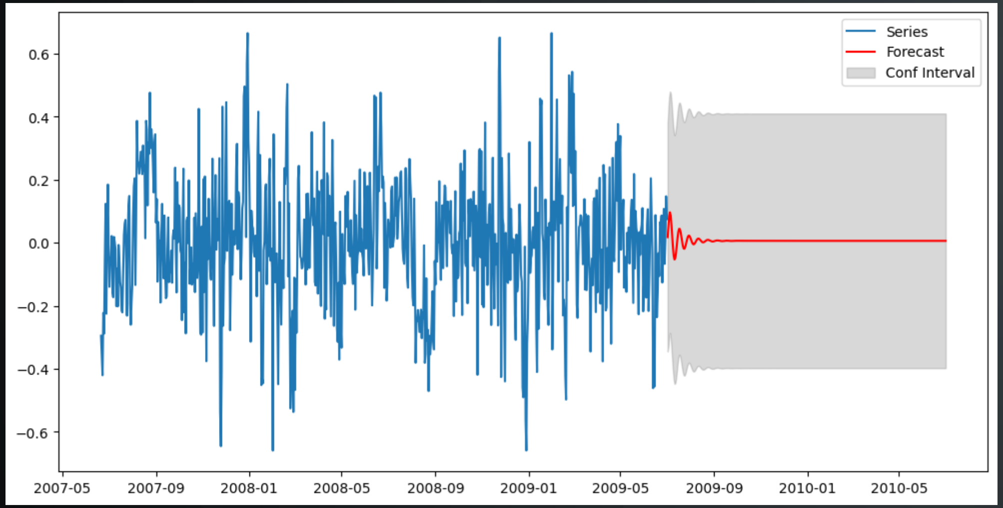
DIAGNOSTICS STATISTICAL TESTS RESULTS

Time Scope	Ljung-Box values*	Shapiro-Wilk values*
Daily	0.67	0.12
Weekly	0.66	0.71
Monthly	0.86	0.67
Quarterly	0.2	0.78

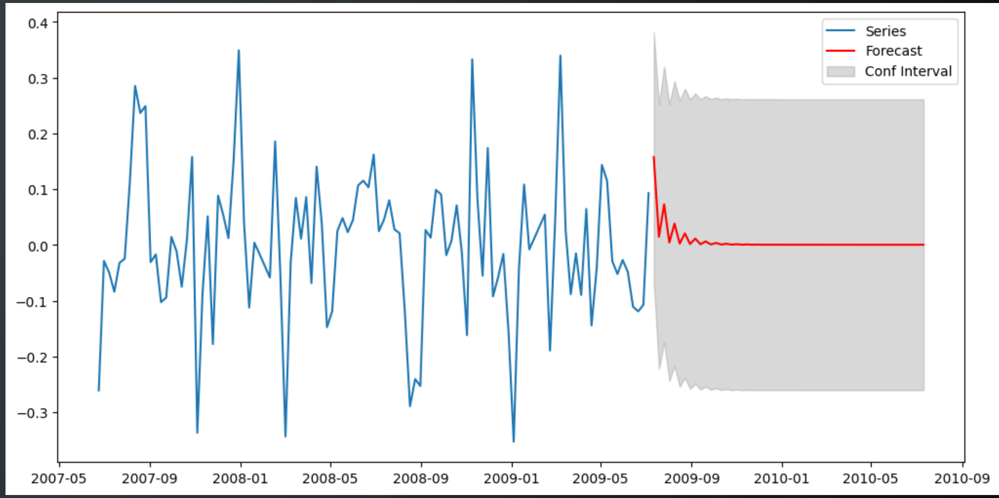
- **Ljung-Box:** H0 Not rejected – values IID
- **Shapiro-Wilk:** H0 Not rejected – values are normally distributed

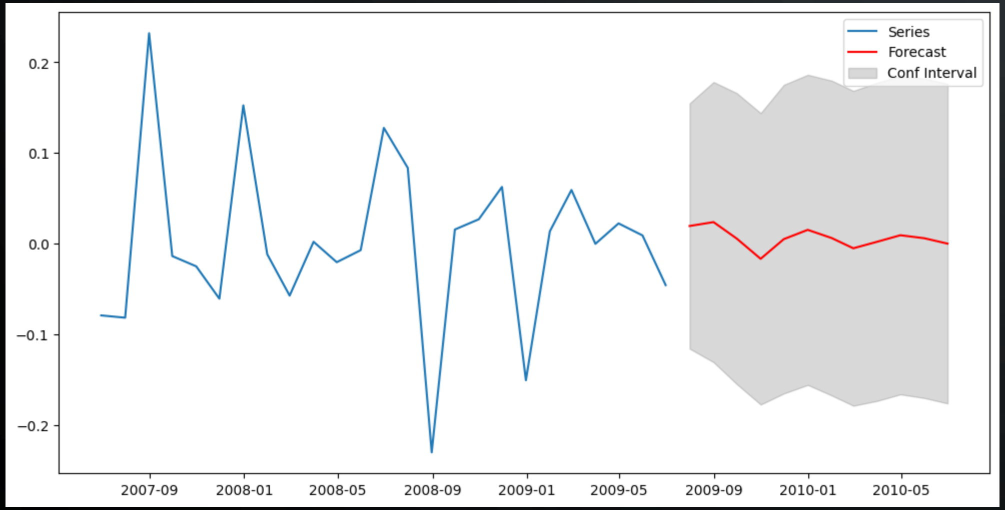
* p-value:0.05

Daily Series Prediction

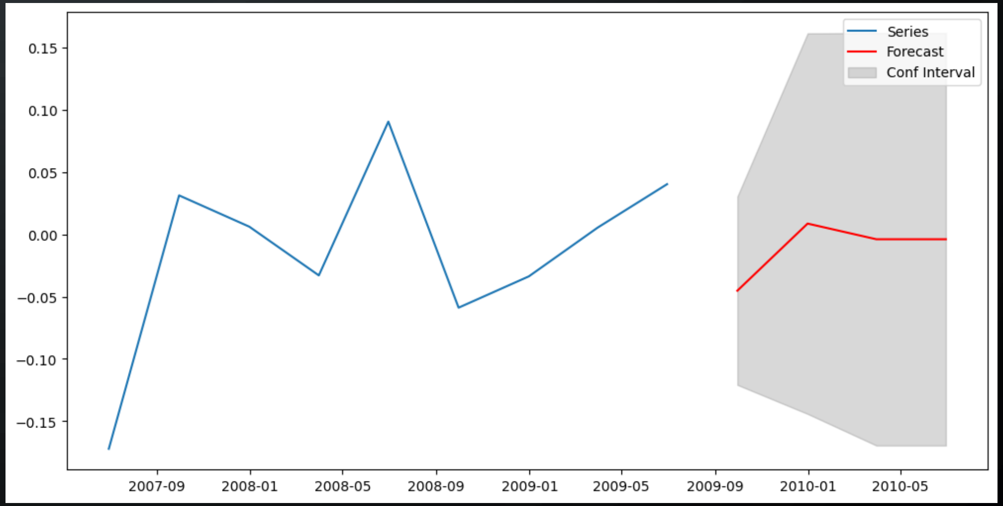


Weekly Series Prediction





Monthly Series Prediction



Quarterly Series Prediction

MODELS EVALUATION

Time Scope	Forecasting Period	Authors Models RMSE	Our Models RMSE
Daily	28	0.29	0.21
Weekly	20	0.18	0.17
Monthly	10	0.09	0.10
Quarterly	3	0.38	0.07

- **Authors Models:** Daily(3,1,3),Weekly(1,0,1),Monthly(0,0,0),Quarterly(0,0,0)
- **Our Models:** Daily(2,0,3),Weekly(3,0,2),Monthly(3,0,0),Quarterly(0,0,2)

SUMMARY

- Our models provided a suitable prediction with a lower RMSE rates.
- Based on our evaluation, we improved the authors model prediction by using a simpler ARMA models for the daily forecast than the one used in the research
- Monthly & Quarterly scopes don't have enough datapoints to provide a good prediction based on a reliable ARMA model which will not be an overfit to the data

THANK YOU!