Università degli Studi di Milano-Bicocca
Master's Degree in Data Science
Academic Year 2022-2023

# Streaming Data Management and Time Series Analysis

2017 energy consumption: analysis and forecasting

# Introduction

- **High frequency** time series: one observation every **10 minutes**, from 01/01/2017 00:00:00 to 30/11/2017 23:50:00.
  - Total number of observations: **48096**.
  - Test-set: from 01/12/2017 00:00:00 to 30/12/2017 23:50:00 (4320 observations).

- Goal: **MAE** (Mean Absolute Error) minimization.

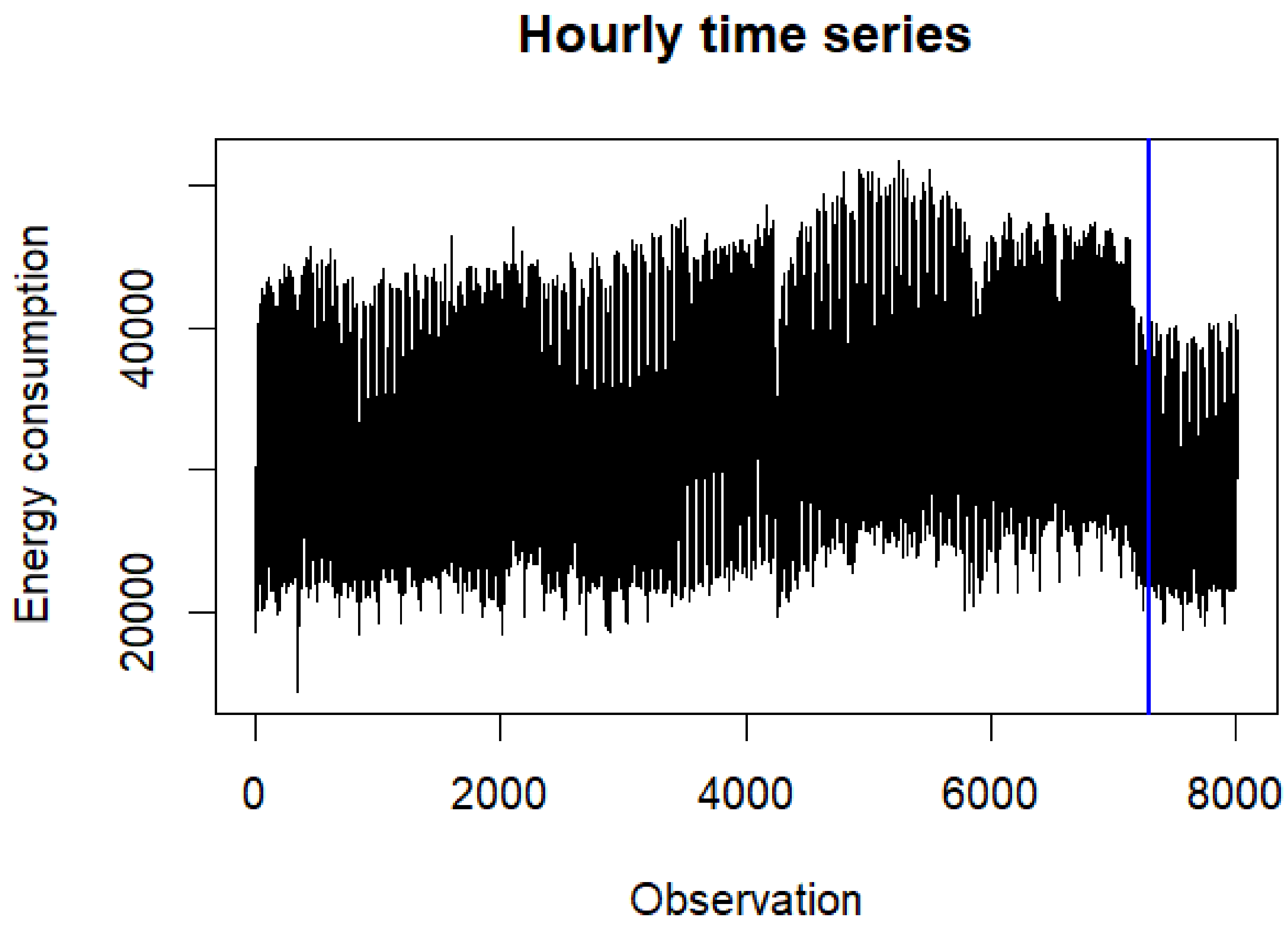
- 3 different approaches:
  - **ARIMA**, **UCM**, **Machine Learning**.
  - Choosing the best model for each approach.

# Methodology

- Dividing into **training** and **validation**:
    - November 2017 as validation-set.
    - Change from daylight saving time to solar time: drastic trend drop (blue line).
    - **Cautious** assessments on validation-set.
    - Trying to keep **generalization** as high as possible.

- Each model re-estimated on training+validation.

- Expecting better performance on the test-set than on the validation-set.

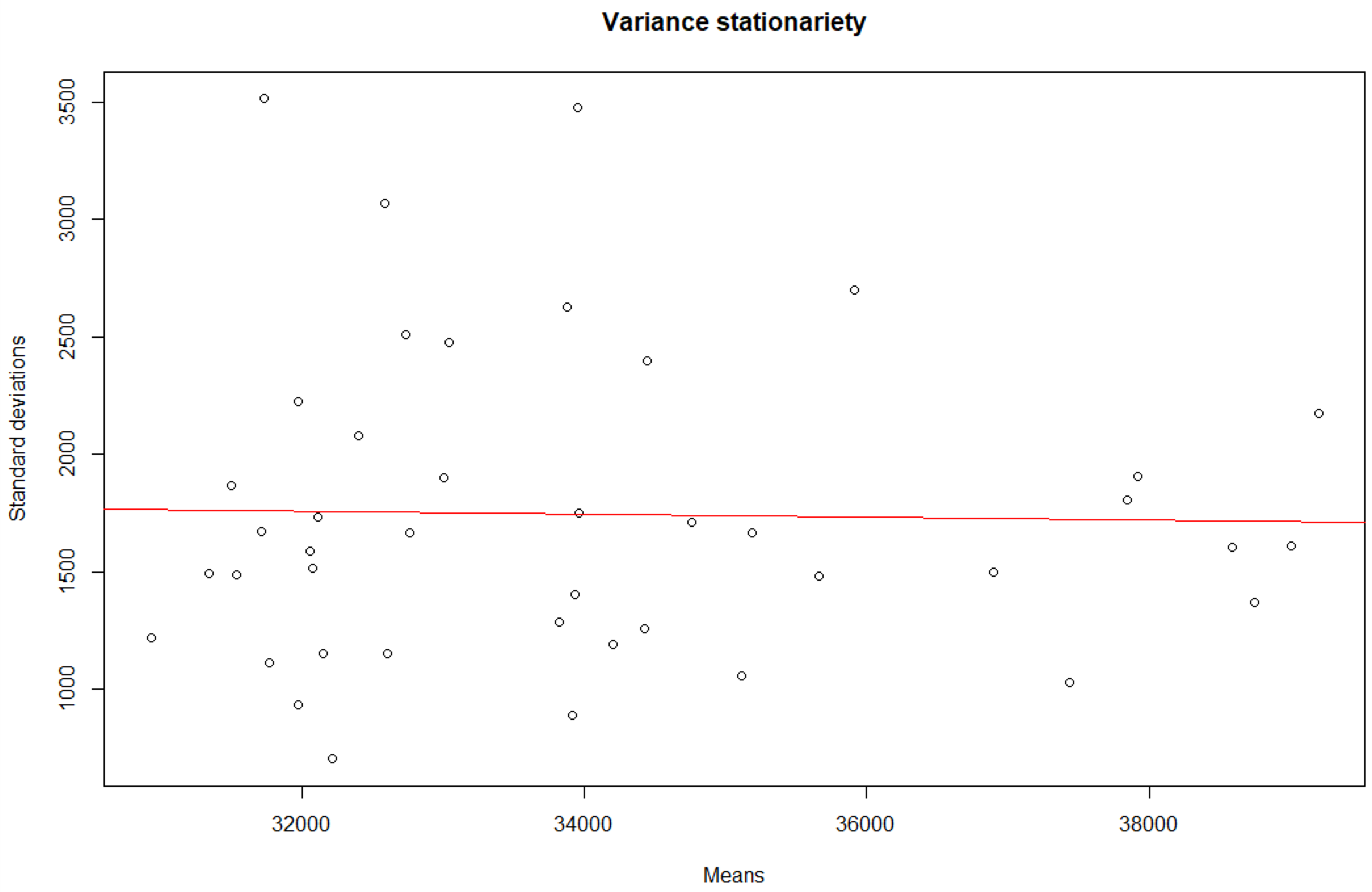


# ARIMA

- Grouping: **24 daily time series**, one for each hour of the day.
    - Losing hour correlation.
    - Improved efficiency.

- **Variance stationariety**:
    - No linear increasing trend.
    - H17, but applies to all other hours.

- **No mean stationariety**:
    - We need two differences.
        - Seasonal diff. (s = 7).
        - Simple diff. (s = 1).



# ARIMA

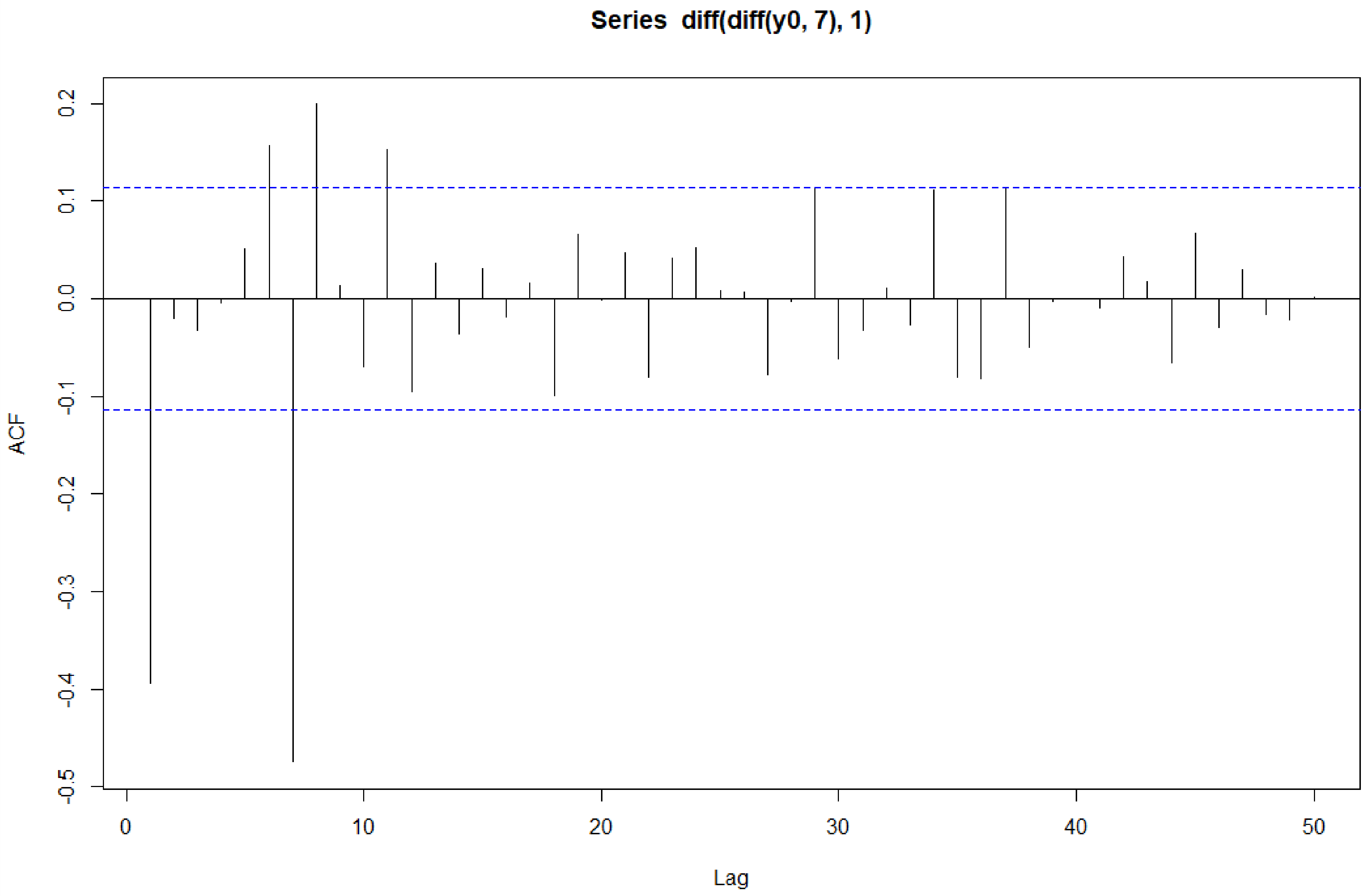
- Starting point for modeling: ARIMA(0,1,0)(0,1,0)[7].

- **MA(1)** and **SMA(1)[7]**.
    - **ARIMA(0,1,1)(0,1,1)[7]**.
    - Airline model.
    - MAE: 1466.737.

- Different consumption **day/night**.

- Residual analysis: many outliers.
    - Most on **holidays**.

- **Time change**: 29 October.



# ARIMA

- 2 different ARIMA models:
  - Night (H21-H3): **ARIMA(0,1,1)**.
  - Day (H4-H20): **ARIMA(0,1,1)(0,1,1)[7]**.

- Dummy for **holidays**.

- Regression variable for **time change** (only for validation-set).

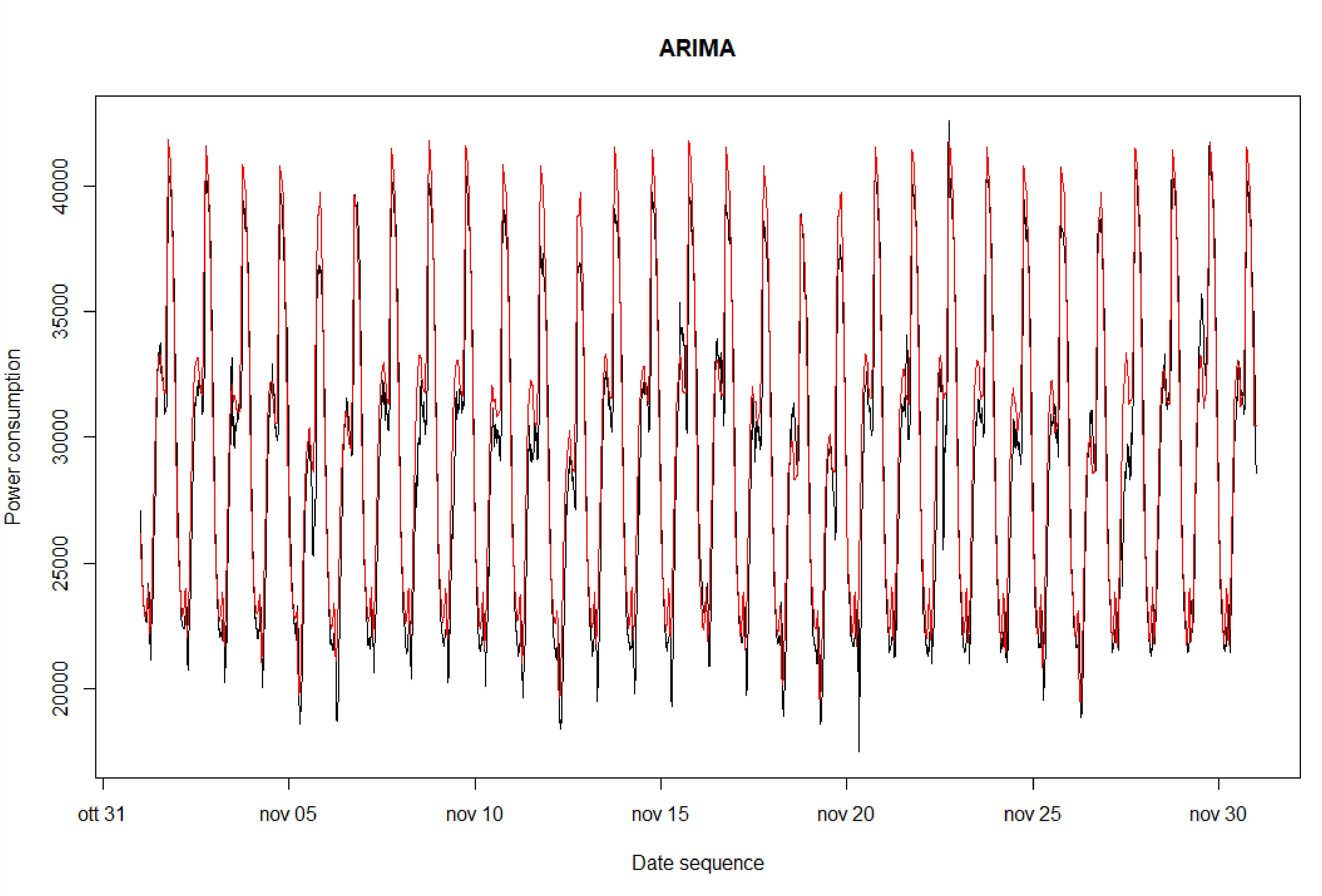
- Insertion of **AR(2)** for daytime: Cleaner ACF+PACF and significativity.

- Re-estimation on training+validation.

# ARIMA

- Final model:
    - Night (H21-H3): **ARIMA(0,1,1)** + **holidays**.
    - Day (H4-H20): **ARIMA(2,1,1)(0,1,1)[7]** + **holidays**.

- **MAE**: 1287.33.



# UCM

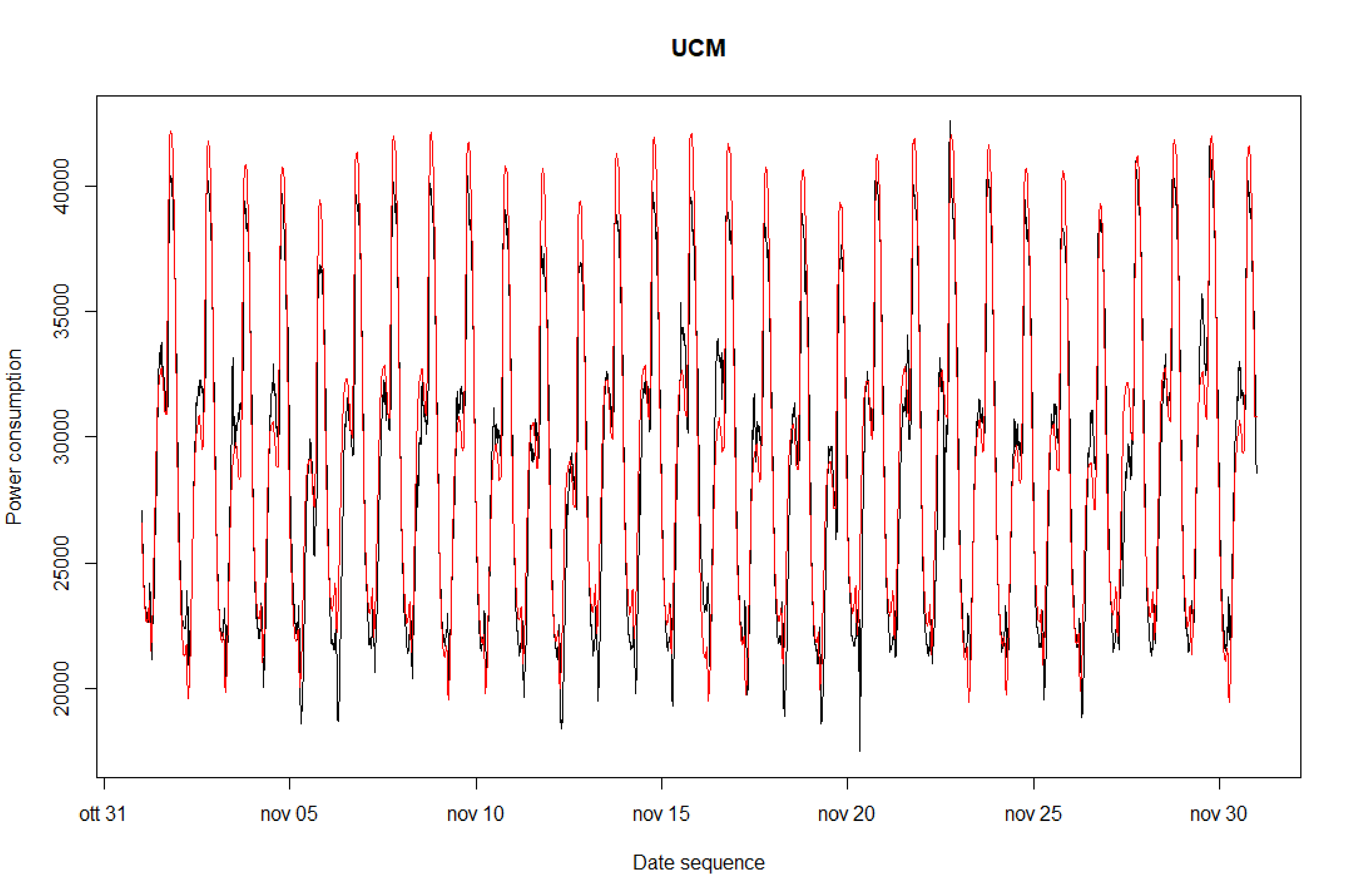
- Grouping data by **hours**.

- Different models and combinations tested:
    - **Local Linear Trend + 24 stochastic dummies**.
        - Very growing trend.
    - Local Linear Trend + 24 stochastic dummies + **stochastic cycle (168 hours)**.
        - Better estimation (MAE = 1873.31), but different period.
    - **Local Linear Trend + 24 stochastic dummies + stochastic sinusoids (period = 168)**.
        - Grid Search on the number of sinusoids.
        - Best n = 6.
    - Holidays dummies brought no improvement.

# UCM

- Final model:
  - **Local Linear Trend.**
  - **24 stochastic dummies.**
  - **6 stochastic sinusoids** (period = 168).

- **MAE** = 1299.51.



# Machine Learning

- 2 different grouping approaches:
    - **Hourly** series (as for UCM).
    - 24 **daily** series (as for ARIMA).

- 3 different algorithms tested:
    - **Random Forest**.
    - **XGBoost**.
    - **Support Vector Machines**.

- **Recursive** method.

- Different **lags** tested as regressors: 2, 8, 16, 24 (1 day), 48, 168 (1 week).

# Machine Learning

- Each model has its own best combination between grouping and lags.

- Generally, the best grouping is the first one (**hourly**).

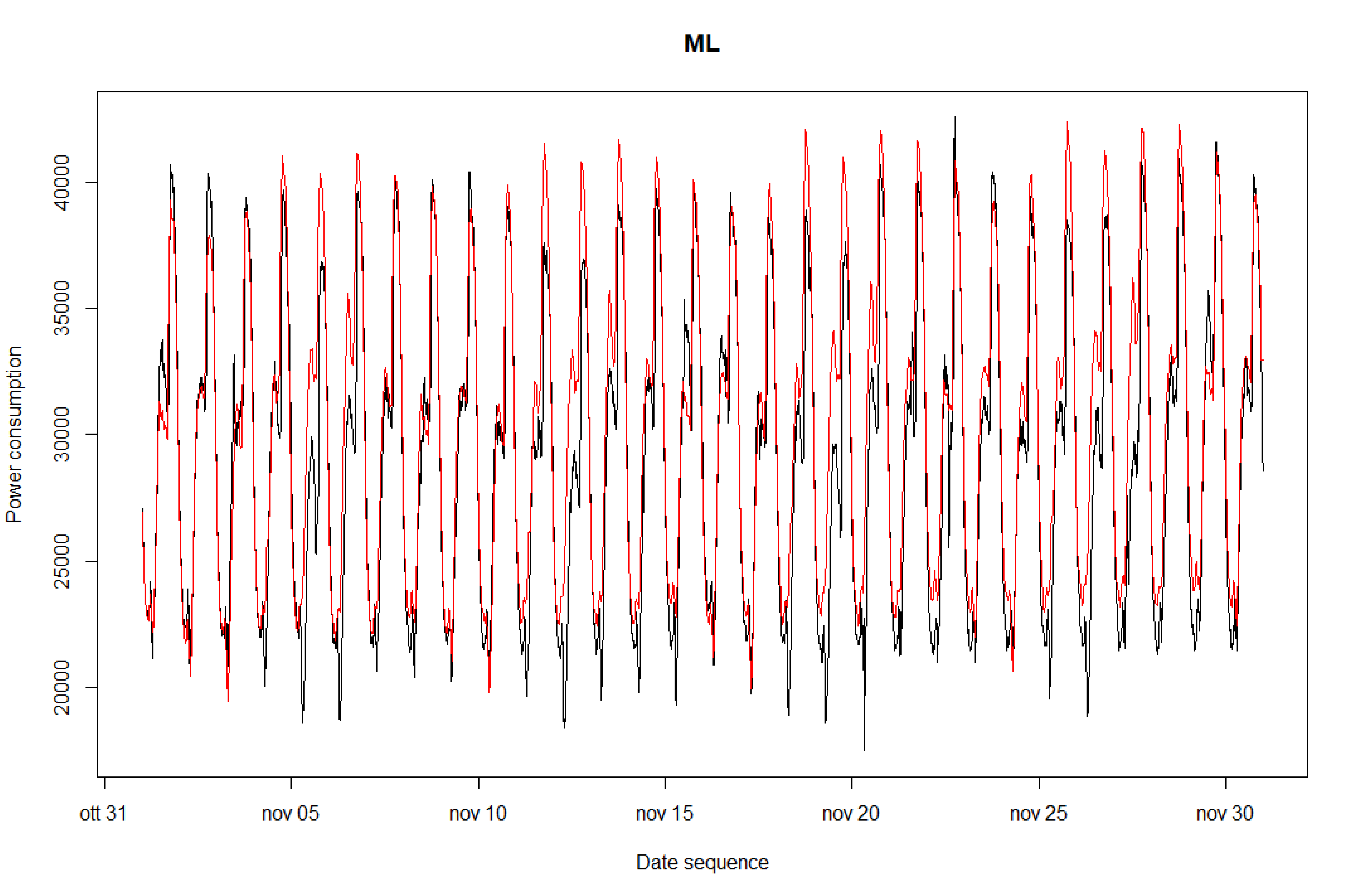
- Best results:
  - **SVM** (linear kernel, 24 lags).
  - **XGBoost** (nrounds = 1000, 168 lags).

- Holidays dummies brought no improvement.

# Machine Learning

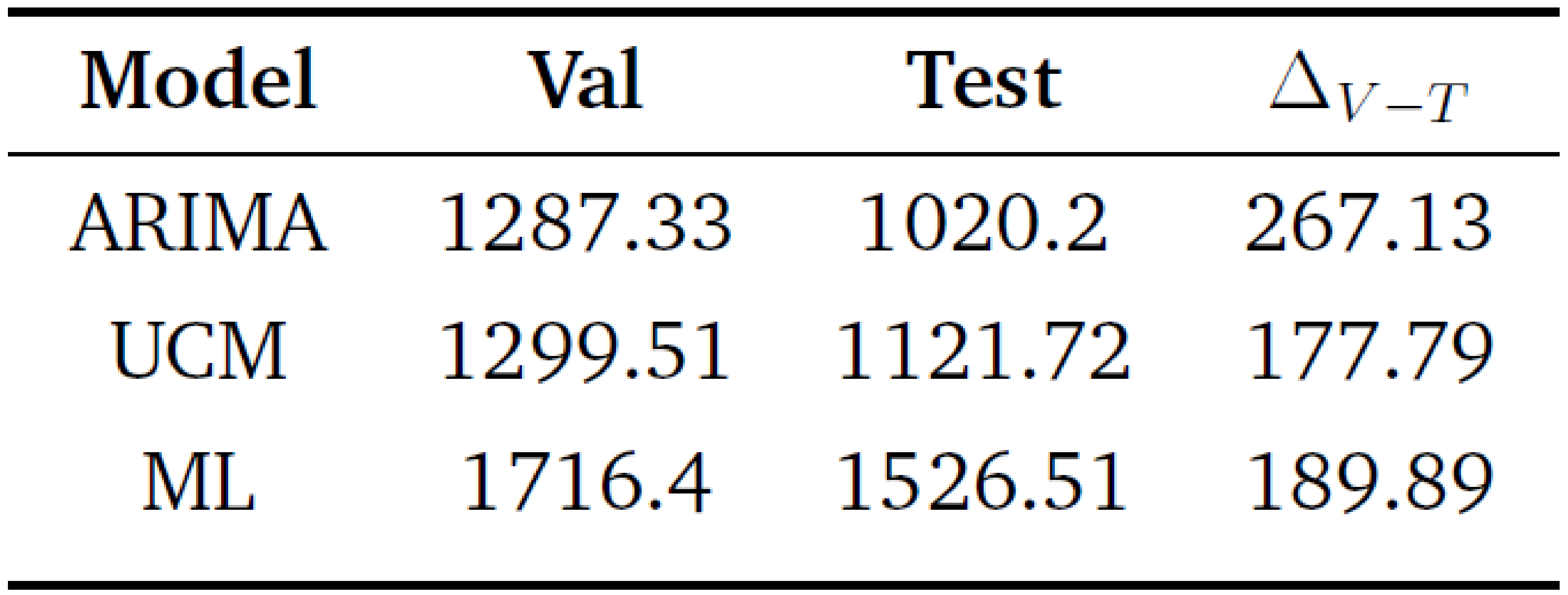
- Final model:
  - **XGBoost** (nrounds = 1000, lags = 168).

- **MAE** = 1716.4.



# Final results

- Better **generalization** on **test-set** than on validation-set.
    - Time change: 29 October.
    - Re-estimation on training+validation.
    - Focus on generalization capabilities during the selection process.



| **Model** | **Val** | **Test** | $\Delta_{V-T}$ |
|---|---|---|---|
| ARIMA | 1287.33 | 1020.2 | 267.13 |
| UCM | 1299.51 | 1121.72 | 177.79 |
| ML | 1716.4 | 1526.51 | 189.89 |