



STREAMING DATA MANAGEMENT & TIME SERIES ANALYSIS

Alfio Lanza

Project Overview

The core **objective** of this project is to forecast missing hourly values for December 2016 using three distinct approaches:



ARIMA



UCM



ML

Data Exploration & Preprocessing

Preprocessing is a critical step in time series forecasting, as raw data often exhibits missing values, non-stationarity, and seasonality. Understanding these properties ensures that models are applied correctly and yield reliable predictions.

Stationarity in Time Series

Stationarity ensures that statistical properties remain stable over time, making models more reliable. The **ADF test** checks for non-stationarity by detecting unit roots. If a series is non-stationary, transformations like differencing are applied to stabilize it.

Decomposing Seasonality and Trend

Time series data often exhibit **trend**, **seasonality**, and **residual** noise. The trend represents long-term movement, while seasonality captures recurring patterns such as daily or weekly cycles. Decomposing these components helps improve forecasting accuracy by isolating key patterns and making the data more interpretable.

Feature Engineering for Machine Learning

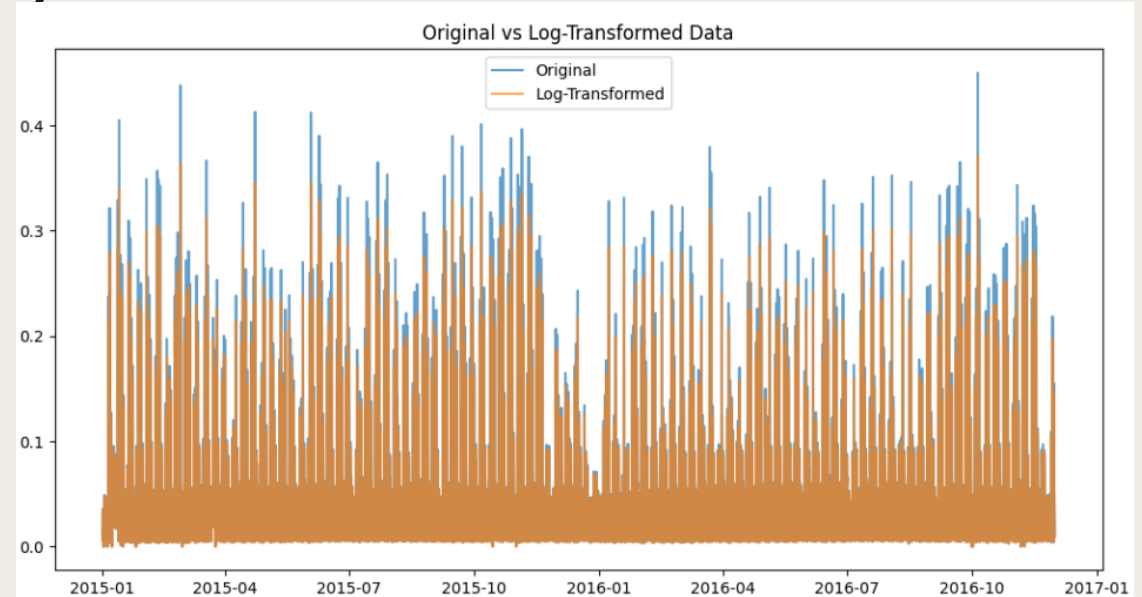
Machine learning models require structured inputs to capture time-based dependencies. **Temporal features** such as hour, day, and month provide context, while **lag features** incorporate past values to detect patterns. **Rolling statistics** help smooth fluctuations and enhance predictive performance.

AutoRegressive Integrated Moving Average (ARIMA) – 1/3

■ Theoretical Foundations

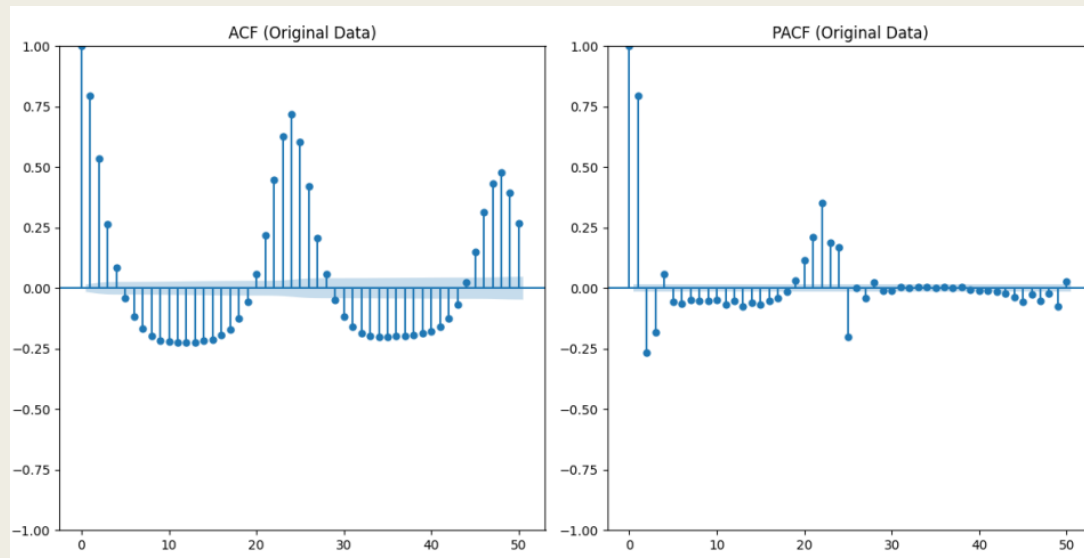
The ARIMA model predicts future values using past observations and errors. It consists of three components:

- **Autoregressive (AR):** Captures relationships with past values.
- **Integration (I):** Differencing is applied if needed to achieve stationarity.
- **Moving Average (MA):** Models dependencies on past errors.



To improve stability, a logarithmic transformation was tested. The Augmented Dickey-Fuller (ADF) test confirmed that the original series was already stationary, making transformation unnecessary.

AutoRegressive Integrated Moving Average (ARIMA) – 2/3



Autocorrelation Analysis

Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots revealed significant seasonal patterns at lag 24, suggesting daily seasonality. The PACF plot indicated that an AR(1) model would be a suitable choice.

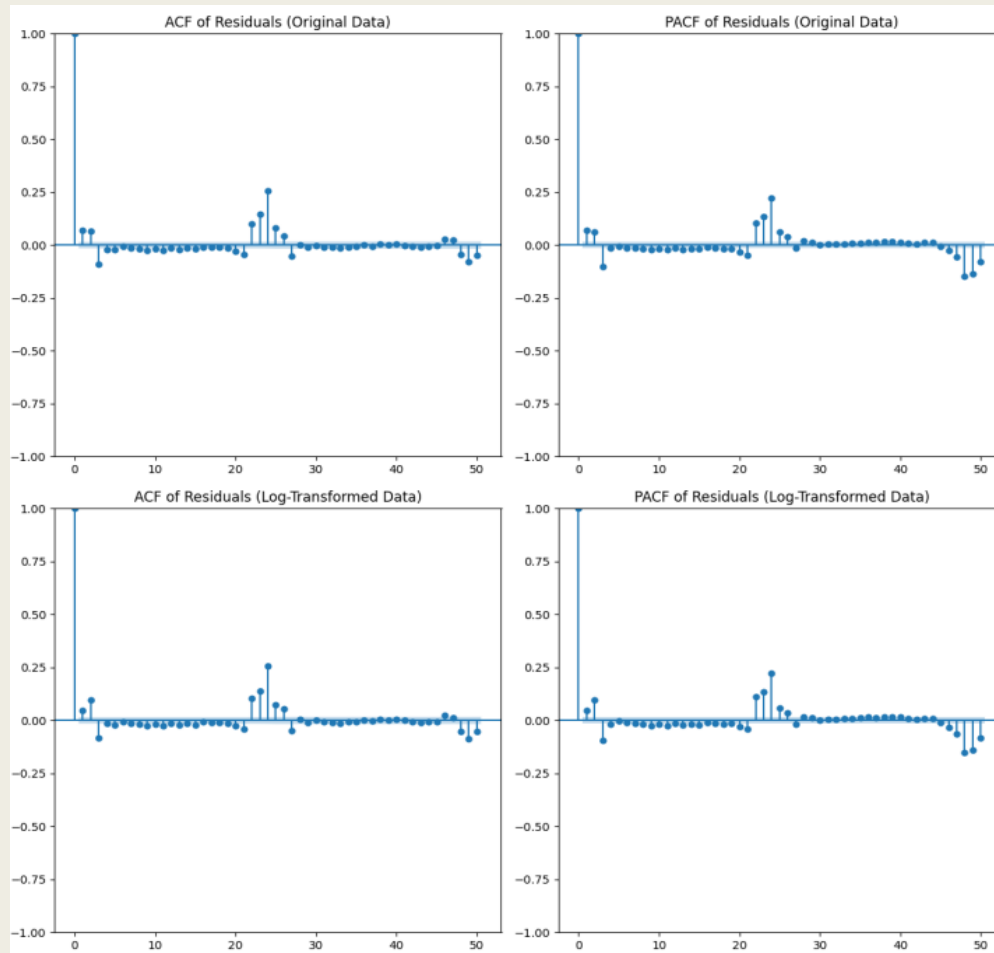
SARIMAX Results						
=====						
Dep. Variable:	X		No. Observations:		16800	
Model:	SARIMAX(1, 0, 1)x(1, 0, 1, 24)		Log Likelihood		38547.819	
Date:	Tue, 04 Feb 2025		AIC		-77085.637	
Time:	15:39:38		BIC		-77046.992	
Sample:	0		HQIC		-77072.886	
	- 16800					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

ar.L1	0.6354	0.004	174.082	0.000	0.628	0.643
ma.L1	0.1147	0.005	21.733	0.000	0.104	0.125
ar.S.L24	0.9846	0.001	1553.116	0.000	0.983	0.986
ma.S.L24	-0.8911	0.002	-434.555	0.000	-0.895	-0.887
sigma2	0.0006	2.14e-06	280.233	0.000	0.001	0.001
=====						
Ljung-Box (L1) (Q):	79.17	Jarque-Bera (JB):	321842.82			
Prob(Q):	0.00	Prob(JB):	0.00			
Heteroskedasticity (H):	0.96	Skew:	1.74			
Prob(H) (two-sided):	0.14	Kurtosis:	24.16			

SARIMA Model Selection

A seasonal ARIMA (SARIMA) model was tested with parameters (1,0,1) for the non-seasonal part and (1,0,1,24) for seasonality. The model on log-transformed data had a slightly better statistical fit (lower AIC/BIC values) but did not significantly improve forecasting accuracy.

AutoRegressive Integrated Moving Average (ARIMA) – 3/3



■ Forecasting Performance

Both the original and log-transformed models were evaluated on a validation set (one month of hourly data). Despite the log-transformed model having lower residual variance, both achieved the same Mean Squared Error ($MSE = 0.0018$). Since log transformation added complexity without improving performance, the simpler approach using the original data was preferred.

Conclusion

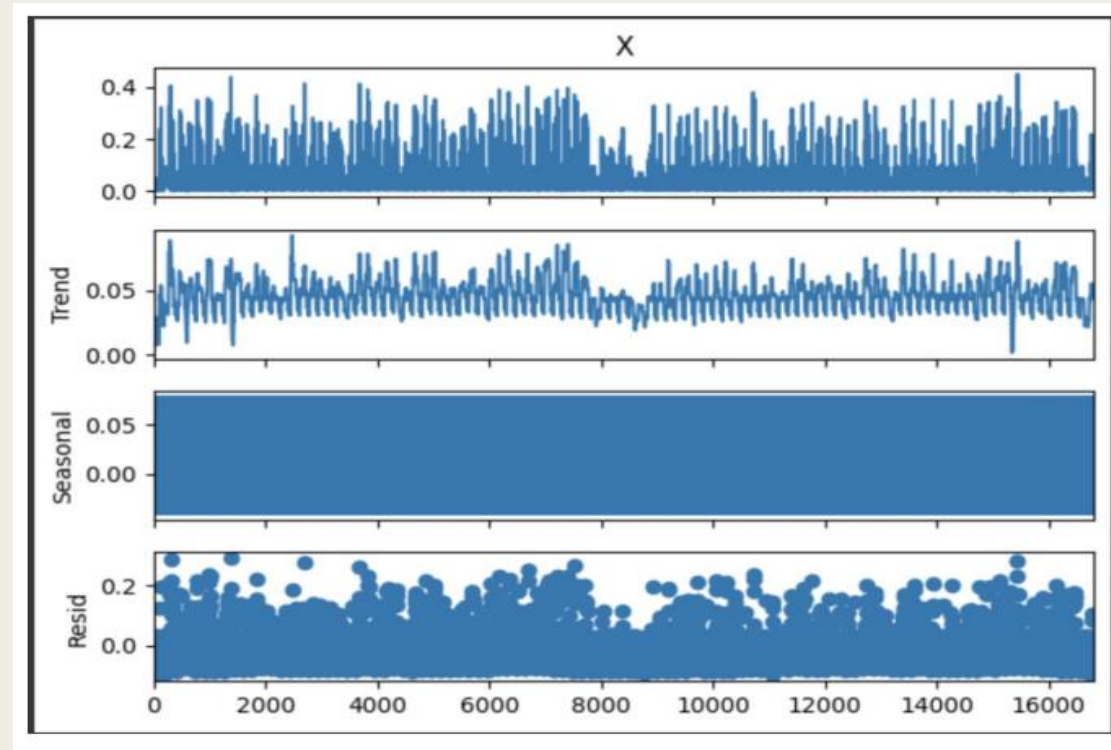
The log-transformed model had a slightly better statistical fit, but its forecasting accuracy was identical to the original model. Given this, using the original data is preferable as it avoids unnecessary complexity while maintaining the same predictive performance.



Unobserved Components Model (UCM) - 1/3

■ Theoretical Foundations

Unobserved Components Models (UCM) decompose time series into trend, seasonality, and residual components within a state-space framework. Unlike ARIMA, UCM explicitly models non-stationary elements, making it ideal for time series with evolving trends.



Modelling Process

The time series was decomposed into:

- Trend: Captures long-term variations.
- Seasonality: Displays a clear 24-hour cycle.
- Residuals: Appeared random, confirming successful isolation of trend and seasonality.

Unobserved Components Model (UCM) - 2/3

Unobserved Components Results						
=====						
Dep. Variable:	X	No. Observations:	16800			
Model:	local level	Log Likelihood	37112.011			
	+ stochastic seasonal(24)	AIC	-74218.021			
Date:	Tue, 04 Feb 2025	BIC	-74194.838			
Time:	15:43:02	HQIC	-74210.371			
Sample:	0					
	- 16800					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

sigma2.irregular	2.543e-11	2.12e-06	1.2e-05	1.000	-4.16e-06	4.16e-06
sigma2.level	0.0007	5.16e-06	126.210	0.000	0.001	0.001
sigma2.seasonal	8.274e-07	3.14e-08	26.378	0.000	7.66e-07	8.89e-07
=====						
Ljung-Box (L1) (Q):	5.20	Jarque-Bera (JB):	293608.31			
Prob(Q):	0.02	Prob(JB):	0.00			
Heteroskedasticity (H):	0.95	Skew:	-0.83			
Prob(H) (two-sided):	0.04	Kurtosis:	23.43			

■ Model Selection & Refinement:

1. **Initial Model:** Included a local level for trend, built-in seasonality, and an irregular component. However, the **Ljung-Box test** detected residual autocorrelation.
2. **Refinement:** Removing the irregular component had no effect, indicating seasonality needed better representation.

Unobserved Components Model (UCM) - 3/3

Unobserved Components Results						
=====						
Dep. Variable:	X	No. Observations:	16800			
Model:	local level	Log Likelihood	36155.121			
Date:	Tue, 04 Feb 2025	AIC	-72294.243			
Time:	15:46:17	BIC	-72232.410			
Sample:	0	HQIC	-72273.841			
	- 16800					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

sigma2.irregular	2.758e-05	2.74e-06	10.060	0.000	2.22e-05	3.29e-05
sigma2.level	0.0007	7.33e-06	100.444	0.000	0.001	0.001
beta.sin_1	0.0262	0.004	6.363	0.000	0.018	0.034
beta.cos_1	-0.0268	0.003	-8.679	0.000	-0.033	-0.021
beta.sin_2	-0.0115	0.002	-4.875	0.000	-0.016	-0.007
beta.cos_2	-0.0221	0.002	-10.080	0.000	-0.026	-0.018
beta.sin_3	-0.0121	0.001	-12.053	0.000	-0.014	-0.010
beta.cos_3	0.0067	0.001	6.669	0.000	0.005	0.009
=====						
Ljung-Box (L1) (Q):	0.19	Jarque-Bera (JB):	341371.33			
Prob(Q):	0.67	Prob(JB):	0.00			
Heteroskedasticity (H):	0.92	Skew:	-0.71			
Prob(H) (two-sided):	0.00	Kurtosis:	25.04			
=====						

- **Final Model:** Replaced the built-in seasonal structure with sinusoidal components (sine & cosine functions) as exogenous regressors.

Results

- Sinusoidal components significantly improved the seasonal fit.
- The Ljung-Box test confirmed residuals were now uncorrelated.
- The final model effectively captured trend and seasonality, reducing autocorrelation and enhancing interpretability.

Machine Learning (XGBoost)


Modeling Process

In the modeling process, feature engineering plays a key role: time-based features such as hour, day, and month are extracted, and lag features are created to capture the past 24 hours of data. Rolling statistics, like the mean and standard deviation over the past week, are also calculated to capture trends and fluctuations.

The model's performance is then evaluated using the Mean Squared Error (MSE), which shows a low value of 0.0004, indicating a good fit to historical data.

Forecasting Step

For forecasting, the same time-based features are extracted from the forecast dataset, and an iterative forecasting approach is used, which updates lag features by incorporating both historical and predicted values. This approach ensures that the model continues to improve its predictions over time.



THANK YOU FOR YOUR
ATTENTION