Trabalho - cadeira de Séries Temporais

Parte 2 - EDA série temporal

Pedro Miguel Sousa Magalhães

2021-11-10

Contents

Summary:

The time series analysed contained a subset of the total number of daily calls received by a customer support call center from 2017-01-01 untill 2018-12-31. From a preliminary exploration we can conclude that the data shows a strong upward trend with a positive shift towards the end of the series. Making use of informal / visual methods we verified that the series is heteroscedastic and contains outliers. Upon cleaning and transforming the series, detrending using first difference and a linear regression (polynomial of 2), the ACF showed signs of a weekly seasonality and eventually monthly suggesting that study using different aggregation ("week", "month") might be needed.

Dataset description and initial transformations

This report explores data containing daily number of calls received by a multinational customer support center. Data extracted from real world operations and includes observations since 2017-01-01 until 2018-11-30. No further contextual information is know or provided which could impact the analysis of the present time series. No seasonality is inherent to the business and product for which the call center provides support. It is know that the number of countries/regions covered increased during the interval in analysis.

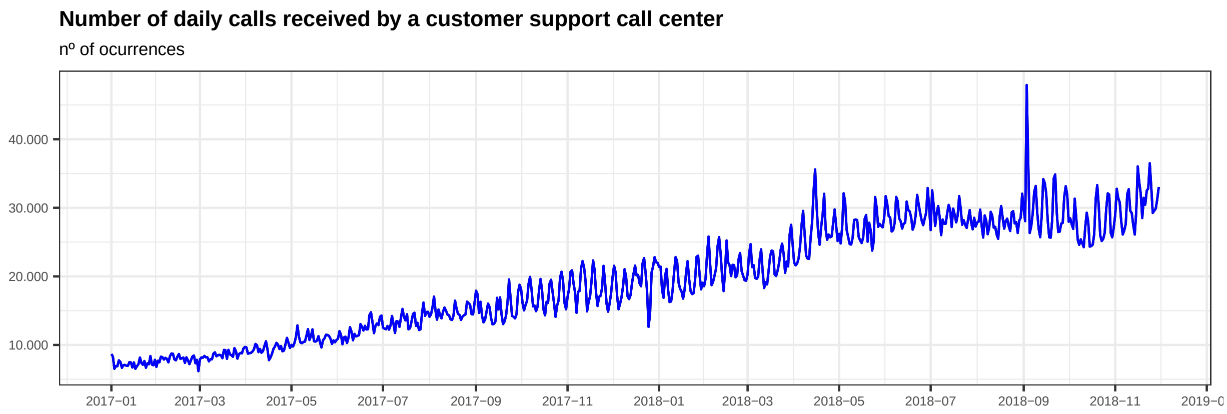


Figure 1: timeseries plot of daily number of calls

From a quick look at *Figure 1* some elements stand out:

- the data shows a clear upward trend with a positive linear relationship despite a upward shift around May 2018.
- the variance increases as time progresses leading to the conclusion that the random variable is **heteroscedastic** (using only visual process).
- the dataset contains a clear outliers specially around April and September 2018.

Dealing with outliers

Outliers can affect the outcome of the analysis and impact the detrending process. Given that no information about operation was given which could help identify a outlier or better characterize then, they were removed using statistical methods. For this analysis the `tsoutlier()` and `tsclean()` functions of the `forecast` package was used to identify outliers and replace then using linear interpolation. *Figure 2* shows the impact of this transformation.

```
##            total_calls
## 2018-04-15       35607
## 2018-09-03       47906
## 2018-09-04       36738
```

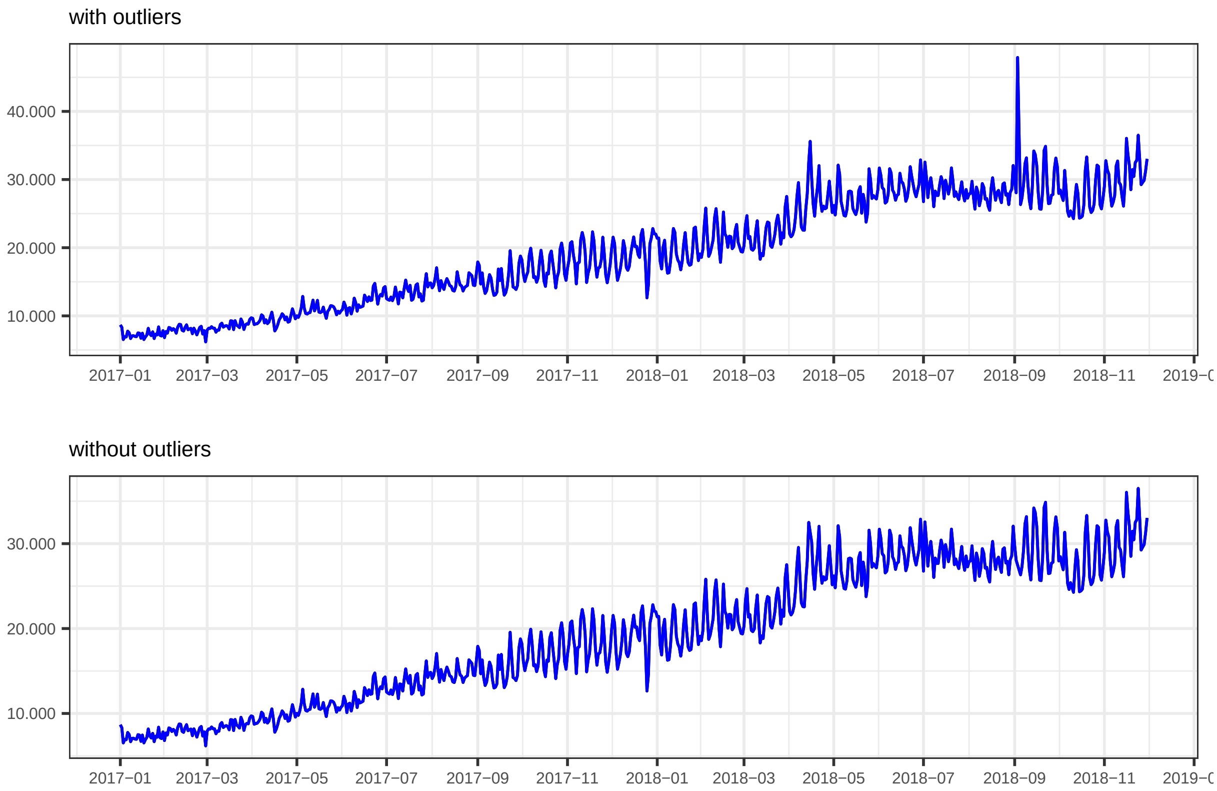
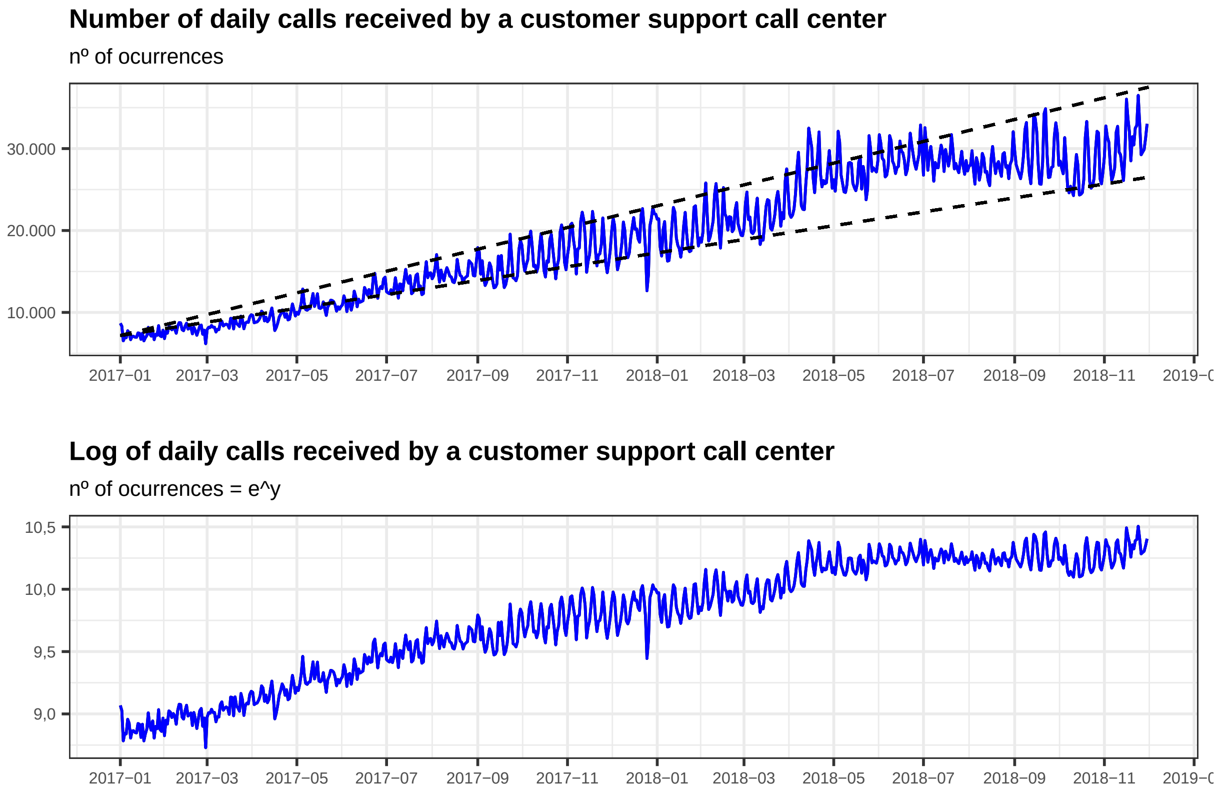


Figure 2: effect of removing outliers

Dealing with heterostacity

The assumption of homoscedasticity is key in many classical models (linear regression) and impacts the analysis. There is limited information for understanding the causes for the increase variation although, as was pointed out in the beginning, the increase number of regions might help explain this behavior (as number of location increases so does the difference between quiet and busy days).

Since the optimal lambda for BoxCox transformation is close to zero (estimated using Guerrero's method and returning -0.22578) a simple log transformation was used for simplicity. *Figure 3* illustrates the impact of this transformation.


Number of daily calls received by a customer support call center

nº of ocurrences

Log of daily calls received by a customer support call center

nº of ocurrences = e^y

Figure 3: log transform

Trend

As it is easily apparent from the plots above, the series under study presents a regular pattern with increasing variance. On one hand it suggests that a weekly season might exist but in other hand it makes it harder to analyse the underlying longer trend upward.

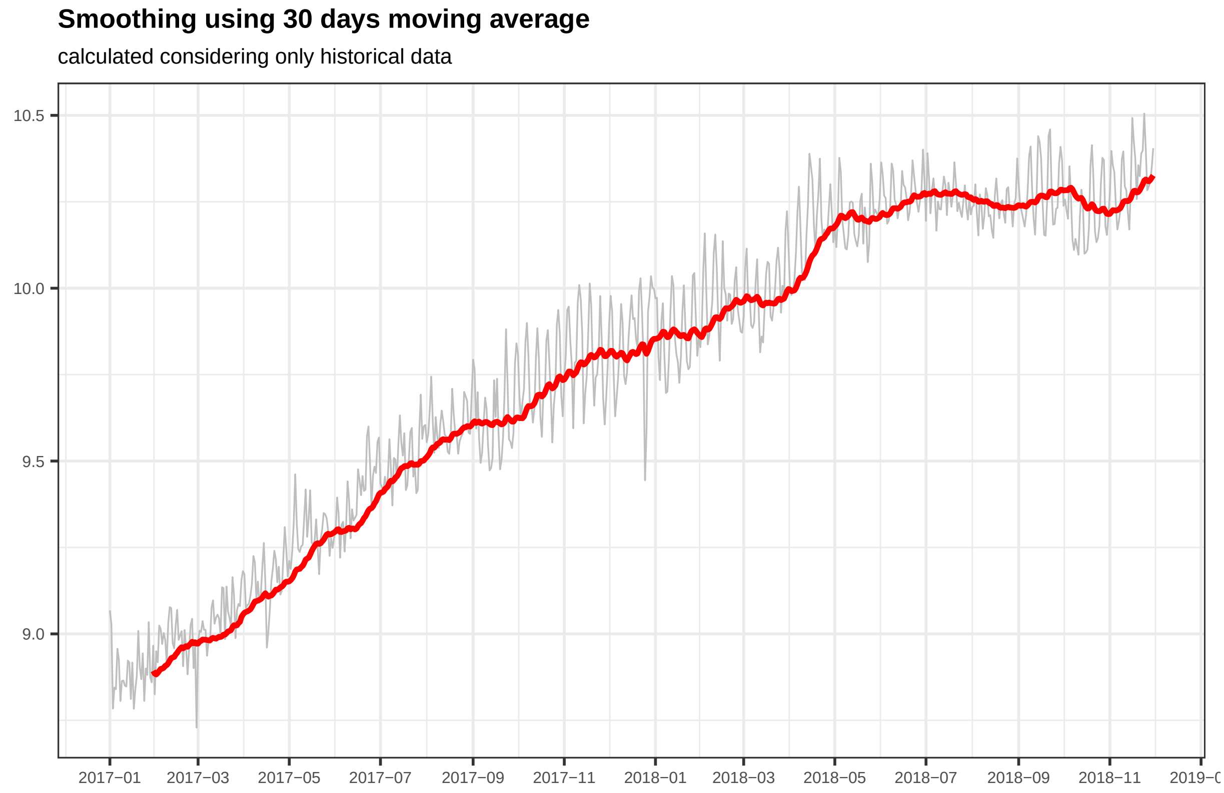


Figure 4: 30 days smoothing

Using the 30 days moving average (∼ 30 days before) the underlying trend becomes more apparent *Figure 4.*. Despite a shift or acceleration around April/May 2018, the trend seems to follow a quadratic function more than a linear relation given the apparent plateau June and October.

Detrending using first difference

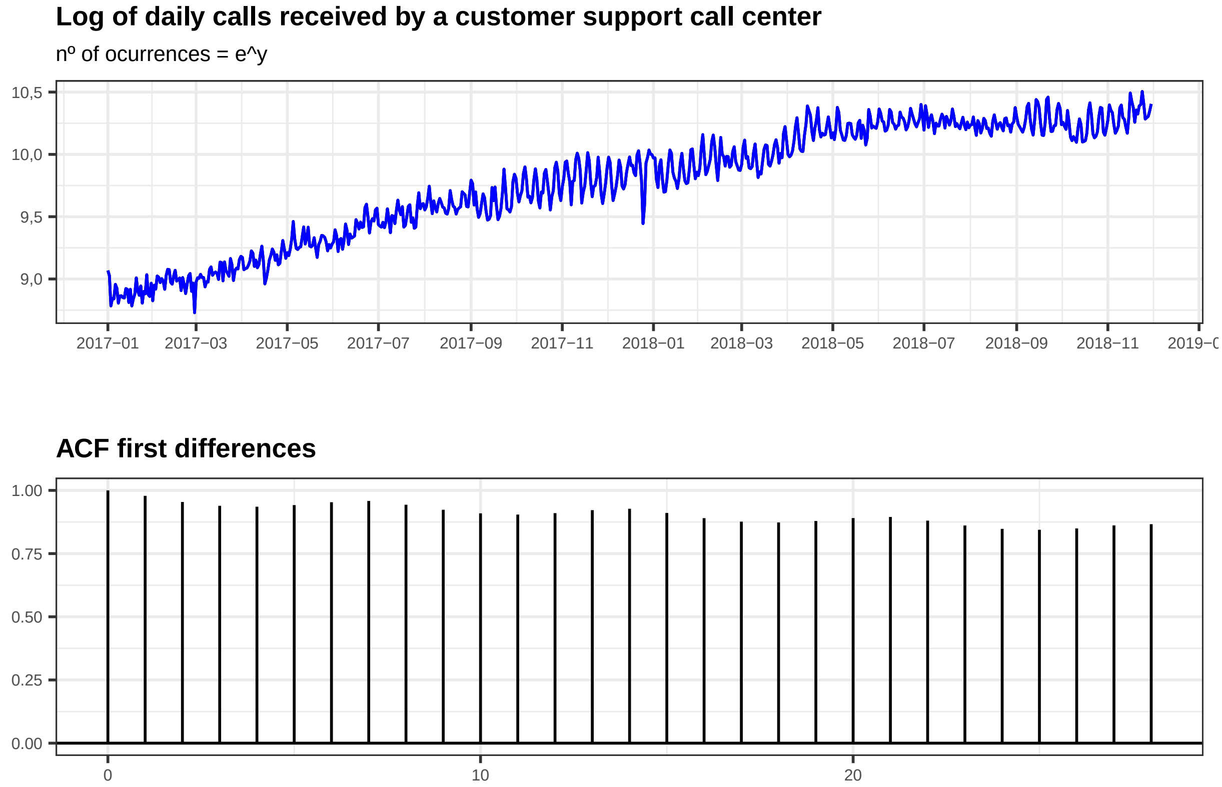


Figure 6: series ACF

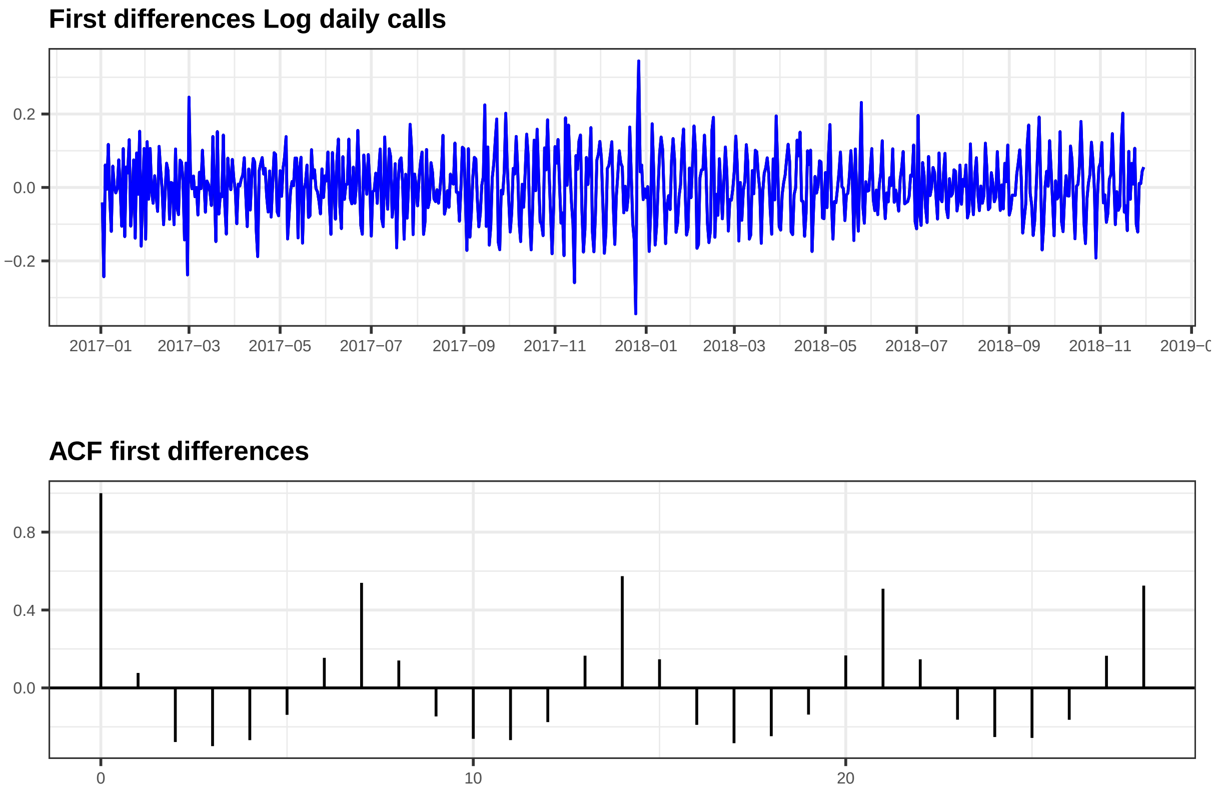


Figure 7: first differences ACF

From *figures 6 and 7* it is can be concluded that a weekly seasonality exists in this time series with the ACF not showing a rapid convergence over time

Detrending fitting a model

As stated before, the trend seems to follow a quadratic function, suspicion which can be better stated on *Figure 7.* The conclusion from this method are in line with the first differences.

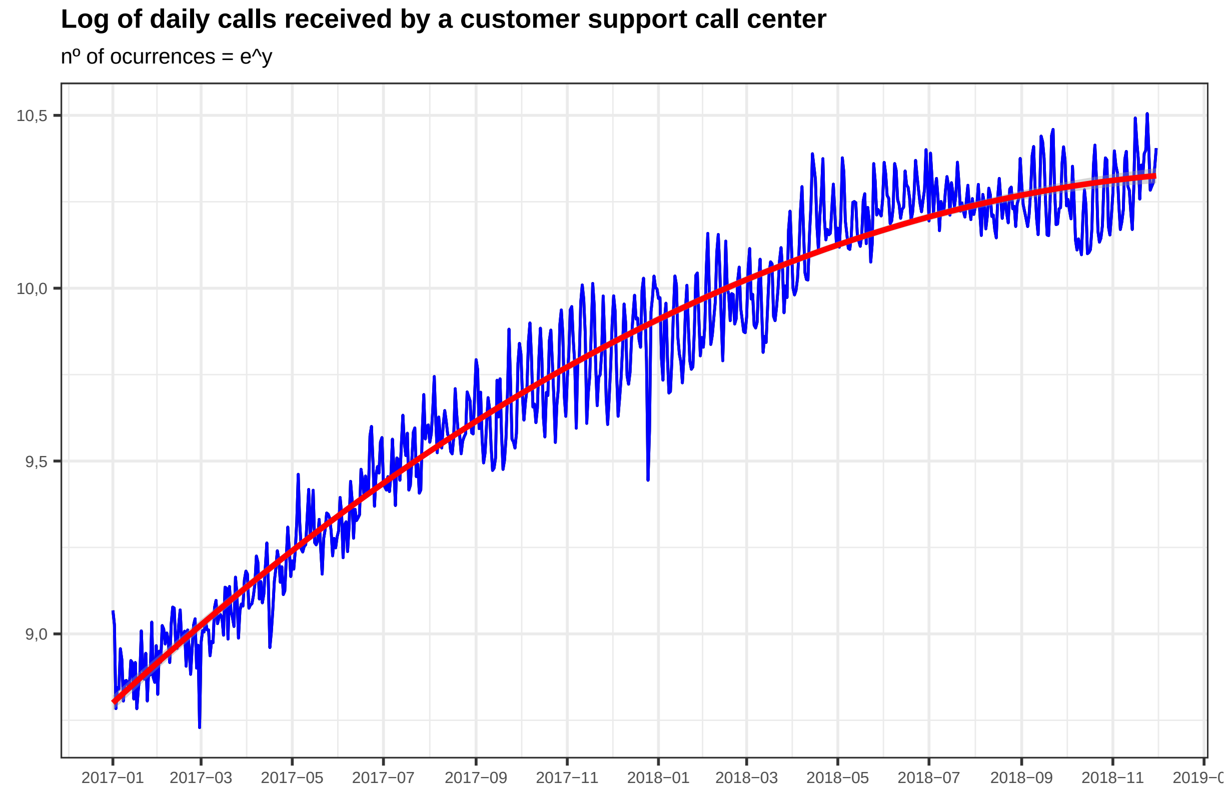
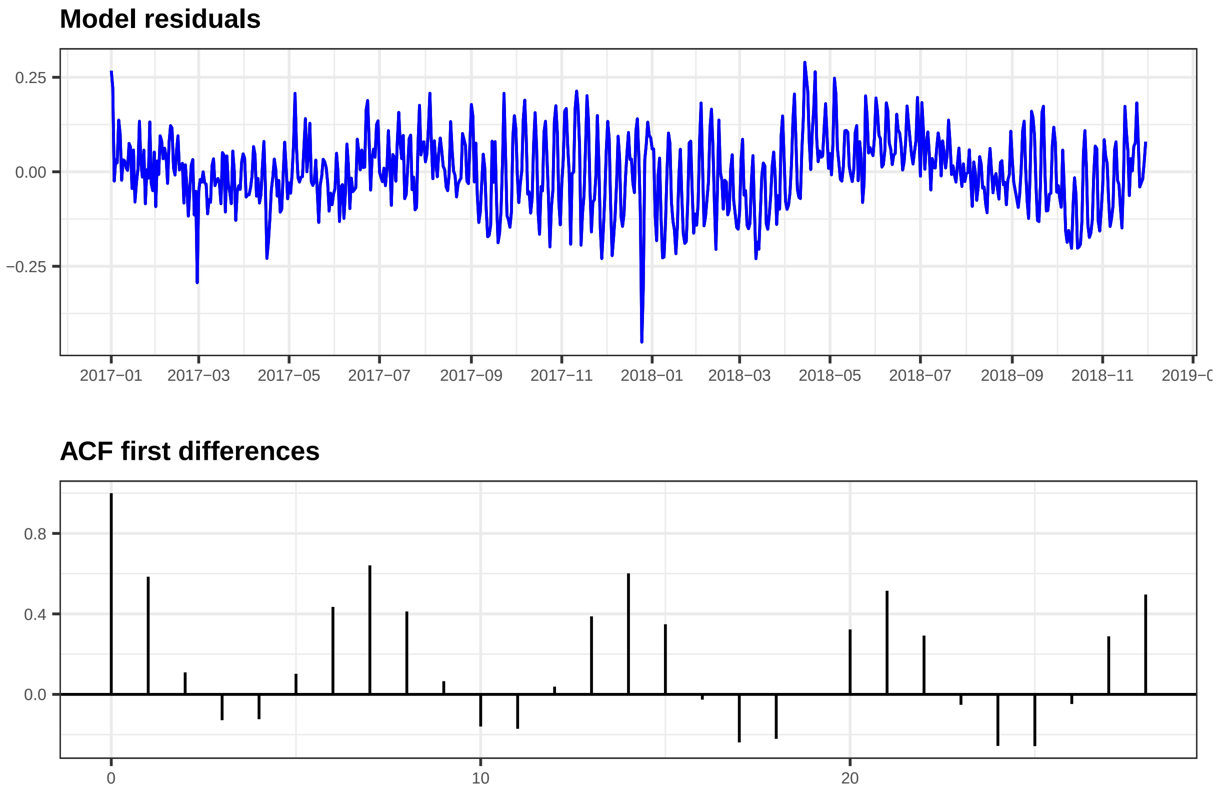


Figure 8: fitted quadratic function



Model residuals

ACF first differences

Figure 9: model residuals ACF