POWER CONSUMPTION FORECASTING USING TIME SERIES ANALYSIS

CA1: REPORT

HIMANSH ARORA **D00233455**

ABSTRACT

This report aims to explore and model Power Consumption data forecasting through applying various Time Series Analysis techniques such as ARIMA/SARIMA and statistical methods such as PCA and Factor Analysis on a time series data of 3 Power Zones in Tetouan City, Morocco. The dataset consisted of 52,416 records including features such as temperature, humidity, wind speed, general diffuse flows and diffuse flows, and target variables are Power Consumption values of 3 different zones. The persistence model was first applied as a baseline, followed by ARIMA, which led to discovery of seasonality in Zone 3 and introduction to SARIMA. PCA was then applied to reduce dimensionality, address multicollinearity and improve model performance. The final ARIMA models demonstrated substantial improvement with very low RMSE values. Future research could explore more complex machine learning approaches such as LSTM networks.

INTRODUCTION

1. Background

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time [1]. It helps in identifying underlying trends and patterns within a given data, allowing in-depth knowledge about how and why specific things change over time. These benefits of Time Series Analysis can be harnessed together with Predictive Modelling by using historical data, and the resulting models can help in predicting future trends and values about that particular data, helping institutions, businesses, etc to make calculated decisions and plan their next steps accordingly [4]. Predictive Modelling can lead to benefit mankind in a number of ways such as improved decision-making, enhanced efficiency, cost reduction, and better resource allocation, etc. One of the many widely used real-world applications of Predictive Modelling using Time Series Analysis is the prediction of Power Consumption trends, which is focused on during this study.

Power Consumption forecasting is quite crucial to ensure a reliable and efficient energy supply, enable decision-making for infrastructure investments, enhance resource management, and reduce potential shortages or overproduction of power [6]. There are several models that can be applied in order to achieve this, such as ARIMA, SARIMA, Factor Analysis, PCA, etc.

2. Research Question

This study investigates the effectiveness of time-series analysis methods such as ARIMA, SARIMA, PCA, Factor Analysis and LSTM in forecasting power consumption in Tetouan City, with a focus on accuracy and model performance.

3. Literature Review

The growth of the human population and the development and application of technology has resulted in a rapid increase in electrical energy consumption. As a result, predicting electrical energy consumption is required when making electrical energy management decisions [3].

The papers chosen for this review focus on time-series models for predicting energy demands and avoiding oversupply or shortage of energy sources. They focus on identifying monthly energy consumptions using different time-series models that perform better on temporal data. The paper titled “Forecasting Electricity Consumption Using Time Series Model” employs 6 time series models, namely Simple Moving Average (SMA), Weighted Moving Average (WMA), Simple Exponential Smoothing (SES), Holt Linear Trend (HLT), Holt-Winters (HW), and Centered Moving Average (CMA), and compares the results using MSE, RMSE, MAPE, and MAE metrics, with HW being the most accurate and reliable model for seasonal data energy predictions [5]. The methodologies used in this paper are focused on both past and present values. Weights are assigned to past values, and trend analysis I performed using SMA, WMA, HLT, and so on, while stationary series are explored through methods like SMA that assign values to present data points. The dataset was obtained over 7 years from Universiti Tun Hussein Onn Malaysia (UTHM) to predict the energy needs that might vary due to factors like infrastructure. It helps in cost optimization and reliable energy forecasting for UTHM. However, the requirement of a more comprehensive and bigger dataset than the one used in this study, which only spans 7 years, is pointed out by the study to increase the accuracy of forecasting.

The other study that forms the main focus of this literature review and proves essential to the project is titled “Model for Predicting Electrical Energy Consumption Using ARIMA Method” [3]. This study is conducted by Muhammad Ridwan Fathin, Yudi Widhiyasana, and Nurjannah Syakrani for predicting energy requirements for medium (tactical decision making) and long-term (strategic decision making) consumptions. The model used in this study is the Auto-Regressive Integrated Moving Average (ARIMA), which is a combination of Auto-Regressive and Moving Average. The authors explore methods like Stationary Testing using methods like the Augmented Dickey-Fuller (ADF) test, Estimating AR and MA values, and selecting the best model by comparing RMSE for different ARIMA variations. The findings state that ARIMA (8,2,0) was good for medium-term predictions of electricity consumption. The data quality of this study is enhanced by a span of a decade. However, it points out challenges for long-term forecasting and points out the need for additional variables and hybrid models. It also opens research possibilities of the SARIMA model for seasonal data.

Both studies are compared with existing models that focus on optimizing smoothing parameters, double exponential smoothing, and the application of time-series models for problems like forecasting Malaysia’s population, predicting water levels, and half-hourly load demand. These comparisons point out the best time-series models for different methodologies and applications and offer insight into hybrid approaches for designing the best model for electrical forecasting. A combination of ARIMA with exponential smoothing or variations of the SARIMA model to forecast quarterly electricity demands might also offer improved accuracy

4. Dataset

The dataset is taken from the UCI Irvine machine learning repository. The dataset is related to Power Consumption of Tetouan, a city situated in North Morocco. The data is Multivariate and consists of 6 features, and 3 target variables. Data cleaning is done through the Python Pandas library. The dataset contains no missing values, identified unique categorical values in each variable this helps how this data will be dealt with.

The dataset can be accessed at:
https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city [2].

METHODS

Dataset Overview

The dataset contains power consumption records from three distribution networks in Tetouan City. Key attributes include timestamps, energy usage, and environmental factors.

Variables

Table 1: Variable Description

Variable Name	Role	Type	Description	Missing Values
DateTime	Feature	Date	Each ten minutes	no
Temperature	Feature	Continuous	Weather Temperature of Tetouan city	no

Humidity	Feature	Continuous	Weather Humidity of Tetouan city	no
Wind Speed	Feature	Continuous	Wind speed of Tetouan city	no
general diffuse flows	Feature	Continuous	general diffuse flows	no
diffuse flows	Feature	Continuous	diffuse flows	no
Zone 1 Power Consumption	Target	Continuous	power consumption of zone 1 of Tetouan city	no
Zone 2 Power Consumption	Target	Continuous	power consumption of zone 2 of Tetouan city	no
Zone 3 Power Consumption	Target	Continuous	power consumption of zone 3 of Tetouan city	no

1. ARIMA (AutoRegressive Integrated Moving Average)

ARIMA is a very popular statistical model used in time-series analysis. It is a model used to analyze and forecast univariate time series data. ARIMA identifies and captures trends and autocorrelation patterns in the given data through three components: AR (AutoRegressive), I (Integrated) and MA (Moving Average) [7]. An ARIMA model is denoted as ARIMA(p,d,q) where:

- **p**: Order of autoregressive terms taken into consideration
- **d**: Order of differencing required to make time series stationary
- **q**: Order of moving average, which is the number of lagged forecast errors taken

Mathematical Expression: $\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$

where $\hat{y}_t$ is the value at time step t, μ is a constant, ϕ is a coefficient of the AR component, θ is a white noise error term.

Assumptions: ARIMA assumes that there is a linear relationship among the variables and that the underlying time series data should be stationary after applying differencing. It also assumes that the errors are white noise with zero mean and constant variance.

Applications: ARIMA models can be used for Stock Price predictions, sales forecasting, weather and climate forecasting, etc.

2. SARIMA (Seasonal AutoRegressive Integrated Moving Average)

The ARIMA models perform well in identifying and capturing non-seasonal trends in a time series data but they are not successful in capturing trends when the data shows seasonal patterns, which introduces us to SARIMA. Seasonal ARIMA is an extension to ARIMA models in which seasonal trends are present [8]. A SARIMA model is represented as SARIMA(p,d,q)(P,D,Q)m where:

- **p**: Order of autoregressive terms taken into consideration
- **d**: Order of differencing required to make time series stationary
- **q**: Order of moving average, which is the number of lagged forecast errors taken
- **P**: Number of Seasonal AR terms
- **D**: Number of Seasonal differences
- **Q**: Number of Seasonal MA terms
- **m**: Number of observations per year

Mathematical Expression: $\Phi(B^m)\phi(B)(1-B)^d(1-B^m)^D.yt=\Theta(B^m)\theta(B)\varepsilon t$

where $\phi(B)$ and $\theta(B)$ are the non-seasonal AR and MA terms, $\Phi(B^m)$ and $\Theta(B^m)$ are the seasonal ARand MA terms, $(1-B)^d$ is the non-seasonal differencing, $(1-B^m)^D$ is he seasonal differencing and εt is the white noise error term.

Assumptions: SARIMA requires non-seasonal stationarity via differencing d and seasonal stationarity via seasonal differencing D. Similar to ARIMA models, SARIMA also assumes linearity.

Applications: SARIMA models can be used for Energy Consumption Analysis, Weather forecasting, traffic pattern analysis, etc.

3. PCA (Principal Component Analysis)

PCA is a widely used statistical method which is used for dimensionality reduction and feature extraction. It transforms a large set of possibly correlated variables into a smaller set of uncorrelated variables which are known as Principal Components, while retaining maximum variance from the original data [9].

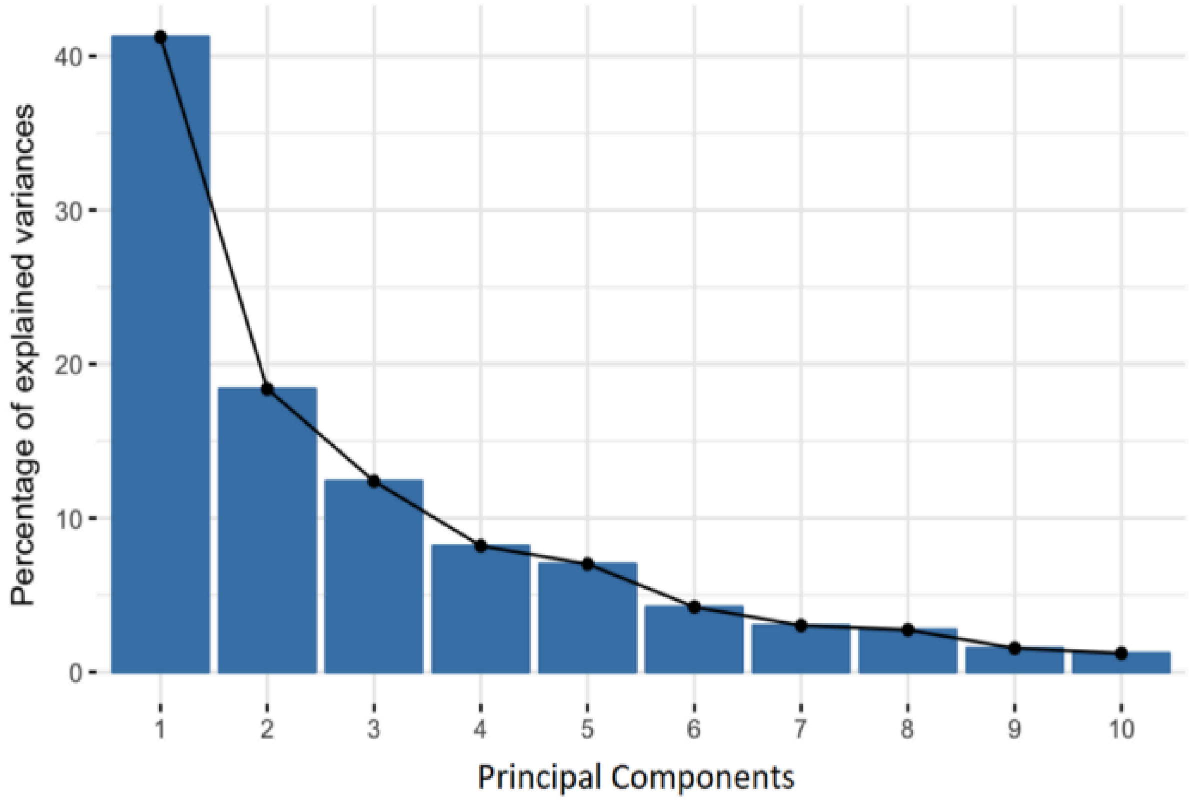


Fig. 1: PCA

Assumptions: PCA assumes that the data lies on a linear subspace and the resulting components are orthogonal and uncorrelated to each other. PCA also assumes that larger variance captures the most important structure.

Applications: PCA can be used for many processes which require dimensionality-reduction such as Noise filtering, Data compression, Preprocessing of data for Machine Learning, Visualization of high-dimensional data, etc.

4. Factor Analysis

Factor Analysis (FA) is a statistical technique which is used to describe variability among a set of observed, correlated variables in terms of a smaller number of unobserved variables called Factors. PCA on one hand focuses on maximizing the variance displayed by Principal Components, FA on the other hand focuses on identifying underlying concepts in the data.

Mathematical Expression: $yt = \alpha + B\xi t + \varepsilon t$

where α is an M-vector of intercept parameters, B is an M × k matrix parameter of factor loadings or simply loadings, and εt is a random M-vector of measurement errors, disturbances, and unique or idiosyncratic factors [10].

Assumptions: Factor Analysis assumes there is linearity present among the observed variables and that the factors are uncorrelated and normally distributed. It also assumes that the errors are uncorrelated with the Factors and with each other.

Applications: FA can be applied in Psychology to identify personality traits, Market research to identify consumer behaviour, and other social sciences and education fields.

5. LSTM (Long Short-Term Memory)

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) design which is created to identify and capture long-term dependencies in a time series data. LSTM involves a unique memory cell structure and gating mechanisms which makes it more special than the standard RNN techniques, that suffer to retain information over long sequences due to vanishing gradients during the stage of backpropagation.
Each LSTM unit maintains a cell state and a hidden state. It uses three main gates: Input gate (i_t), Forget gate (f_t), and Output gate (o_t) [11].

Mathematical Expressions:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

$$\begin{aligned}
i_t &\rightarrow represents\ input\ gate.\\
f_t &\rightarrow represents\ forget\ gate.\\
o_t &\rightarrow represents\ output\ gate.\\
\sigma &\rightarrow represents\ sigmoid\ function.\\
w_x &\rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.\\
h_{t-1} &\rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).\\
x_t &\rightarrow input\ at\ current\ timestamp.\\
b_x &\rightarrow biases\ for\ the\ respective\ gates(x).
\end{aligned}$$

Assumptions: LSTM assumes there is no stationarity required unlike ARIMA/SARIMA models. And LSTM networks also require large datasets to generalize well.

Applications: LSTM networks can be used for stock predictions, weather forecasting, NLP methods like text generation or translation, Speech recognition, etc.

RESULTS

1. EDA (Exploratory Data Analysis)

The dataset consists of 52,416 values across 9 columns with no missing values. There is one DateTime index and 8 of the other variables are of float type.

Table 2: Dataset Description

Column	Non-Null Count	Dtype
Temperature	52,416	float64
Humidity	52,416	float64
Wind Speed	52,416	float64
General Diffuse Flows	52,416	float64
Diffuse Flows	52,416	float64

Zone 1 Power Consumption	52,416	float64
Zone 2 Power Consumption	52,416	float64
Zone 3 Power Consumption	52,416	float64

DatetimeIndex: 52,416 entries, from 2017-01-01 00:00:00 to 2017-12-30 23:50:00

Dataset Statistics

Table 3: Dataset Statistics

Statistic	Temperature	Humidity	Wind Speed	General Diffuse Flows	Diffuse Flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
Count	52,416	52,416	52,416	52,416	52,416	52,416	52,416	52,416
Mean	18.81	68.26	1.96	182.70	75.03	32,344.97	21,042.51	17,835.41
Standard Deviation	5.82	15.55	2.35	264.40	124.21	7,130.56	5,201.47	6,622.17
Min	3.25	11.34	0.05	0.004	0.011	13,895.70	8,560.08	5,935.17
25% (Q1)	14.41	58.31	0.08	0.062	0.122	26,310.67	16,980.77	13,129.33
50% (Median, Q2)	18.78	69.86	0.09	5.04	4.46	32,265.92	20,823.17	16,415.12
75% (Q3)	22.89	81.40	4.92	319.60	101.00	37,309.02	24,713.72	21,624.10
Max	40.01	94.80	6.48	1,163.0	936.0	52,204.40	37,408.86	47,598.33

Correlation Matrix

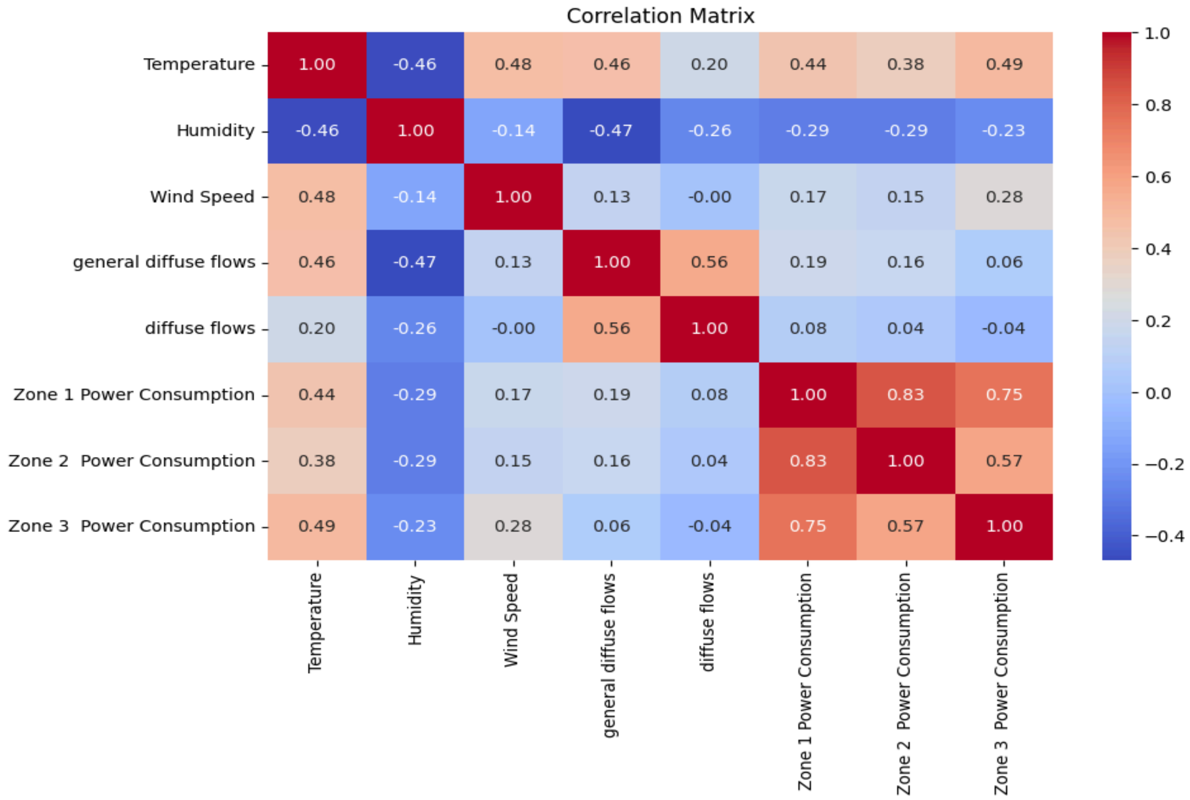


Fig. 2: Correlation Matrix

Detecting Outliers

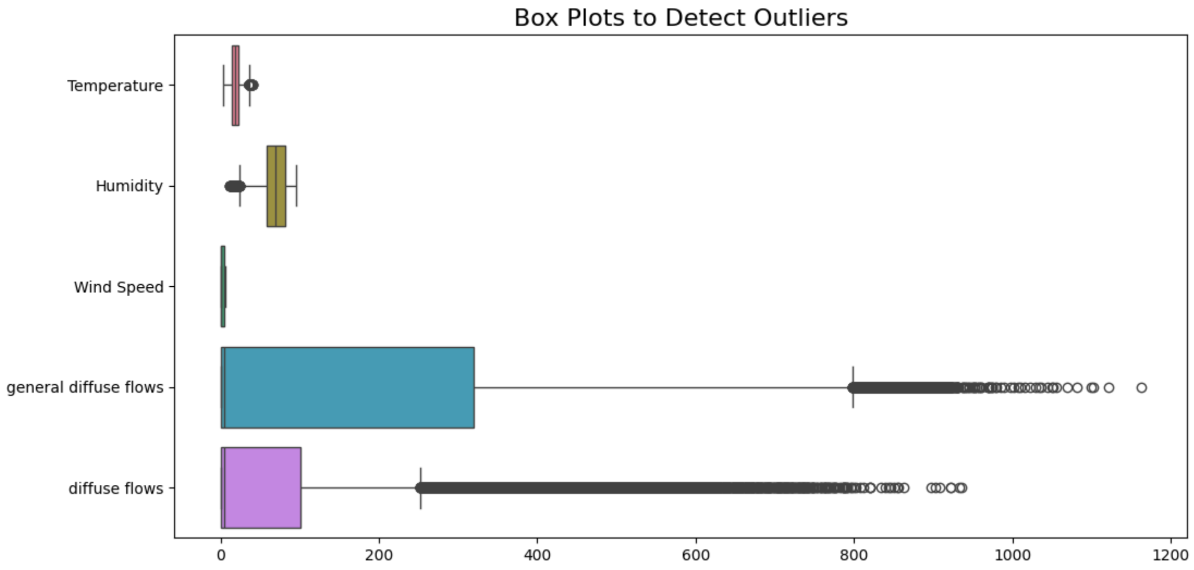


Fig. 3: Boxplots to Detect Outliers

2. Time Series Plots

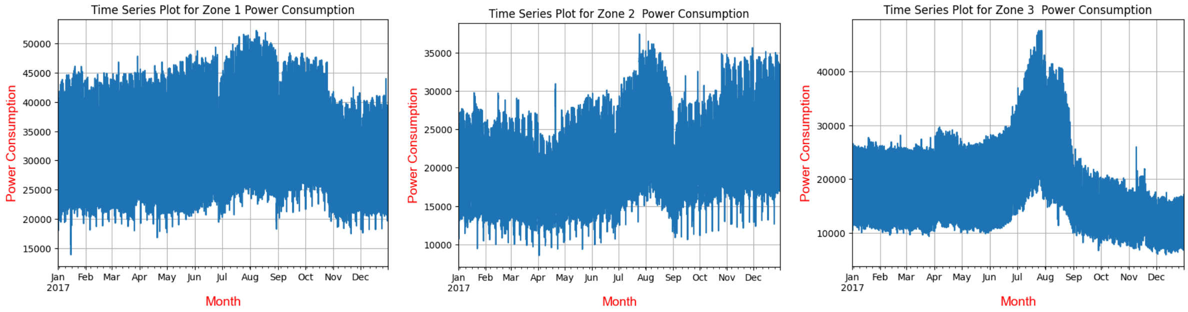


Fig. 4: Time Series Plots

3. Density Plots

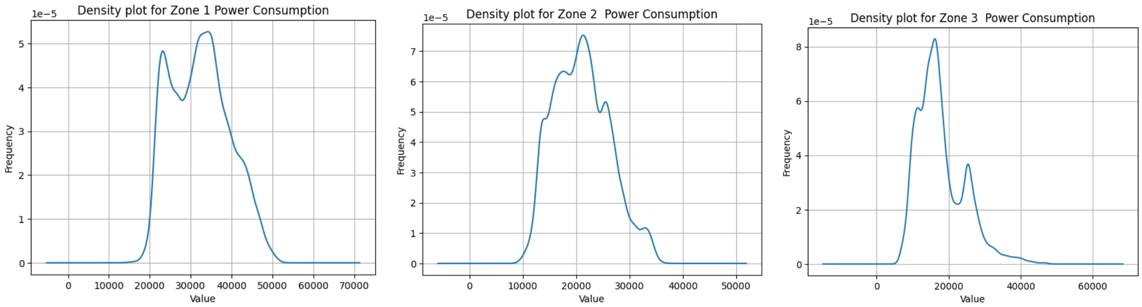


Fig. 5: Density Plots

4. Autocorrelation Plots

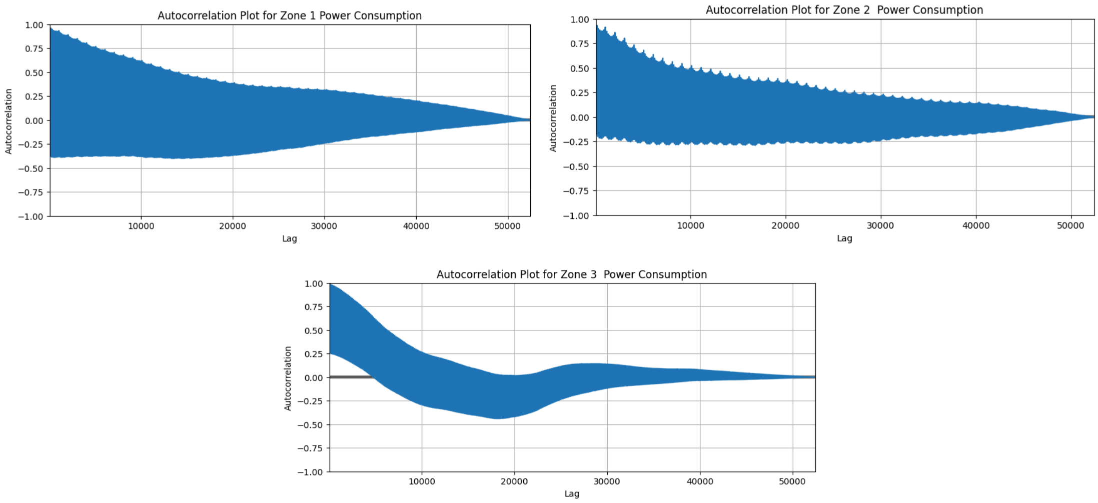


Fig. 6: Autocorrelation Plots

5. ACF Plots

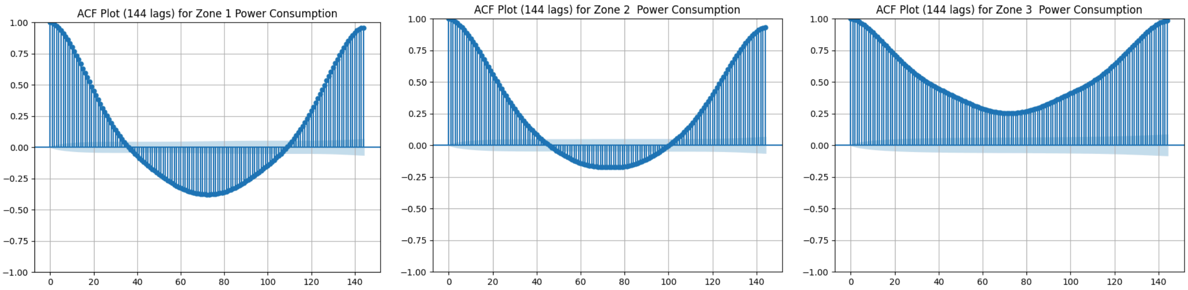


Fig. 7: ACF Plots (Daily)

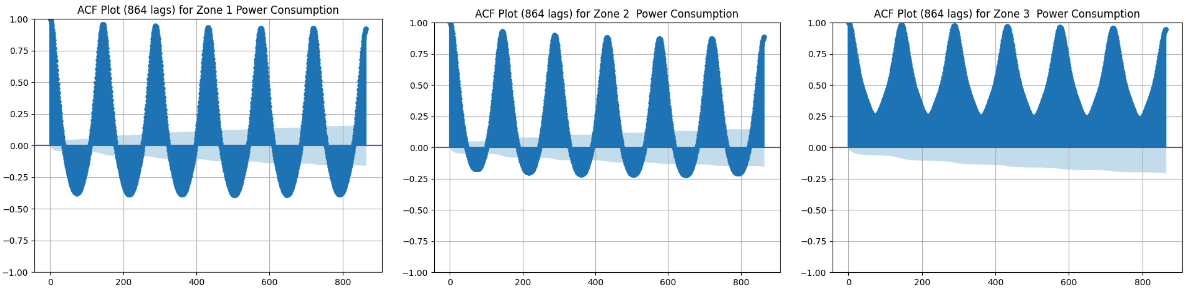


Fig. 8: ACF Plots (Weekly)

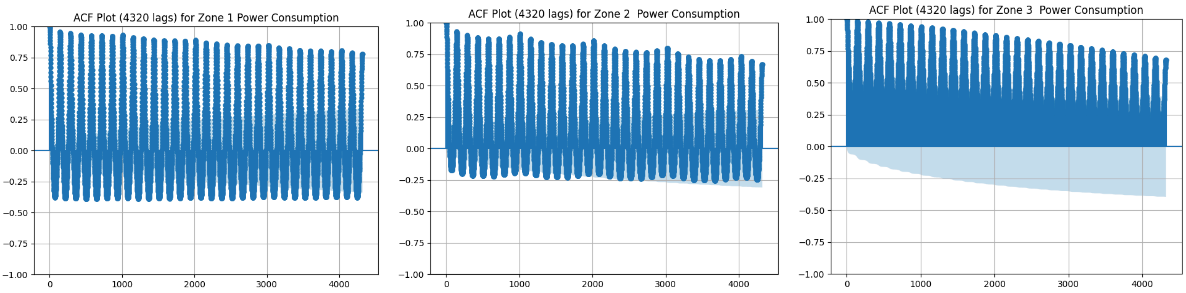


Fig. 9: ACF Plots (Monthly)

6. Persistence Model

```
--- Persistence Model for: Zone 1 Power Consumption ---
Test RMSE (Persistence Model): 733.131

--- Persistence Model for: Zone 2  Power Consumption ---
Test RMSE (Persistence Model): 576.081

--- Persistence Model for: Zone 3  Power Consumption ---
Test RMSE (Persistence Model): 327.909
```

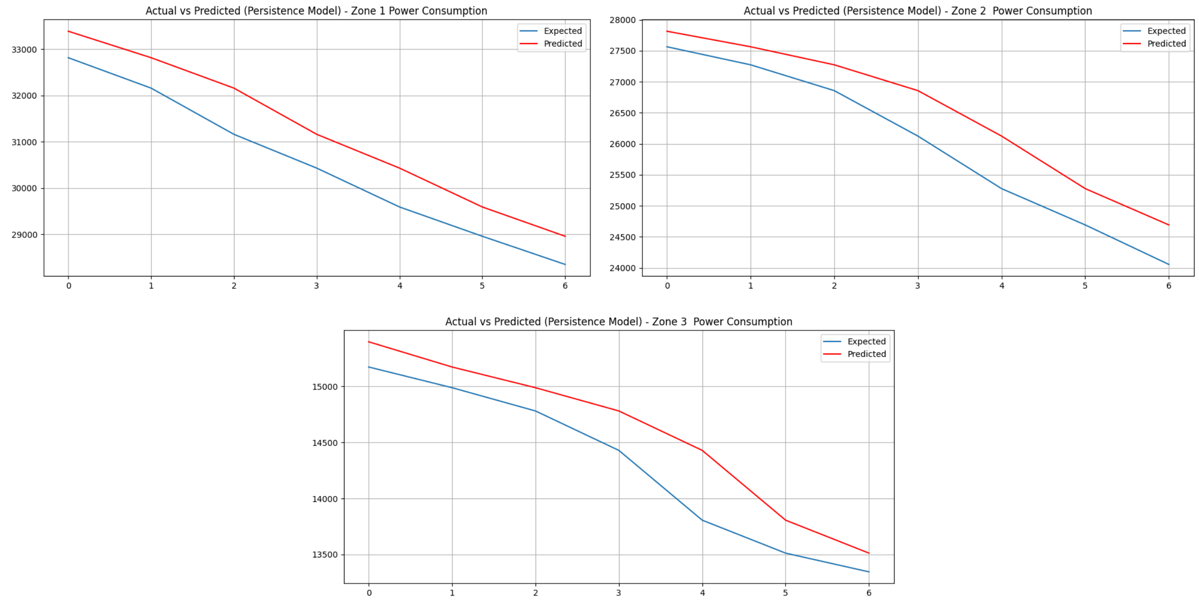


Fig. 10: Actual vs Predicted Values (Persistence Model)

7. ARIMA Model

Best model: ARIMA(1,1,1)(0,0,0)[12]
Total fit time: 9.612 seconds
Optimal ARIMA order for Zone 1 Power Consumption: (1, 1, 1)
RMSE for ARIMA Model (Zone 1 Power Consumption): 461.594

Best model: ARIMA(1,1,1)(0,0,0)[12]
Total fit time: 5.021 seconds
Optimal ARIMA order for Zone 2 Power Consumption: (1, 1, 1)
RMSE for ARIMA Model (Zone 2 Power Consumption): 286.000

Best model: ARIMA(0,0,1)(2,0,2)[12] intercept
Total fit time: 25.698 seconds
Optimal ARIMA order for Zone 3 Power Consumption: (0, 0, 1)
RMSE for ARIMA Model (Zone 3 Power Consumption): 2327.934

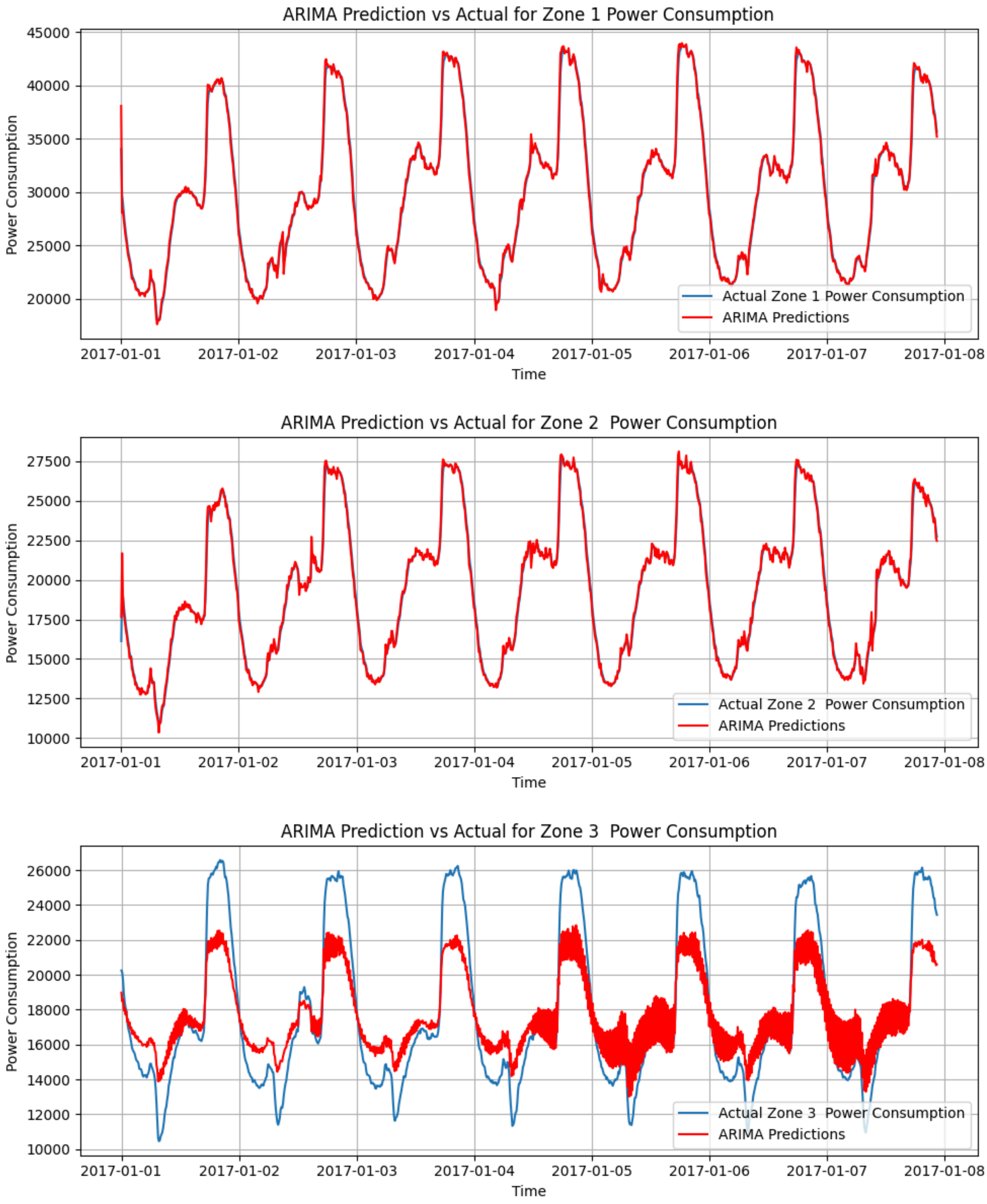


Fig. 11: Actual vs Predicted Values (ARIMA without PCA)

8. SARIMA Model

```
SARIMAX Results
==============================================================================
Dep. Variable:          Zone 3  Power Consumption          No. Observations:            1000
Model:            SARIMAX(0, 0, 1)x(2, 0, [1, 2], 12)       Log Likelihood          -10452.281
```

```
Date:                    Fri, 18 Apr 2025          AIC                  20916.562
Time:                            18:15:34          BIC                  20945.851
Sample:                        01-01-2017          HQIC                 20927.708
                             - 01-07-2017
Covariance Type:                      opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.9555      0.197      4.852      0.000       0.569       1.342
ar.S.L12       0.1754      1.350      0.130      0.897      -2.470       2.821
ar.S.L24       0.8439      1.154      0.732      0.464      -1.417       3.105
ma.S.L12       0.5155      2.003      0.257      0.797      -3.409       4.440
ma.S.L24      -0.2590      1.570     -0.165      0.869      -3.336       2.818
sigma2      3.279e+08   7.27e-08   4.51e+15      0.000    3.28e+08    3.28e+08
==============================================================================
Ljung-Box (L1) (Q):              218.92   Jarque-Bera (JB):              284.61
Prob(Q):                           0.00   Prob(JB):                        0.00
Heteroskedasticity (H):            0.50   Skew:                            0.10
Prob(H) (two-sided):               0.00   Kurtosis:                        5.64
==============================================================================
```

SARIMA RMSE for Zone 3: 928.4612

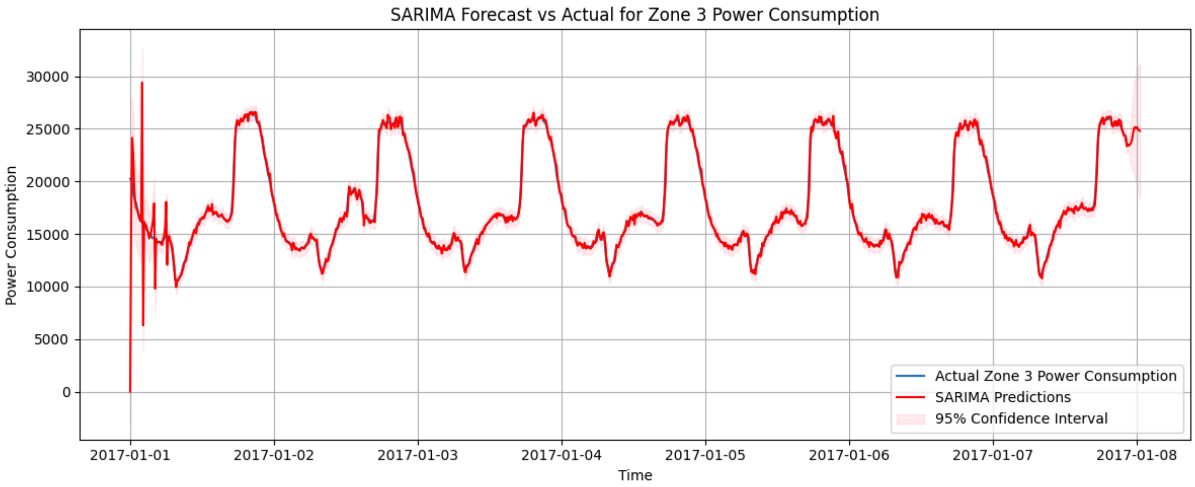


Fig. 12: Actual vs Predicted Values (SARIMA Model for Zone 3)

9. PCA

Table 4: Explained Variance Ratio Table

Component	Explained Variance Ratio
1	0.883947

2	0.113477
3	0.002239
4	0.000291
5	0.000046

Number of important components: 2

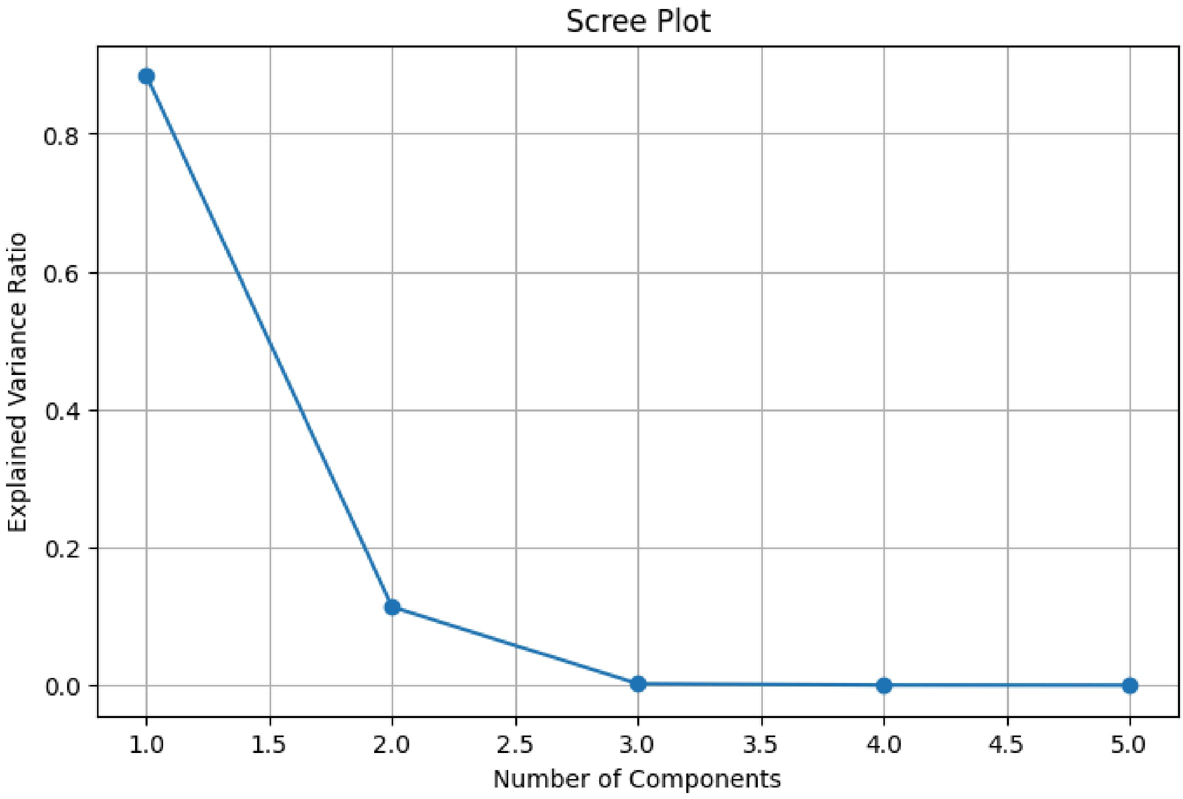


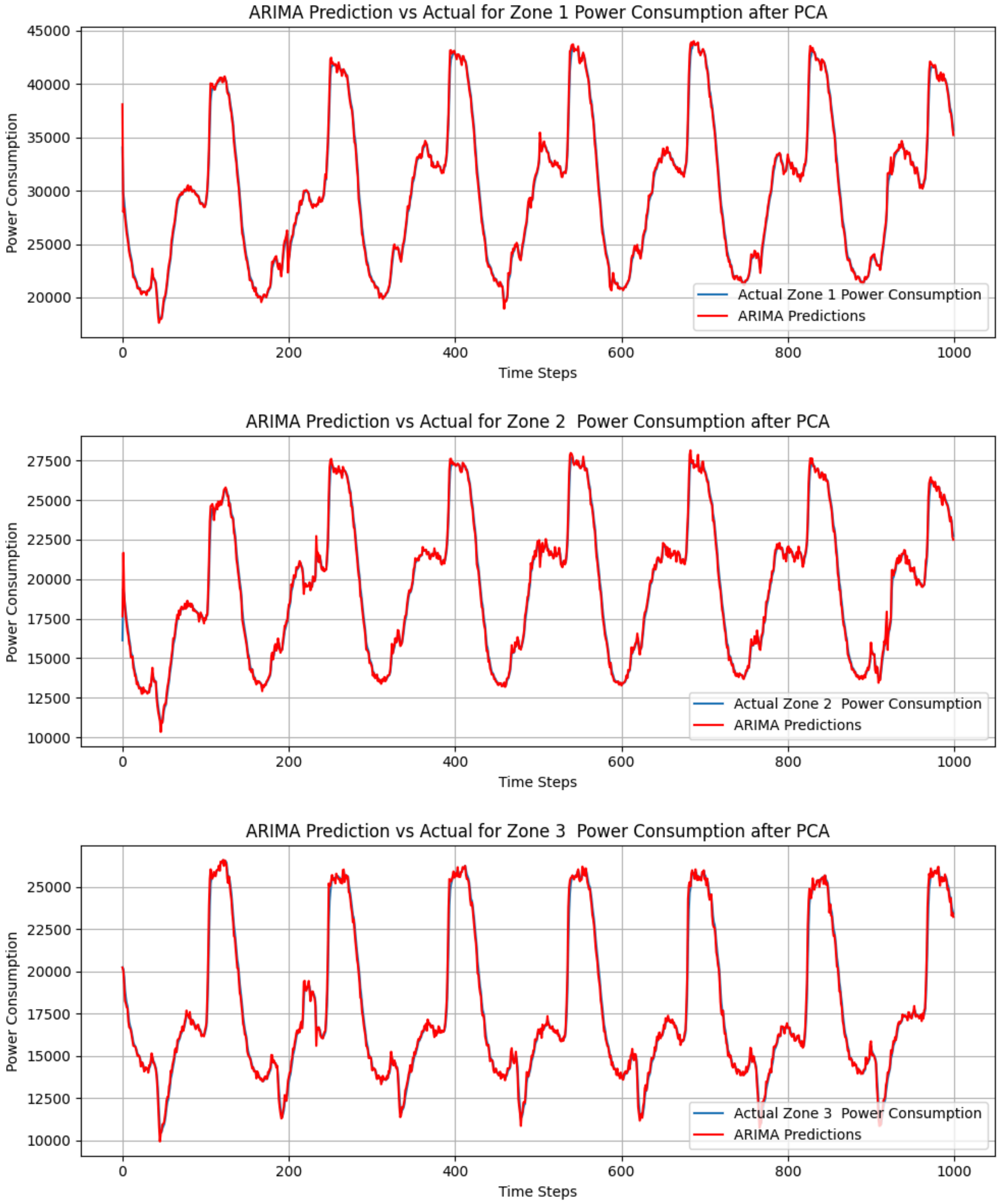
Fig. 13: Scree Plot for PCA

10. ARIMA Model after PCA

Best model: ARIMA(2,1,2)(0,0,0)[0] intercept
Total fit time: 2.015 seconds
Optimal ARIMA order for Zone 1 Power Consumption: (2, 1, 2)
RMSE for ARIMA Model (Zone 1 Power Consumption): 461.234

Best model: ARIMA(1,1,2)(0,0,0)[0]
Total fit time: 2.724 seconds
Optimal ARIMA order for Zone 2 Power Consumption: (1, 1, 2)
RMSE for ARIMA Model (Zone 2 Power Consumption): 285.546

```
Best model:  ARIMA(2,0,1)(0,0,0)[0] intercept
Total fit time: 1.371 seconds
Optimal ARIMA order for Zone 3  Power Consumption: (2, 0, 1)
RMSE for ARIMA Model (Zone 3  Power Consumption): 248.554
```



ARIMA Prediction vs Actual for Zone 1 Power Consumption after PCA

ARIMA Prediction vs Actual for Zone 2 Power Consumption after PCA

ARIMA Prediction vs Actual for Zone 3 Power Consumption after PCA

Fig. 14: Actual vs Predicted Values (ARIMA after PCA)

DISCUSSION

1. EDA (Exploratory Data Analysis)

The exploratory data analysis discovers the dataset in detail through density plots, correlation matrix, outlier detection plots, autocorrelation plots, and ACF plots on a daily, weekly, and monthly time period. Table 2 displays the dataset description, whereas Table 3 shows the dataset statistics for each variable in the time series.

Correlation Matrix(Fig. 2)**:** The correlation matrix is important in identifying the relationship between the 3 output variables, and the feature variables. The correlation matrix points relation between temperature and humidity and the 3 output variables. However, it is also important to check whether there is multicollinearity in the data by studying the correlation between other feature variables. It can be noted that temperature has high correlation with almost every other feature variable. There is also high correlation between Humidity and diffuse flows, and general diffuse flows and diffuse flows. This high multicollinearity can be problematic for various machine learning models because of the risk of overfitting and learning noise. It can be dealt with through the PCA application. More will be discussed about this.

Box Plots(Fig. 3)**:** There are a high number of outliers observed in General Diffuse Flows, and Diffuse Flows. There is potential for improvement in the plots by scaling them to the same range.

2. Time Series Plots (Fig. 4)

The Time Series line plots for each Power Consumption Zone show us how the Power Consumption changes over time for different zones. The x-axis consists of the time showcasing each month and the y-axis contains the power consumption values in KWh.

The plots help us identify if there is any seasonality present in the power consumption values over the year.

3. Density Plots (Fig. 5)

The Density plots for each Power Consumption Zone show us where the power consumption values are concentrated over the given time period of a year.

The density plot for Zone 2 Power consumption shows a normal distribution. However, the distribution of Zone 1 and Zone 3 Power Consumption shows 2 peaks. Zone 3 also appears to be slightly left skewed due to potential outlier values present in the dataset.

4. **Autocorrelation Plots** (Fig. 6)

The autocorrelation plots show the correlation between a variable's current value and its lagged versions. The x-axis represents lag, while the y-axis represents the autocorrelation at a particular lag (ranges from -1 to 1).

The autocorrelation plots tell us how strongly the time series depends on its past values and how quickly the autocorrelation drops to zero. A slow decay is observed in the autocorrelation which hints towards long memory.

5. **ACF Plots** (Fig. 7,8,9)

There are 3 variations of ACF plots generated for each Power Consumption Zone, Daily trend (Fig. 7), Weekly trend (Fig. 8) and Monthly trend (Fig. 9). The y-axis is the autocorrelation coefficient and x-axis is the number of lags, 144 lags for daily, 864 lags for weekly and 4320 lags for capturing monthly trend in a year long data of every 10 minutes (52,416 lags in total).

Autocorrelation spikes can be noticed clearly at multiples of 144 lags, suggesting a strong daily pattern.

6. **Persistence Model** (Fig. 10)

A persistence model assumes that the future value of a time series is calculated under the assumption that nothing changes between the current time and the forecast time [12]. The blue line describes the actual values and the red line shows the predicted values of the persistence model.

The prediction line varies by a significant amount of difference and with RMSE values of 733.131 for Zone 1, 576.081 for Zone 2, and 327.909 for Zone 3.

7. **ARIMA Model** (Fig. 11)

ARIMA models offer a range of advantages for forecasting time series, including the flexibility to capture various types of patterns and behaviors in the data, such as seasonality, cycles, or trends [13].

The ARIMA model works well for Zone 1 and Zone 2 Power Consumptions and captures most of their behaviour with RMSE values of 461.594 and 286.000 respectively. The prediction line very closely follows the actual values line. But for Zone 3, it fails to capture the behaviour with a very high RMSE value of 2327.934. This is because Zone 3 also shows seasonality, so it is better to apply SARIMA model for that zone.

8. **SARIMA Model** (Fig. 12)

The SARIMA model for zone 3 works much better in capturing the trend of the data and shows a considerable improvement in the RMSE value 928.4612. The SARIMA prediction line almost closely follows the actual values line but there still seems to be some uncaptured patterns present which the SARIMA model fails to capture since the RMSE value is still quite high.

9. PCA

Principal Component Analysis is applied to reduce dimensionality and identify underlying patterns in the series. Table 4 displays the explained variance ratio and there are 2 important Principal Components present. The Scree Plot for PCA (Fig. 13) also confirms that there are only 2 main Principal Components that explain the most variance, with a clearly noticeable 'elbow point' after PC2.

Since only a few components explain most of the variance, it indicated that most of the original variables are highly correlated. The 2 Principal Components recently discovered can be used instead of using all the original variables to further models to reduce complexity.

10. ARIMA Model after PCA (Fig. 14)

After applying PCA, the ARIMA models for each Power Consumption Zone show huge improvement. The RMSE values come down to 461.234 for Zone 1, 285.546 for Zone 2, 248.554 for Zone 3. The Actual vs Predicted ARIMA plots also display that the Predicted ARIMA line very closely follows the Actual values line for each zone.

CONCLUSION

This study aimed to explore and model Power Consumption data forecasting through applying various Time Series Analysis techniques such as ARIMA/SARIMA and statistical methods such as PCA and Factor Analysis on a time series data of 3 Power Zones in Tetouan City, Morocco. The dataset consisted of 52,416 entries including features such as temperature, humidity, wind speed, general diffuse flows and diffuse flows, and target variables were Power Consumption values of 3 different zones.

After applying EDA, several key insights were gained related to the characteristics of the data and suitability of the time series methods. The correlation matrix highlighted strong relationships between key variables, particularly between temperature and humidity, as well as among different flow measures. This multicollinearity was addressed through the application of Principal Component Analysis (PCA), which identified two principal components that explained the majority of the variance. This dimensionality reduction significantly improved the performance of the subsequent models by reducing the complexity of the input features and mitigating the risk of overfitting. The time series and autocorrelation plots revealed strong seasonal patterns, especially on a daily scale, and highlighted the importance of capturing these patterns in the modeling process.

The persistence model, which assumes no change from one time step to the next, was initially used as a baseline. The RMSE values for the persistence model were high across all zones, meaning that simply predicting the previous value was not an effective strategy in this case for forecasting power consumption.

After that, ARIMA models were fit to all the zones. For Zones 1 and 2, ARIMA models were found to perform well, capturing the underlying trends and producing reasonable RMSE values of 461.594 and 286.000, respectively. The ARIMA models' ability to fit the data with minimal error demonstrated their effectiveness for these zones, where the patterns were relatively straightforward. However, for Zone 3, the ARIMA model struggled due to the presence of significant seasonality, with an RMSE value of 2327.934. This led to the adoption of the SARIMA model, which is better suited for handling seasonal components. Although the SARIMA model significantly improved the RMSE for Zone 3, it still failed to fully capture all the seasonal behavior, resulting in an RMSE of 928.4612.

PCA was then applied to reduce dimensionality and improve model performance further. After incorporating the 2 important principal components, ARIMA models for all three zones showed substantial improvement. RMSE values decreased to 461.234 for Zone 1, 285.546 for Zone 2, and 248.554 for Zone 3, showcasing the impact of dimensionality reduction on prediction accuracy.

In conclusion, this analysis demonstrated the importance of selecting appropriate models based on the specific characteristics of the data. While ARIMA models performed well for Zones 1 and 2, the seasonality in Zone 3 required the use of SARIMA. The application of PCA further enhanced model performance by addressing multicollinearity and reducing complexity. This study underscores the value of combining exploratory data analysis, statistical modeling, and dimensionality reduction to produce effective forecasting models for time series data. Future work could explore the use of more advanced models, such as LSTM, to capture non-linear patterns and further improve prediction accuracy.

REFERENCES

1. Tableau. (2024). Time Series Analysis: Definition, Types, Techniques, and When It's Used. [online] Available at: https://www.tableau.com/analytics/what-is-time-series-analysis/
2. Salam, A. & El Hibaoui, A. (2018). Power Consumption of Tetouan City [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5B034.
3. Fathin, M.R., Widhiyasana, Y. and Syakrani, N. (2021). Model for Predicting Electrical Energy Consumption Using ARIMA Method. [online] www.atlantis-press.com. doi:https://doi.org/10.2991/aer.k.211106.047.
4. Preset.io. (2023). Available at: https://preset.io/blog/time-series-forecasting-a-complete-guide/
5. Lee, Y.W., Tay, K.G. and Choy, Y.Y. (2018). Forecasting Electricity Consumption Using Time Series Model. International Journal of Engineering & Technology, 7(4.30), p.218. doi:https://doi.org/10.14419/ijet.v7i4.30.22124.

6. Saylor Academy. (2016). Forecasting Electricity Demand: Why are electricity-demand forecasts important? | Saylor Academy | Saylor Academy. [online] Available at: https://learn.saylor.org/mod/book/view.php?id=61397/
7. Ariton, L. (2021). A Thorough Introduction To ARIMA Models. [online] Medium. Available at: https://medium.com/analytics-vidhya/a-thorough-introduction-to-arima-models-987a24e9ff71.
8. Ritu Santra (2023). Introduction to SARIMA Model - Ritu Santra - Medium. [online] Medium. Available at: https://medium.com/%40ritusantra/introduction-to-sarima-model-cbb885ceabe8 [Accessed 16 Apr. 2025].
9. Built In. (2022). Principal Component Analysis (PCA) Explained | Built In. [online] Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis? [Accessed 16 Apr. 2025].
10. Gilbert, P.D. and Meijer, E. (2025). Time Series Factor Analysis with an Application to Measuring Money. the University of Groningen research portal. [online] doi:https://hdl.handle.net/11370/d7d4ea3d-af1d-487a-b9b6-c0816994ef5a.
11. Thakur, D. (2018). LSTM and its equations - Divyanshu Thakur - Medium. [online] Medium. Available at: https://medium.com/%40divyanshu132/lstm-and-its-equations-5ee9246d04af [Accessed 16 Apr. 2025].
12. Marius Paulescu, Paulescu, E. and Viorel Bădescu (2021). Nowcasting solar irradiance for effective solar power plants operation and smart grid management. Elsevier eBooks, pp.249–270. doi:https://doi.org/10.1016/b978-0-12-817772-3.00009-4.
13. www.linkedin.com. (n.d.). What are the advantages and disadvantages of using ARIMA models for forecasting? [online] Available at: https://www.linkedin.com/advice/0/what-advantages-disadvantages-using-arima.