Streaming Data Management and Time Series Analysis project

Samir Doghmi - 897358

Contents

01 Data exploration and preprocessing

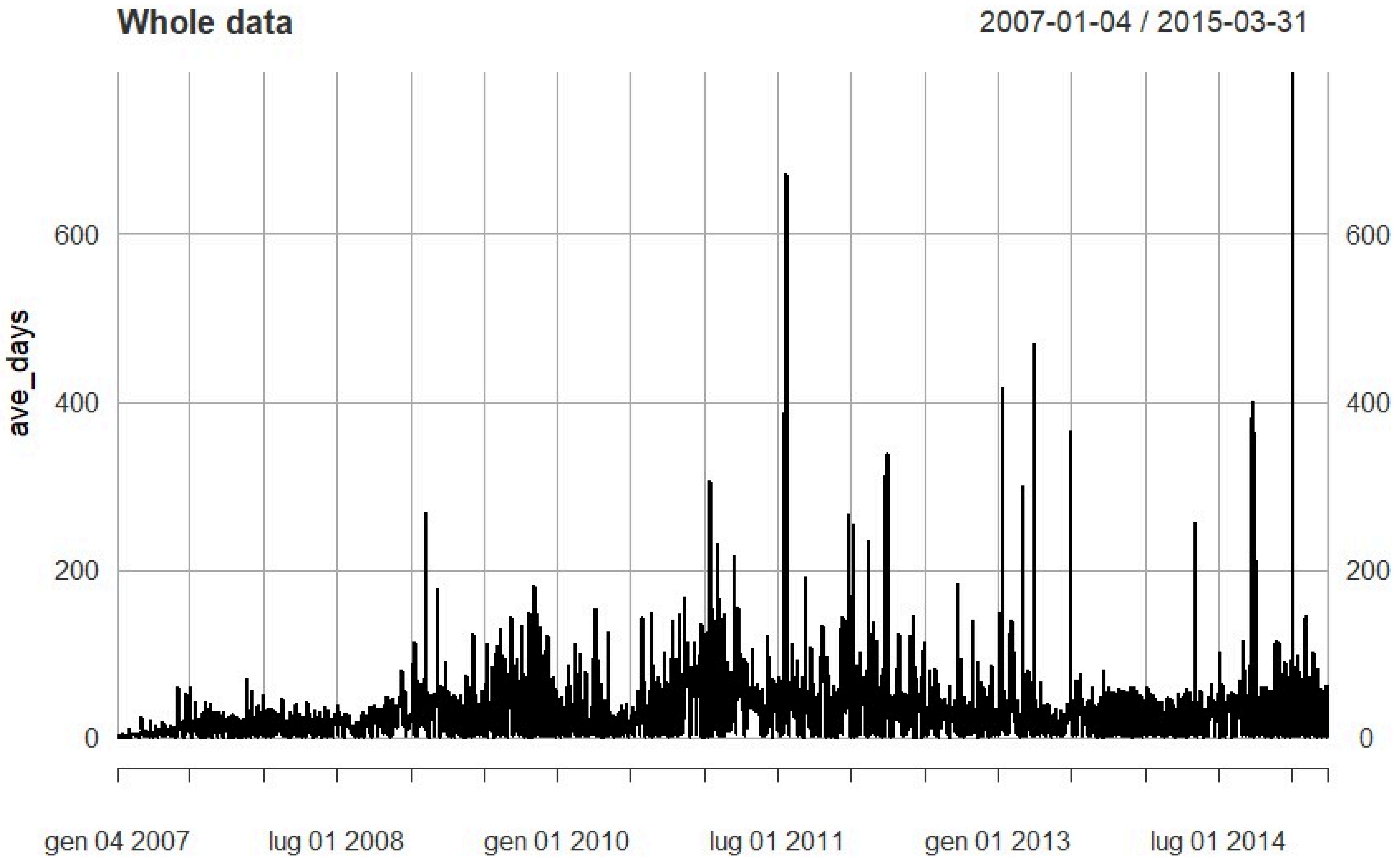
02 Models

03 Conclusions and future development

Dataset

Dataset is made of 3009 records each with the following three columns:

- **Date**: string with date in format yyyy-mm-dd ranging from 2007-01-04 to 2015-03-31;
- **weekday**: string with the name of the weekday;
- **ave_days**: floating point number representing the average number of days needed to close the requests that were closed that day.



Objective

The project aims to predict the same time series with a ARIMA model, and UCM model and a machine learning model. Thus, the purpose is to forecast daily values for 2015-04-01\2015-11-07 period and evaluate predictions using the Mean Absolute Error (MAE) metric.

The dataset was then divided as follows:

- **Train**: 2007-01-04 to 2014-08-22
- **Test**: 2014-08-23 to 2015-03-31

01 Data exploration and preprocessing

01 Outliers and null values

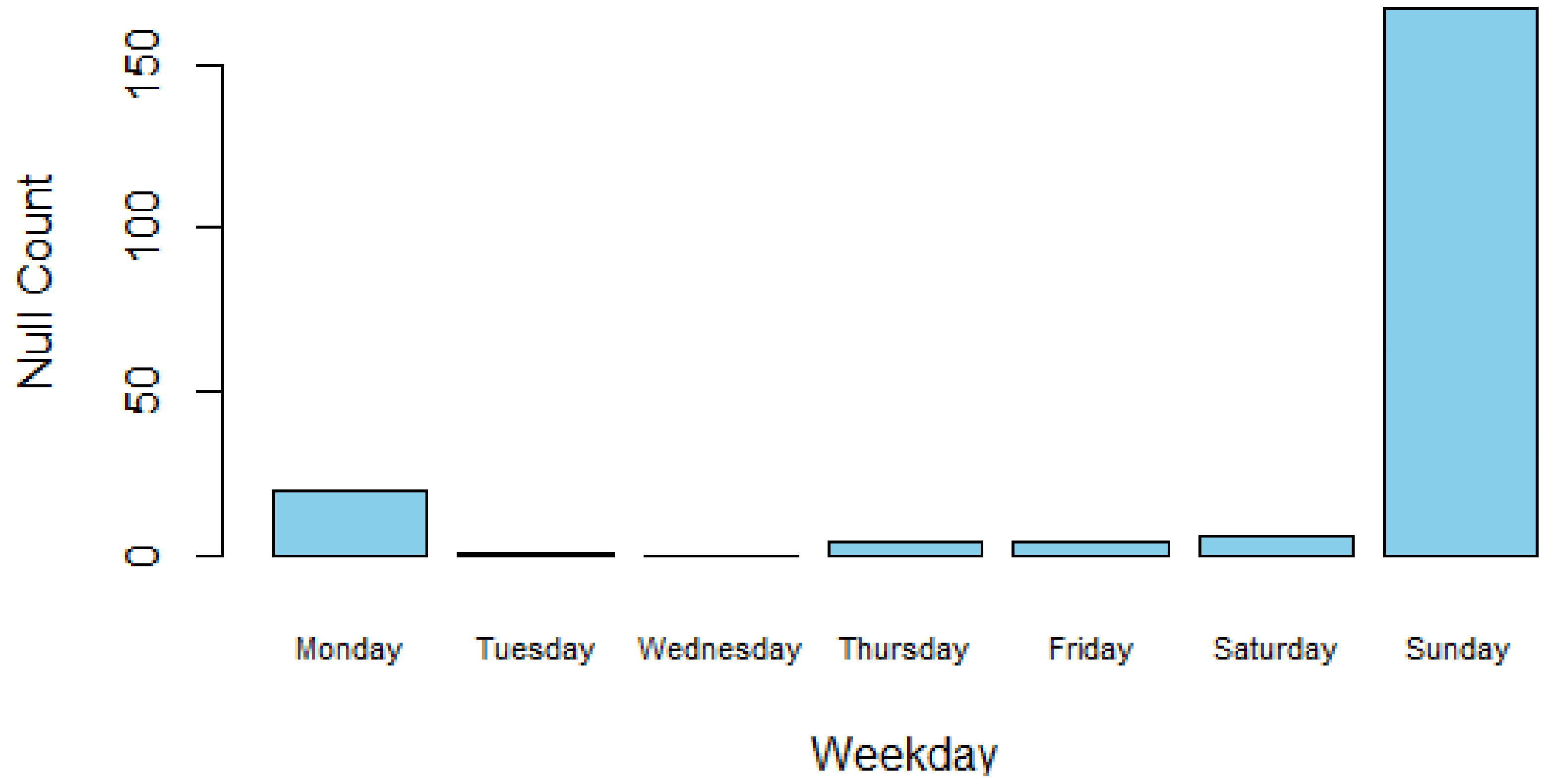
02 Seasonality

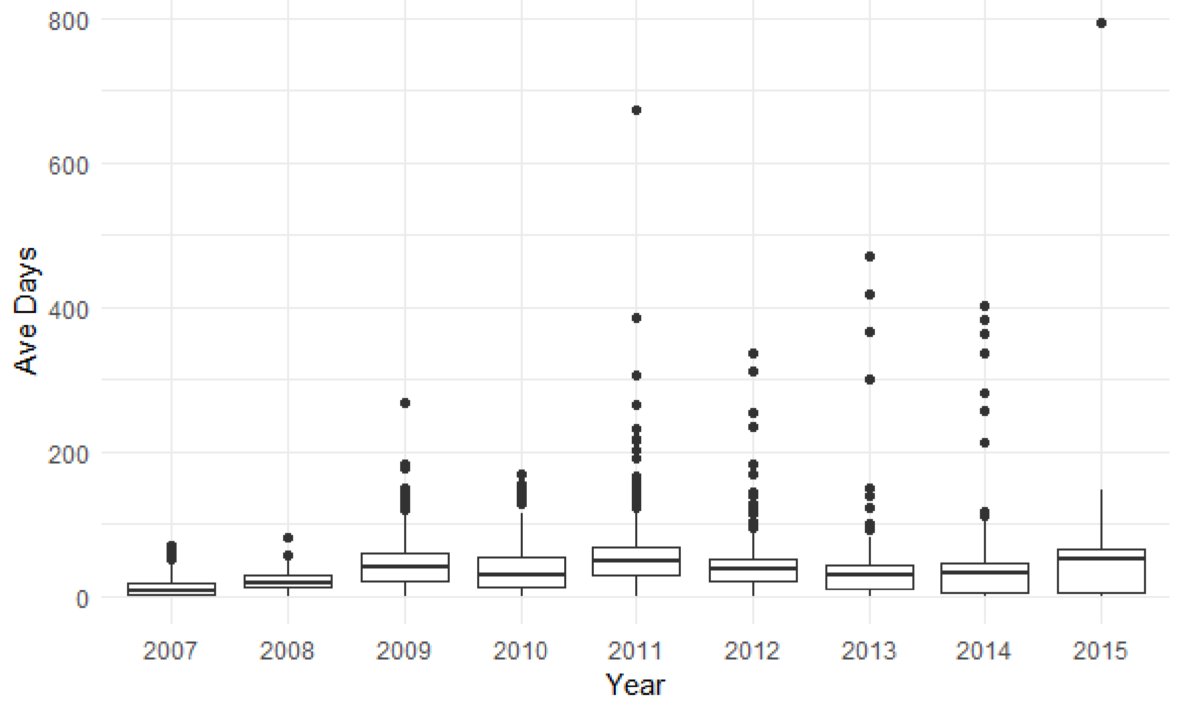
03 Stationarity analysis

1.1 Outliers and null values

The first step was to verify the temporal continuity of the series by checking for the missing values. We noticed that there are 202 null values, most of them appearing on **Sundays.** We handled by replacing them with the last observed value in the series.

To handle the outliers, we adjusted the values by replacing the value exceeding the maximum threshold of the IQR with the value of the threshold itself.

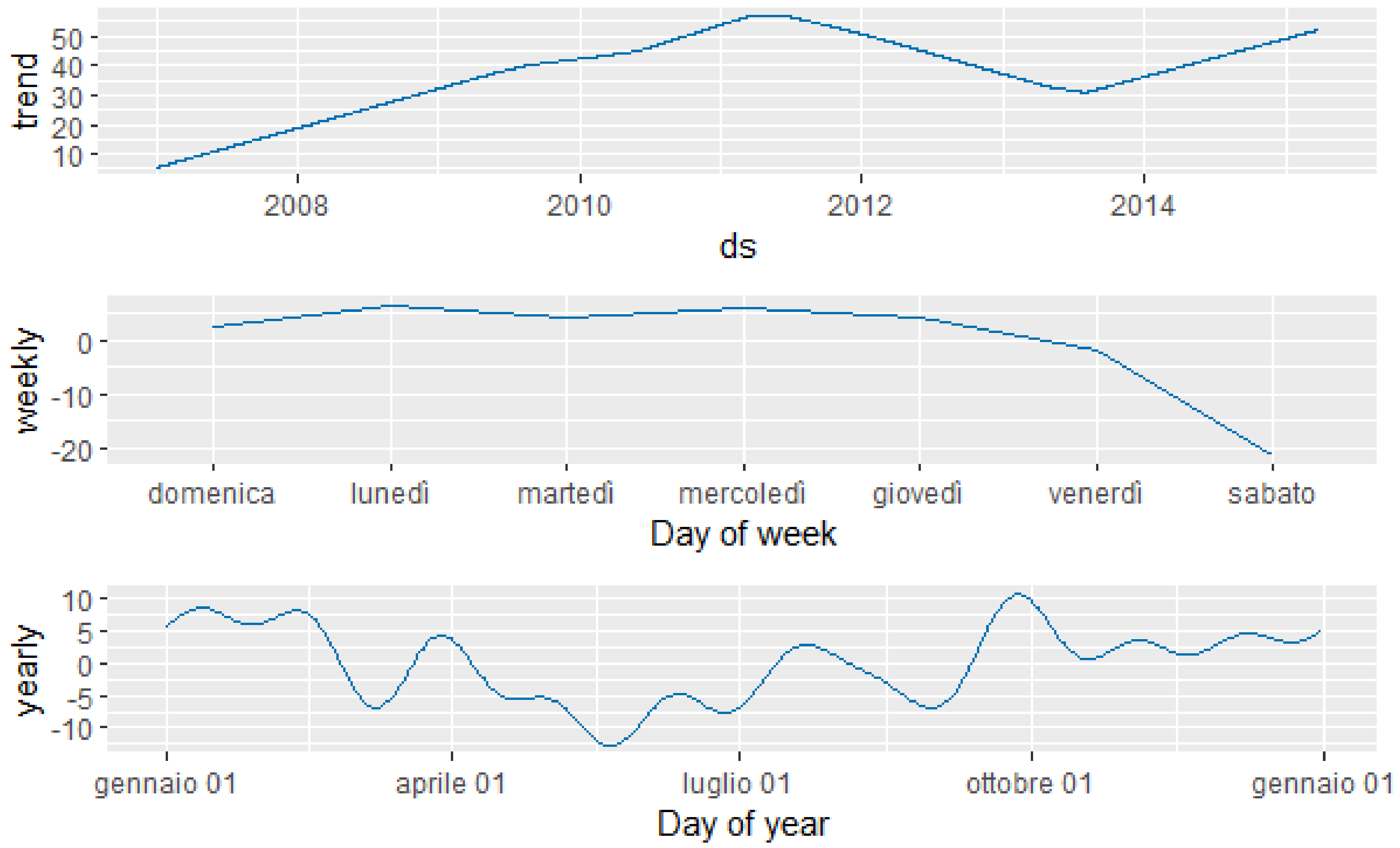




1.2 Seasonality

We observed a generally increasing trend over the years in the average number of days required to close a request, except for the period between 2011 and mid-2013.

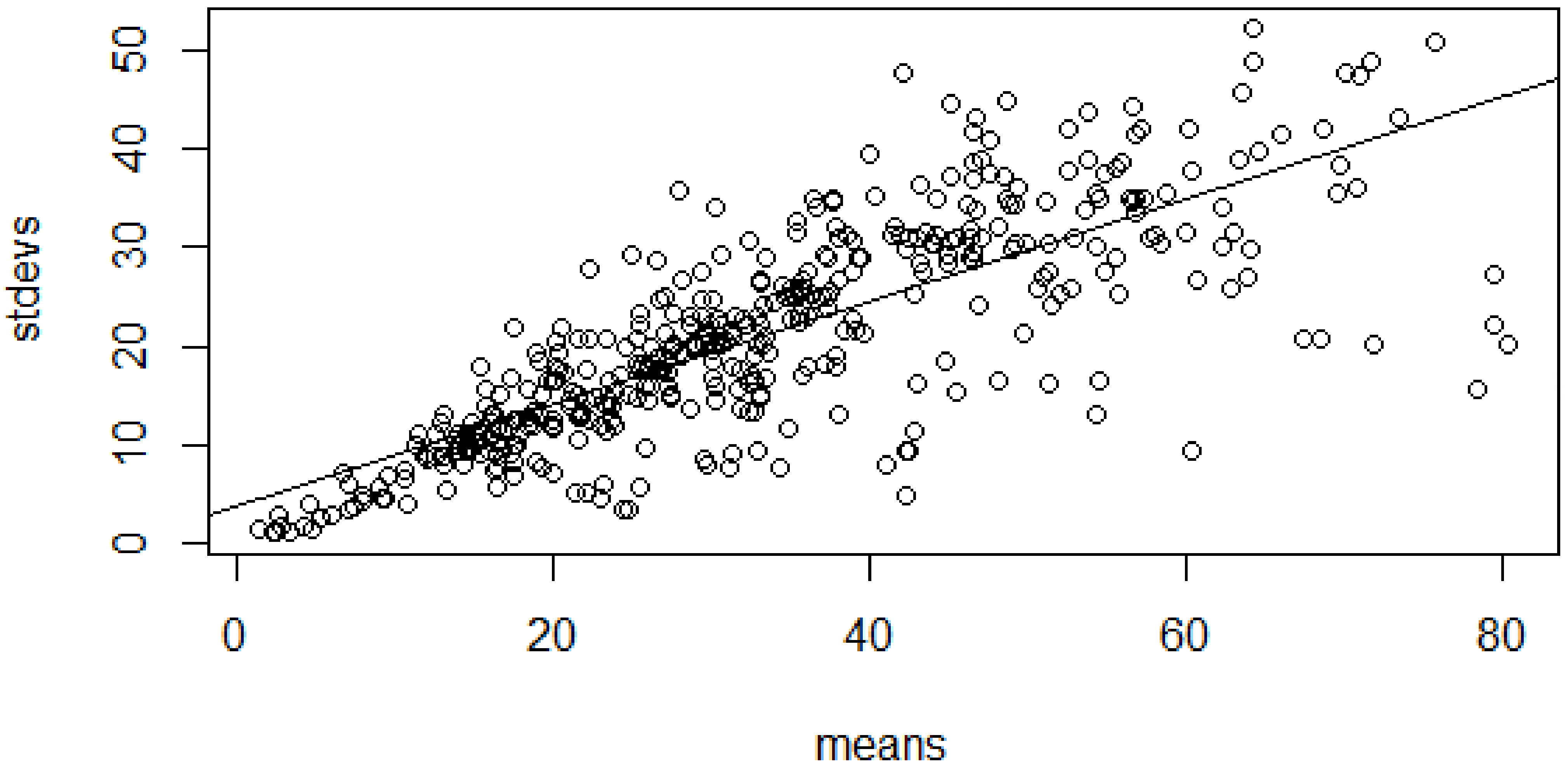
Additionally, the data exhibits a clear **weekly seasonality**.



1.3 Stationarity Analysis

The ADF test indicated that the time series is **stationary** (p-value = 0.01)

The mean and variance change over time showing a linear growth trend, so we considered implementing a **logarithmic transformation** to stabilise the variance.

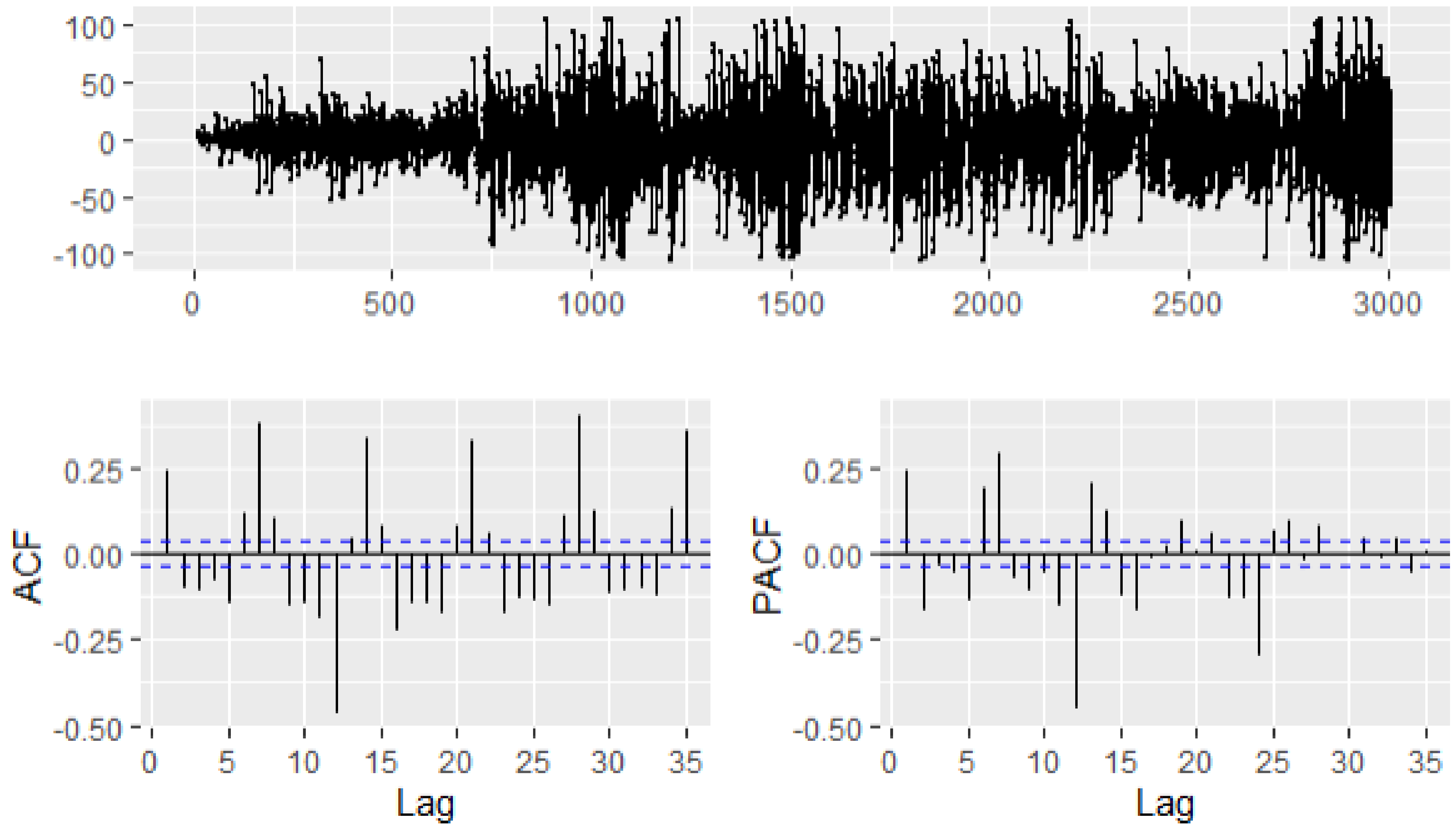


02 Models

01 ARIMA

02 UCM

03 ML



2.1 ARIMA

We experimented with several models by examining the model coefficients and analyzing the ACF and PACF of the residuals. A specific test set was built for each model to incorporate all the regressors used for the predictions.

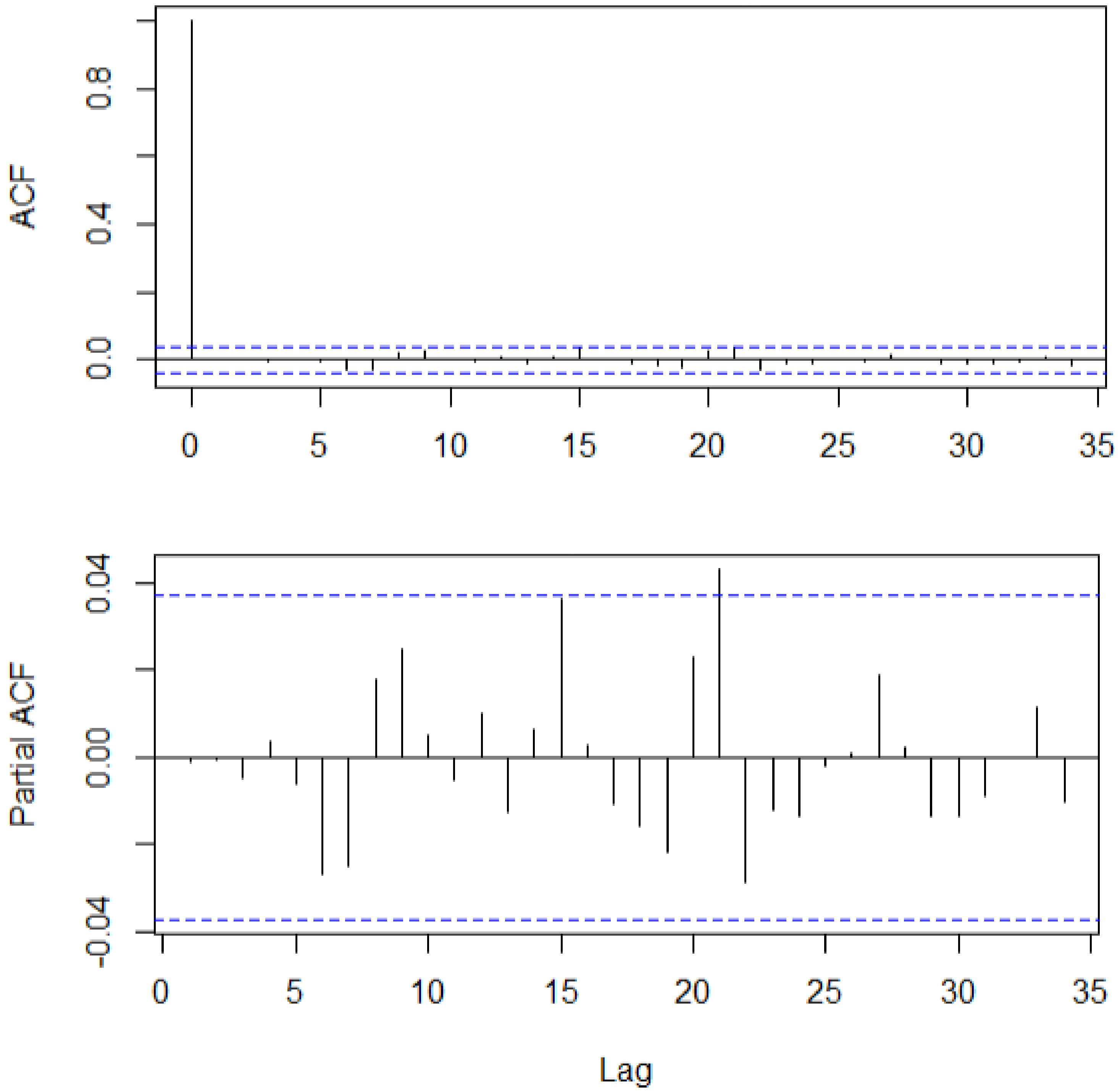
By comparing the results using mean absolute error, we determined that the ARIMA (3,0,1)(1,0,1)[7] model performed the best.

Model	Mae
(2,1,1)(1,0,1)[7]	19.86
(3,0,1)(1,0,1)[7] with dummies	16.41
(2,0,1)(1, 1, 2)[7] with sinusoids	19.38

2.1 ARIMA (3,0,1)(1,0,1)[7] with dummies

This result was achieved by including five dumy variables:

- dum_sunday;
- dum_saturday;
- dum_holy_feriale;
- dum_holy_saturday;
- dum_holy_sunday.



2.2 UCM

For the implementation of these models, it was it was decided to keep the same split of the time series into training and test series as has been done for the ARIMA models

It can be seen from the table that the first model performed better than the other two, obtaining the lowest Mae value.

Model	Mae
dummy variables as regressors, weekly seasonal dummies and integrated random walk	16.76
dummy variables as regressors, weekly seasonal dummies and local linear trend	18.87
dummy variables as regressors, weekly seasonal dummies and annual trigonometric component and local linear trend	19.38

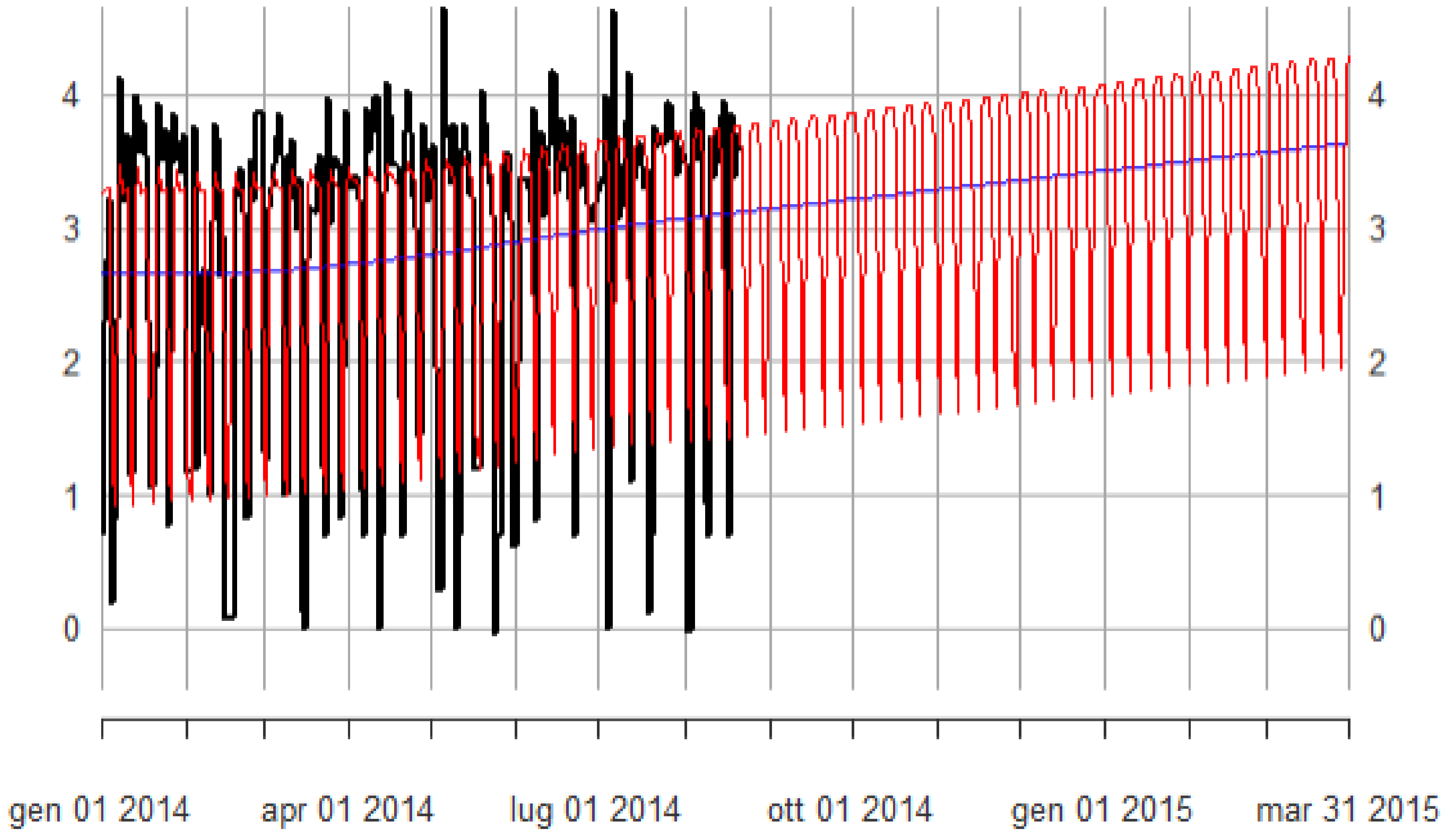
2.2 UCM

This result was achieved by including three dumy variables:

- dum_holy_feriale;
- dum_holy_saturday;
- dum_holy_sunday.

And by considering the following factors:

- integrated random walk
- weekly seasonal dummy



2.3 ML

For machine learning models, it is essential to have a set of highly explanatory covariates. We considered three different models-

By comparing the results using mean absolute error, we determined that the SVM model performed the best, achieving an MAE value of 16.25

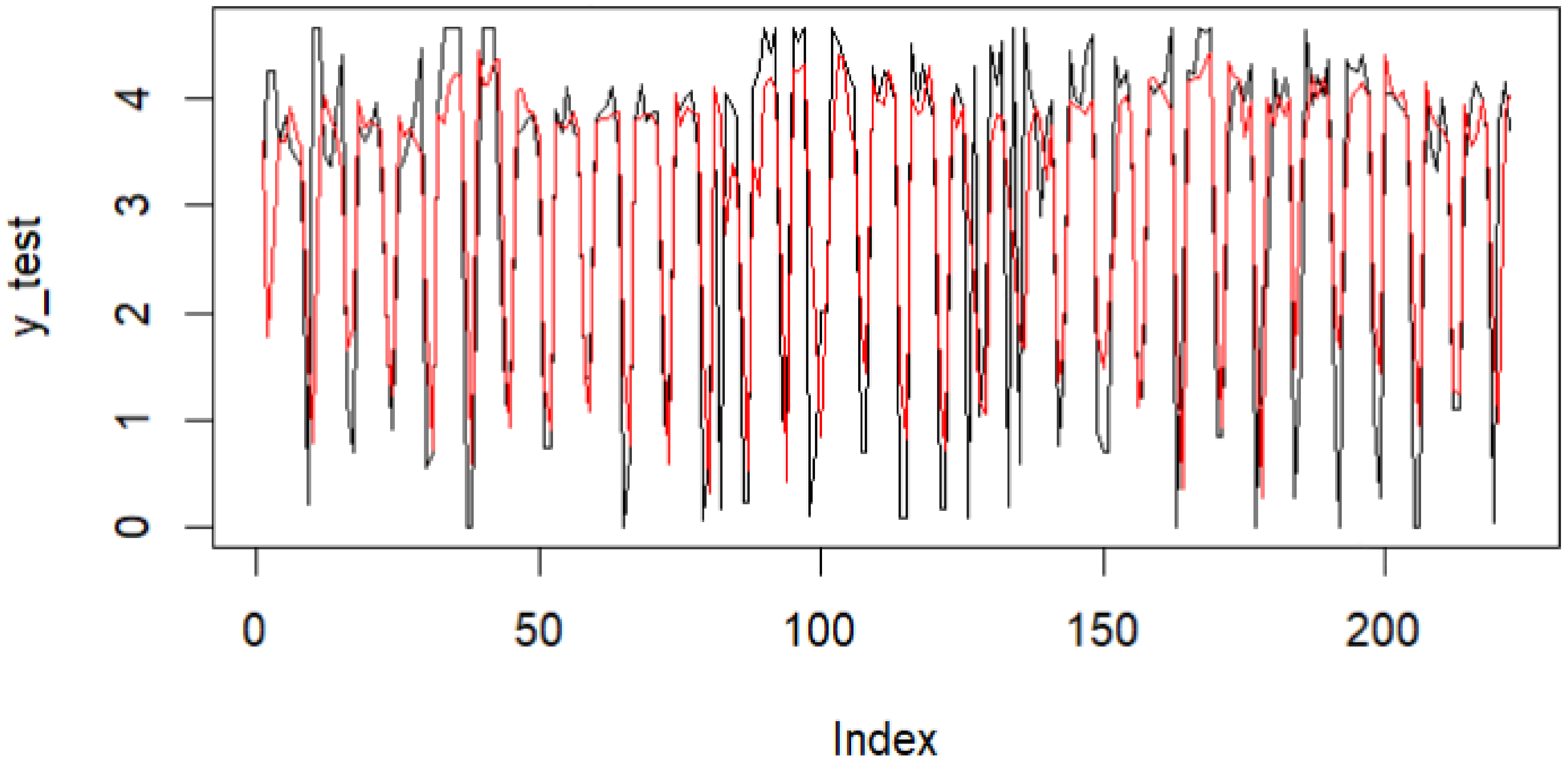
Model	Mae
Random Forest	17.10
SVM	16.25
XGBoost	18.71

2.3 ML

We employed radial kernel that is a used in to measure the similarity between two data points in a way that allows the model to handle non-linear relationships.

This result was achieved by including these dummy variables:

- dum_holy_feriale;
- dum_holy_saturday;
- dum_holy_sunday.
- dum_month
- day_of_year



03 Conclusions and future development

3.1 Conclusions

Our results showed that all three models, ARIMA, UCM , and ML, achieved similar Mae values around 16.50 when tested on the period from 2014-08-22 to 2015-03-31. This indicates that no single model consistently outperformed the others across all metrics.

For future developments, we recommend exploring different models such as deep neural networks ,with various parameter combinations.
Additionally, better handling of NaN values and considering more regressors could provide deeper insights and enhance model performance.

THANKS!