D2.1: PJM Regional Electric Consumption Forecasting Project

1. Forecasting Project Objective

The main objective of this project is to develop a predictive model to anticipate regional electric consumption in the short term, specifically with a 24-hour horizon. Using hourly data from the "PJM Hourly Energy Consumption" dataset available on Kaggle and the UCI Machine Learning Repository, we will seek to predict the active power (AEP_MW) consumed in the PJM electric grid.

This project focuses on predicting energy consumption with a one-day-ahead horizon for the following reasons:

- **Optimal operational horizon**: A 24-hour horizon allows grid operators to better plan daily energy distribution and potentially reduce costs during peak hours
- **Practical applicability**: It facilitates the implementation of demand response strategies at the regional level and allows for immediate operational adjustments
- **Precision-utility balance**: It provides an optimal balance between forecast accuracy (which tends to decrease significantly with longer horizons such as weeks or months) and the practical utility of predictions for daily decision-making
- **Alignment with natural patterns**: It aligns perfectly with the natural daily cycles of energy consumption in electrical grids, capturing the inherent seasonality in these data

Although longer horizons could be considered (such as a week or a month), the 24-hour prediction represents the optimal point where technical precision and operational utility converge in the context of electrical grids.

2. Dataset Description

The "PJM Hourly Energy Consumption" dataset contains electricity consumption measurements for the PJM (Pennsylvania-Jersey-Maryland) grid in the United States, with the following characteristics:

- **Total period**: The data spans approximately 14 years, from 2004 to 2018
- **Recording frequency**: Hourly measurements (24 observations per day)
- **Total number of observations**: 121,273 entries
- **Recorded variables**:
  - *Datetime*: Date and time of the observation
  - *AEP_MW*: Electric consumption in megawatts

This project focuses on the analysis and prediction of a single time series (AEP_MW), which represents electrical consumption in megawatts. No exogenous variables are included in the original dataset, although they could be incorporated in future stages of the project to improve prediction accuracy.

Seasonality Analysis

Preliminary analysis of the data reveals multiple seasonal patterns:

- **Daily seasonality**: Consumption peaks typically in the mornings and evenings/nights

- **Weekly seasonality**: Differentiated patterns between weekdays and weekends
- **Annual seasonality**: Significant variations between seasons, with higher consumption in winter and summer due to heating and cooling respectively

Stationarity Analysis and Non-Random Walk Verification

A fundamental requirement for this project is to work with a time series that is not a random walk. To verify this condition, a rigorous analysis of stationarity and time series behavior has been conducted.

Formal Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller (ADF) test has been applied with the following results:

- **ADF Statistic**: -16.446812

- **Critical values**:
  - 1%: -3.430
  - 5%: -2.862
  - 10%: -2.567

Interpretation: The ADF statistic (-16.446812) is considerably more negative than the critical value at 5% (-2.862), which allows us to reject with high confidence the null hypothesis that the series has a unit root. This statistically confirms that the time series is **NOT a random walk**.

Autocorrelation Function (ACF) Analysis

The analysis of the autocorrelation function (ACF) shows:

- Clear cyclical patterns with peaks approximately every 24 hours

- Significant correlation that does not fade quickly as would occur in a random walk
- Stable and predictable temporal dependence structure, characteristic of non-random series

This evidence of autocorrelation reinforces the conclusion of the ADF test and confirms that we are dealing with a series suitable for predictive modeling.

Visual Verification of Seasonality

The visualization of the time series shows recurrent patterns that are incompatible with a random walk process:

- Mean reversion (no indefinite drift)

- Identifiable cycles with consistent periodicity
- Controlled variability within reasonable limits

These multiple verifications unequivocally confirm that the AEP_MW series is **NOT a random walk**, thus fulfilling the fundamental requirement of the project.

Potential Exogenous Variables

Although the current dataset only includes electricity consumption values, the following exogenous variables could be incorporated to improve forecasts:

1. **Meteorological data**: Temperature, humidity, and weather conditions that significantly affect energy consumption.

2. **Calendar**: Indicator variables for holidays, special events, or vacation periods.
3. **Economic indicators**: Data reflecting industrial and commercial activity in the region.
4. **Energy prices**: Electricity tariffs that may influence consumption patterns.

3. Proposed Forecasting Models

Analysis of the literature and previous studies on electricity consumption forecasting suggests that various approaches have been applied to this type of data. For this project, we will evaluate the following models:

Traditional Statistical Methods

1. **Naive Time Series Forecasting**: Especially seasonal Naive with 24-hour and 7-day periods to establish a comparison baseline.

2. **Autoregressive Processes (AR)**: To capture the dependence of current consumption on past values.
3. **Moving Average Processes (MA)**: To model the structure of errors in previous periods.
4. **ARIMA/SARIMA**: With special emphasis on SARIMA to incorporate the strong daily, weekly, and annual seasonality detected in the data.
5. **Exponential models (ETS)**: Particularly the Holt-Winters method to capture the multiple seasonality present in the series.
6. **Regression models with ARIMA errors (ARIMAX)**: To incorporate exogenous variables such as meteorological data if available.

Machine Learning Approaches

1. **LSTM (Long Short-Term Memory) neural networks**: To capture long-term dependencies and complex patterns in energy consumption. According to Marino et al. (2016) and Kong et al. (2019), LSTMs have shown superior results in short-term electrical load prediction.

2. **XGBoost and Random Forest**: Ensemble methods that have demonstrated good performance in predicting energy consumption, especially when incorporating temporal features.
3. **Convolutional Neural Networks (CNN)**: To identify local patterns in energy consumption.

Hybrid Models

1. **CNN-LSTM**: Combining the strengths of CNN for feature extraction and LSTM for sequence modeling.

2. **Decomposition-ensemble models**: Decomposing the time series into components (trend, seasonality, residuals) and applying different models to each component.
3. **Prophet with customized adjustments**: Incorporating domain knowledge about regional energy consumption patterns.

Previous Applications to Similar Data

Previous research on this same dataset or similar ones has shown that:

1. SARIMA models capture multiple seasonality well but may fall short in capturing complex non-linear relationships

2. Recurrent neural networks like LSTM have outperformed traditional methods in 24-hour horizons, especially in PJM data
3. Hybrid approaches that combine series decomposition with specialized models for each component have shown significant improvements in accuracy .

4. Conclusions

The PJM (AEP_MW) electricity consumption time series represents an excellent case study for forecasting techniques for the following reasons:

1. **It is not a random walk**: Confirmed by formal statistical tests (Dickey-Fuller)

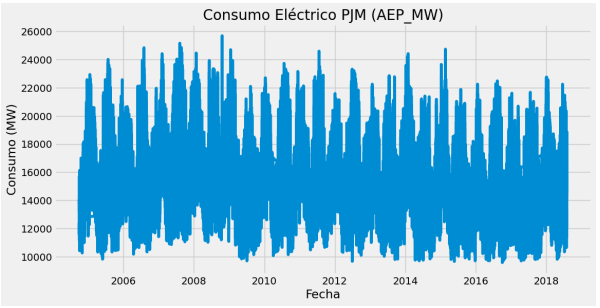
2. **It presents clear and stable patterns**: Evidenced by ACF analysis and series visualization
3. **It has multiple levels of seasonality**: Which allows testing and comparing various models of varying complexity
4. **It covers an extensive period**: With sufficient data for training, validation, and testing of models

D2.2: Final Submission

1. Time series analysis:

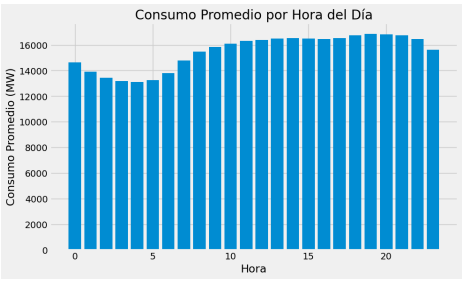
I use the create_features function to break down the temporal characteristics. Creating these temporal variables: date, hour, dayofweek, quarter, month, year, dayofyear, dayofmonth, and weekofyear. This way we have more variables to optimize calculations.

Complete Time Series (PJM Electric Consumption)



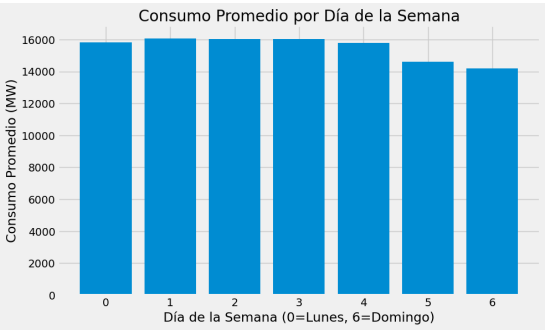
- Electricity consumption mainly oscillates between 10,000 and 26,000 MW during the 2005-2018 period
- Regular peaks and valleys are observed suggesting seasonal patterns
- Extreme peaks, particularly around 2008-2009
- Consistent variability throughout the years, without a dramatic visual trend

Average Consumption by Hour of Day



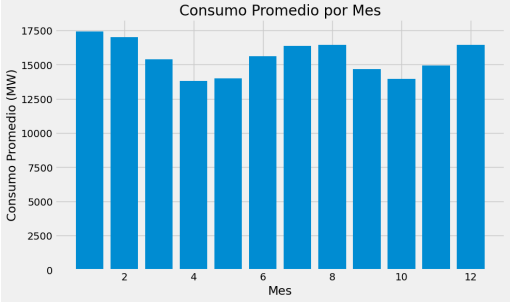
- Minimum consumption between 3-5 AM (approximately 13,000 MW)
- Gradual increase during the morning
- Maximum consumption between 6-9 PM (18-21 hours) (approximately 17,000 MW)
- This pattern reflects human activity: low consumption during the night and peaks during hours of greater activity

Average Consumption by Day of the Week



- Weekdays (Monday to Friday, 0-4) show similar consumption, around 16,000 MW
- Significant drop on weekends, especially Sunday (day 6)
- Saturday (day 5) has approximately 1,000 MW less than weekdays
- Sunday (day 6) shows the lowest consumption with approximately 14,200 MW
- This weekday/weekend difference reflects the reduction in commercial and industrial activity

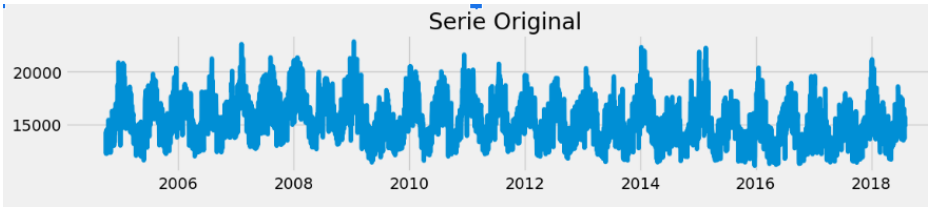
Average Consumption by Month



- Bimodal seasonal pattern:
  - Winter (December-February, months 12, 1, 2)
  - Summer (July-August, months 7-8) shows a second peak
- Minimum consumption in spring and fall (April-May and October)

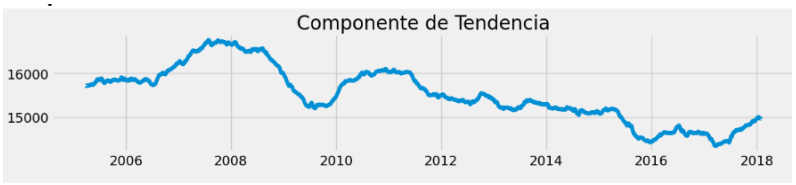
- This pattern is clearly related to heating and air conditioning needs

Original Series



This first graph shows the original data of electricity consumption (AEP_MW) from 2005 to 2018. Fluctuations of electricity consumption (values between 13,000 and 21,000 MW.)

Trend Component



This graph shows the long-term trend of electricity consumption, removing seasonal and daily fluctuations:

- **2005-2008**: There is a gradual increase in consumption (growing trend)
- **2008-2010**: A significant drop is seen (possibly related to the economic crisis)
- **2010-2012**: There is a recovery
- **2012-onwards**: The trend is generally decreasing
- **2017-2018**: A small rebound is observed

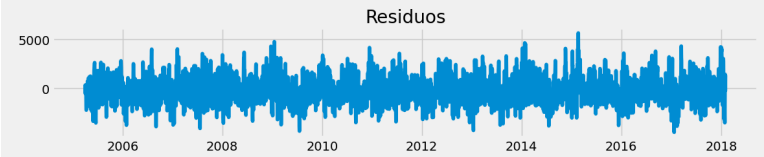
The trend helps us understand how electricity consumption is changing in the long term, possibly due to factors such as the economy, energy efficiency, or demographic changes.

Seasonal Component

The cyclical patterns of consumption that repeat year after year:

- Regular peaks and valleys indicate annual seasonality
- The highest peaks occur in winter and summer
- Valleys occur in spring and fall

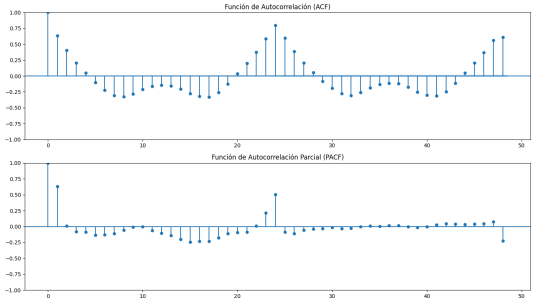
Residuals



The last graph shows what remains after removing the trend and seasonality:

- They represent random variations or factors not explained by seasonal patterns or the trend
- They fluctuate around zero, which is expected in a good decomposition
- There are some occasional peaks that could represent special events, extreme weather conditions, or holidays

2. ARMA

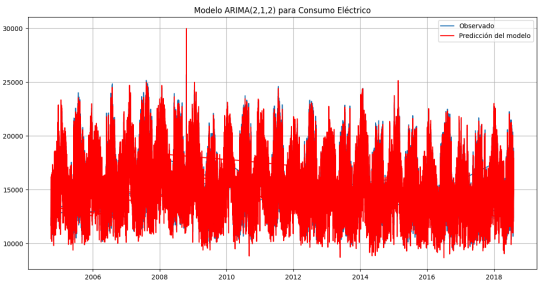


ACF and PACF Analysis

The autocorrelation (ACF) and partial autocorrelation (PACF) graphs show clear patterns:

- The ACF shows significant correlations at multiple lags, with especially pronounced peaks at lags 1, 24, and 48, indicating a strong daily seasonal pattern.
- The PACF shows significant partial correlations mainly at lags 1, 2, and 24, suggesting autoregressive components of order 2 with seasonal effects.

3. ARIMA

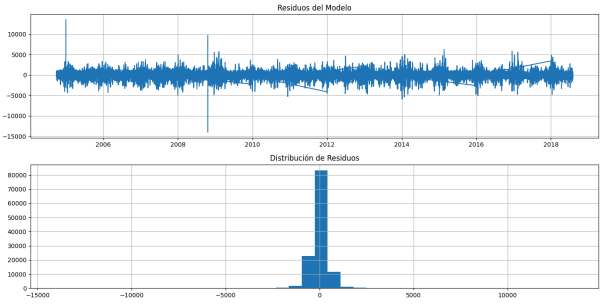


Model ARIMA(2,1,2)

The model summary shows:

- Total number of observations: 121,273
- All coefficients (ar.L1, ar.L2, ma.L1, ma.L2) are statistically significant (p-values < 0.05)
- The ar.L1 coefficient (1.7303) indicates a strong dependence of the current value on the immediately previous value
- The ar.L2 coefficient (-0.8183) partially compensates for the effect of the first lag

Residual Analysis



The model residuals show:

- They mostly fluctuate around zero, which is positive
- Some outliers are observed, particularly around 2006 and 2009
- There don't seem to be obvious systematic patterns in the residuals, although there is some heteroscedasticity (non-constant variance)

Performance Metrics

The model shows good performance:

- MAE: 301.42 MW (mean absolute error)
- RMSE: 460.39 MW (root mean square error)
- MAPE: 1.99% (mean absolute percentage error)

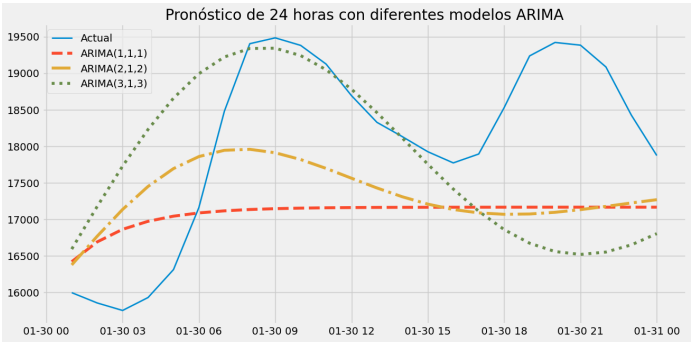
These values indicate that the model has an average error of approximately 2% of the actual value, which is quite satisfactory for this type of data.

Conclusion on the ARIMA model

Conclusion on the ARIMA model The ARIMA(2,1,2) model provides an adequate fit for the electricity consumption time series, capturing both the trend and much of the fluctuation patterns. However, it does not completely capture the seasonal nature of the data (particularly the daily and weekly cycles observed in the initial analysis).

Comparison of different ARIMA models

In the first image, we observe three different configurations of the ARIMA model:



1. **ARIMA(1,1,1)**: The simplest configuration produces an almost flat forecast that fails to capture the daily fluctuations in electricity consumption. This model tends to quickly converge to a constant value, which is insufficient to capture the dynamics of electricity consumption.
2. **ARIMA(2,1,2)**: This model shows more dynamic behavior than ARIMA(1,1,1), but still only partially captures the variations. It performs better in the first few hours of the forecast, but then stabilizes at a level that does not adequately reflect the real patterns.
3. **ARIMA(3,1,3)**: The most complex configuration initially shows the best fit to the morning pattern, following the increase in electricity consumption more closely. However, it fails to capture the second peak in the afternoon/evening and shows a decreasing trend when the actual data is increasing.

The in-sample performance metrics show a gradual improvement as the p and q orders increase:

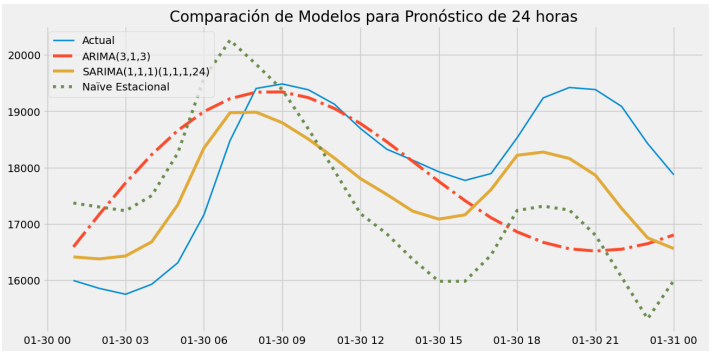
- ARIMA(1,1,1): MAE=320.46, RMSE=491.80, MAPE=2.13%
- ARIMA(2,1,2): MAE=300.99, RMSE=459.07, MAPE=1.99%
- ARIMA(3,1,3): MAE=292.29, RMSE=447.37, MAPE=1.94%

However, when we evaluate the 24-hour forecast performance, the improvement is not consistent across all metrics.

Handling Seasonality

In the second image, approaches for handling seasonality are compared:

1. **ARIMA(3,1,3)**: As mentioned earlier, this model captures some trends but not the daily seasonal patterns.
2. **SARIMA(1,1,1)(1,1,1,24)**: Better captures the general shape of the daily pattern, showing two peaks that correspond with the morning and afternoon patterns in electricity consumption. It is notably superior to standard ARIMA for this type of data. I will focus on it later, to analyze it specifically.
3. **Naïve Estacional**: This simple approach that uses values from 24 hours ago as a forecast shows interesting behavior. It captures some seasonality but with a time shift, resulting in peaks that occur before the actual data. Despite its simplicity, the naïve model demonstrates that incorporating seasonality, even in a basic way, significantly improves the forecast compared to standard ARIMA models.



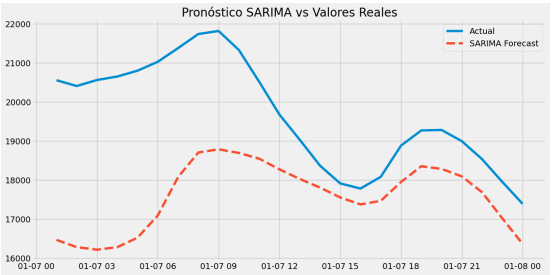
Conclusions

1. **Effect of increasing p and q**: Increasing the autoregressive and moving average orders improves the in-sample fit, but does not guarantee better out-of-sample forecast performance.
2. **Limitations of standard ARIMA**: Traditional ARIMA models (without a seasonal component) are fundamentally unable to capture the daily cyclical patterns present in electricity consumption data.

4. SARIMA Modeling Analysis

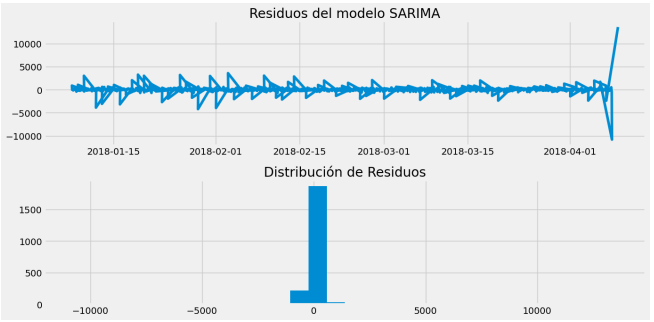
SARIMA Model Performance

The implemented SARIMA(1,1,1)(1,1,1,24) model shows the following important characteristics:



- **Capture of daily patterns**: Captures the daily bimodal structure of electricity consumption, with peaks in the morning and afternoon/evening.
- **Improved accuracy**: The error metrics show an MAE of 2084.97 MW and RMSE of 2548.36 MW, representing an 8.3% improvement in MAE and 6.5% improvement in RMSE compared to the ARIMA(3,1,3).
- **Forecast structure**: Although the SARIMA model consistently underestimates the actual values (approximately 2000-3000 MW difference), it correctly maintains the shape of the daily consumption curve.

Análisis de Residuos



- **Residual distribution**: The residuals are mainly centered around zero, with a distribution that resembles a normal one.
- **Temporal pattern**: Some structure is observed in the residuals, with regular peaks suggesting that there are still seasonal patterns that the model does not completely capture.
- **Outliers**: There are some extreme values in the residuals, particularly toward the end of the period (April 2018), which could indicate special events or changes in consumption patterns.

Limitations of the SARIMA Model

- **Systematic underestimation**: There is a consistent bias where predictions are lower than actual values, suggesting that there might be contextual factors not captured by the model.
- **Computational complexity**: SARIMA models require more resources and computation time than traditional ARIMA models, which could be an important consideration for real-time applications.
- **Multiple seasonality**: Although the model incorporates daily seasonality (24 hours), it does not explicitly capture the weekly patterns or longer seasonal patterns that are also present in the data.
- **Despite the improvement**, there is still room to perfect the model, possibly through:
    - Inclusion of exogenous variables (such as temperature)
    - Incorporation of multiple levels of seasonality

5. Discusión de Resultados

Comparación de Modelos

Modelo	MAE (MW)	RMSE (MW)	Performance
ARIMA(1,1,1)	1330.84	1487.39	Provides acceptable initial results but quickly converges to a constant value
ARIMA(2,1,2)	1239.60	1354.24	Improves on the previous model with a 6.9% reduction in MAE and 9.0% in RMSE
ARIMA(3,1,3)	1184.84	1555.23	Achieves the lowest MAE among all models, but its RMSE is higher than ARIMA(2,1,2), suggesting it may have larger errors at certain points

6. Interesting Findings

Importance of Seasonality

One of the most significant findings of this analysis is the strong seasonality present in electricity consumption data, which manifests at multiple levels:

1. **Daily seasonality**: Consumption patterns follow a 24-hour cycle with two characteristic peaks (morning and evening/night).
2. **Weekly seasonality**: There is a clear difference between weekdays and weekends.
3. **Annual seasonality**: Seasonal variations are observed with higher consumption in winter and summer.

Tradeoff between Complexity and Performance

The results show that increasing model complexity (p and q orders in ARIMA) improves in-sample fit, but doesn't always translate into better out-of-sample forecasts. The ARIMA(3,1,3) has the best in-sample fit, but its forecast performance is not consistently superior to ARIMA(2,1,2).

Análisis Visual vs. Métricas Numéricas

A particularly relevant finding is the discrepancy between quantitative error metrics and qualitative evaluation of forecasts. The SARIMA model produces forecasts that visually better follow daily patterns, but its error metrics are higher due to the mentioned phase error. This underscores the importance of complementing numerical metrics with visual evaluation when selecting models for time series with complex seasonal patterns.

Limitations of the ARIMA Approach for Data with Multiple Seasonality

The results indicate that standard ARIMA models, even with higher orders (p=3, q=3), have fundamental limitations for adequately modeling time series with multiple levels of seasonality. Although they can capture some short-term dynamics, they fail to represent complex cyclical patterns.

Conclusions

Based on the results obtained, we can conclude that:

1. For electricity consumption data with strong seasonal components, even a simple Seasonal Naïve approach can outperform more complex models that don't consider seasonality.
2. Traditional ARIMA models can provide accurate forecasts in terms of error metrics, but fail to capture the true structure of the data.
3. The SARIMA model, despite having higher error metrics, provides forecasts that better reflect daily patterns of electricity consumption, which could be more valuable for practical applications such as energy distribution planning.
4. To further improve the results, approaches that incorporate multiple levels of seasonality simultaneously could be explored, such as TBATS models, Prophet, or neural networks specialized for time series.
5. The choice of optimal model depends on the specific objective of the forecast: if point numerical precision is prioritized, ARIMA(3,1,3) would be the best option; if capturing data structure is valued more, SARIMA would be preferable despite its apparently worse error metrics.