CIE4831 - Empirical Analyses of Transport and Planning (2021-2022)

Time Series Analyses

David Matheus (5242223)

TU Delft

Contents

1

Introduction

Ideally, when analyzing a data set an analyst can rely on it having favorable characteristics to perform easy analysis on the data and elaborate forecasts. These characteristics relate to statistical properties, such as the mean, variance, co-variance, and autocorrelation, which enable the analyst to characterize the entire process from observing a smaller set of time steps, that it, with tools such as regression.

However, not all data exhibits these properties, and more importantly, all of them simultaneously, a phenomenon known as stationarity. With non-stationary data, traditional tools do not provide a reliable analysis and forecast, for which an analyst is forced to use more complex time series analyses.

In this report, a data set from the Rijkswaterstaat is used and described in Chapter 2, the time series is checked for stationarity and the relevant analyses are performed in Chapter 3, and lastly some conclusions are drawn in Chapter 4.

2

Data Set

In this chapter a characterization of the data set will be given, as well as a description of the steps taken to make the data set usable for time series analysis for this assignment.

2.1. Data Description

For the purposes of this assignment, publicly available data from the Rijkswaterstaat has been used. As this is a specialized government entity which is a reputable source, it can be expected that the data is exhaustively and thoroughly collected. This particular data set pertains information about traffic jams on the main road network of the Netherlands, and it includes from every month since January 2015, information about the beginning and end time of a traffic jam, as well as its duration, location in the network measured as a distance from various reference points, direction, length, and its identified cause. The data can be downloaded from the Rijkswaterstaat's Data Register.

Since the data includes individual instances of traffic jams that are separated by month, the traffic jams can be aggregated per month, obtaining the monthly number of traffic jams, and then it can be used as a time series.

2.2. Data Preparation

The data is fragmented into several files, one for each month for which there is available data. Moreover, it is not intuitive how to download the data. A little bit of scraping is necessary to obtain all the data files. Other than that, however, there are no big problems with the data. It appears that the Rijkswaterstaat already put some work and attention into making it usable.

The data set is adequately formatted in a way that can be easily processed with a computer, and there are no missing entries. It is worth noting that the way in which missing entries for the traffic jam causes (there is no extraordinary cause) are dealt with is by assuming that it corresponds to high traffic intensity, which is then filled in the empty entry. This also seems to be the most comprehensive data set for traffic jams in the country, which means that it is not possible to verify with alternate data whether or not there are traffic jams in main roads that were not present in the data, so it is assumed that the data is complete in this regard.

Because of COVID-19 pandemic-related disruptions, it can be expected that data from 2020 until at least the present data can be significantly different from the previous trend, so for this reason that data has been excluded, as it is not useful to understand the previous data. Also, to prevent any anomalies from showing up in the data, only traffic jams caused by high intensity were considered, which means that traffic jams caused by traffic accidents, weather conditions, or other causes were excluded from the data set, which resulted in a total of 8095 traffic jam entries. This data was aggregated per month, obtaining a monthly count, and then only the columns that correspond to the count, the year, and the month were kept, as that is the relevant data for the time series. The result is monthly data from 2015 to 2019, which corresponds to 60 points in the time series, from 5 yearly cycles, which should b sufficient to run the analyses required. In Figure 2.1 the time series is graphically represented.

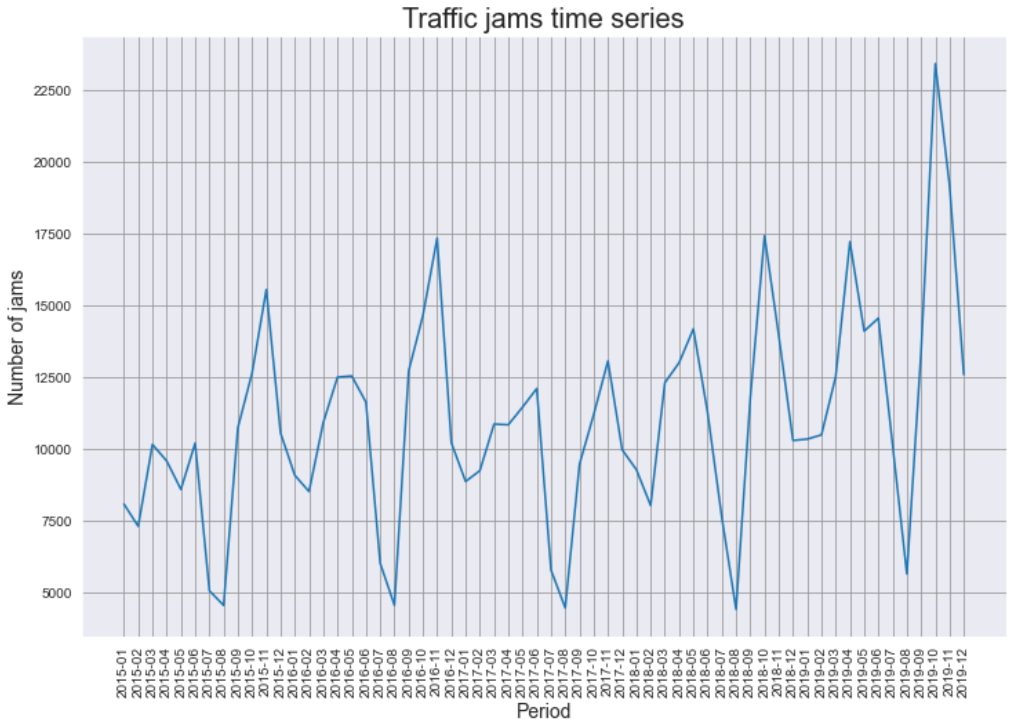


Figure 2.1: Time series of traffic jams per monthly period.

3

Time Series Analyses

In this section different analyses are performed on the time series. In every section a different method is applied, and the method description and results are shown.

3.1. Stationarity Check

With the simple visualization of the data in Figure 2.1, we can already see that the data does not look stationary, that is, that its statistical properties vary over time. For starters, even if the mean of the data seems to be constant over the first few years, by the end of the time series it appears to have an upward trend. A similar thing can be said about the variance, where even if it would appear that the data is moving between roughly the same two points, we see that the difference between the maximum and minimum in the first year is slightly lower, while in the last year it is higher.

More importantly, we can already appreciate some pattern being repeated in every yearly cycle, and in Figure 3.1 the range of the data for every month across different years can be evidenced, showing the appearance of a pattern that peaks around April and October, and falls around February and August. This yearly cycle, or seasonality, makes the time series definitely not stationary.

3.2. Smoothing

Because the data is not stationary, series analyses are necessary to understand and forecast the time series. In this case a single exponential smoothing has been performed, which would allow to flatten the series a little, without losing too much detail, by calculating a smoothed value that takes into account the past values. The importance of the past values in comparison to the most recent values in the smoothed curve is given by a smoothing factor $0 < \alpha < 1$. The smoothing process is described by equation 3.1.

$$S_t = \begin{cases} X_1 & t = 1 \\ \alpha \cdot X_t + (1 - \alpha) \cdot S_{t-1} & t > 1 \end{cases} \tag{3.1}$$

The result of applying this method with a smoothing factor $\alpha = 0.5$ can be appreciated in Figure 3.2, where the original curve X_t and the smoothed curve S_t are represented for each time step t. There it can be seen that the noise in the data has mostly been dealt with, making the smoothed curve easier to analyze.

3.3. Trend detection

The data can also have exhibit a broader trend, and to detect this trend the method of double exponential smoothing is suitable. This method, besides smoothing the curve based on past values, also accounts for these trends. Double exponential smoothing uses a smoothing factor $0 < \alpha < 1$ and a trend smoothing factor $0 < \beta < 1$. This process is described by equations 3.2 and 3.3.

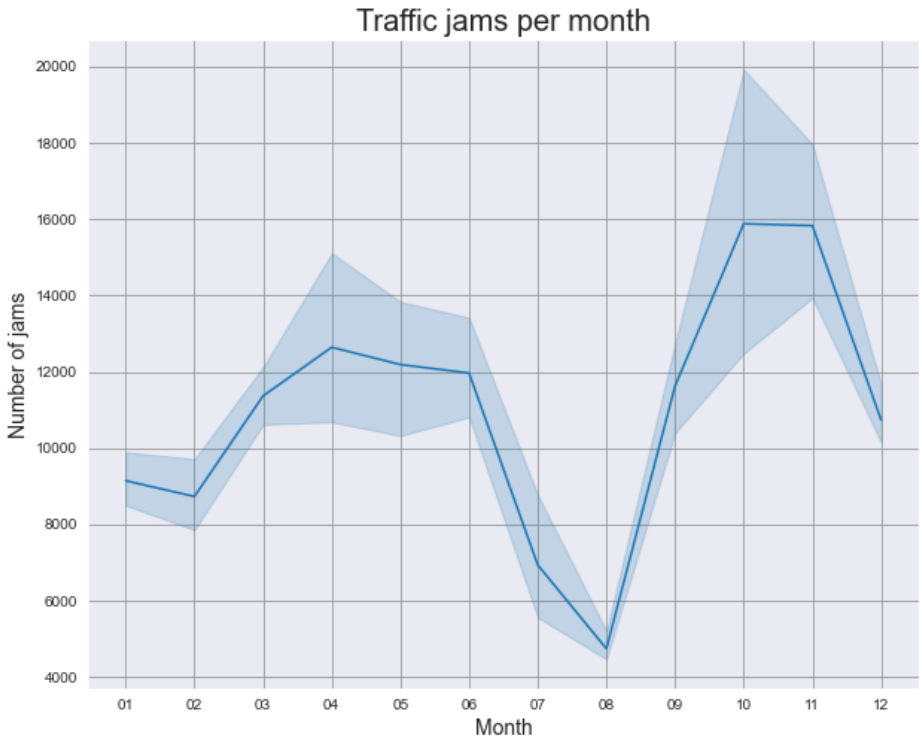


Figure 3.1: Visual representation of the yearly cycle. It can be seen that the data follows a very similar pattern every year.

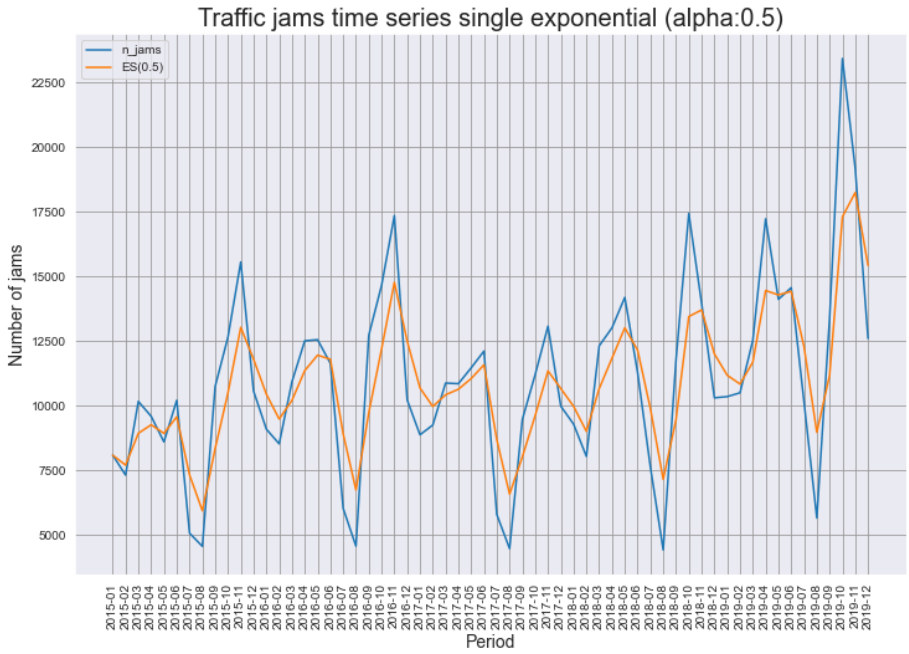


Figure 3.2: Plot of the original time series and the smoothed time series.

$$S_t = \begin{cases} X_1 & t = 1 \\ \alpha \cdot X_t + (1-\alpha) \cdot (S_{t-1} + b_{t-1}) & t > 1 \end{cases} \tag{3.2}$$

$$b_t = \begin{cases} X_1 - X_0 & t = 1 \\ \beta \cdot (S_t - S_{t-1}) + (1-\beta) \cdot b_{t-1} & t > 1 \end{cases} \tag{3.3}$$

The result of applying this method with a smoothing factor $\alpha = 0.5$ and a trend smoothing factor $\beta = 0.3$ can be appreciated in Figure 3.3, where the original curve X_t, the smoothed curve S_t, and the trend b_t are represented for each time step t. There it can be appreciated that not only the noise in the data has been dealt with, but that the general trend over time can also be appreciated.

Figure 3.3: Plot of the original time series, the smoothed time series, and the trend over time.

3.4. Seasonality detection

Lastly, the data can also exhibit seasonality, or a cyclical behavior that repeats over time, which as we saw in Figure 3.1 appears to be the case in this data set. To detect and analyze this seasonality the method of triple exponential smoothing is suitable. This method, besides smoothing the curve based on past values, and accounting for broader trends, also reveals and compensates for the seasonal variation. Triple exponential smoothing uses a smoothing factor $0 < \alpha < 1$, a trend smoothing factor $0 < \beta < 1$, and a seasonal smoothing factor $0 < \gamma < 1$. With a cycle length L and a number of full cycle in the data N, this process is described by equations 3.4, 3.5, and 3.6.

$$S_t = \begin{cases} X_t & t \leq L \\ \alpha \cdot \frac{X_t}{c_{t-L}} + (1-\alpha) \cdot (S_{t-1} + bt - 1) & t > L \end{cases} \tag{3.4}$$

$$b_t = \begin{cases} \frac{1}{L}\sum_{i=1}^{L} \frac{X_{L+i} - Xi}{L} & t = 1 \\ \beta \cdot (S_t - S_{t-1}) + (1-\beta) \cdot b_{t-1} & t > 1 \end{cases} \tag{3.5}$$

$$c_t = \begin{cases} \frac{1}{N}\sum_{j=1}^{N} \frac{X_{L(j-1)+t}}{A_j} & t \leq L \\ \gamma \frac{X_t}{S_t} + (1-\gamma) \cdot c_{t-L} & t > L \end{cases} \tag{3.6}$$

where

$$A_j = \frac{\sum_{i=1}^{L} X_{L(j-1)+i}}{L} \qquad j \leq N \tag{3.7}$$

The result of applying this method with a smoothing factor $\alpha = 0.4$, a trend smoothing factor $\beta = 0.2$, and a seasonal smoothing factor $\gamma = 0.2$ can be appreciated in Figure 3.4, where the original curve X_t, the smoothed curve S_t, the trend b_t, and the seasonality c_t are represented for each time step t. In this case, the noise in the data has been eliminated, the general trends over time identified, and the pattern of seasonality is exposed, leaving a smoothed curve that more closely approximates stationarity. In this case, the curve really flattened, suggesting that seasonality was the main driving factor in the behavior shown in the data.

3.5. Notes on the choice of smoothing factor values

For the purposes of this assignment, the values for the different smoothing factors were chosen by trial and error, testing different values and ultimately choosing the ones that would smooth the curve without compromising the information given by it. This process is cumbersome, especially when having to test multiple factors simultaneously.

This process can be made more efficiently with optimization solvers on a computer, by minimizing the error of the forecast against available data. However, given that the solver will have many degrees of freedom to ”fit the curve” (three factors in the case of triple exponential smoothing), the analyst must be wary of overfitting the curve, which is counterproductive for the reliability of any forecast.

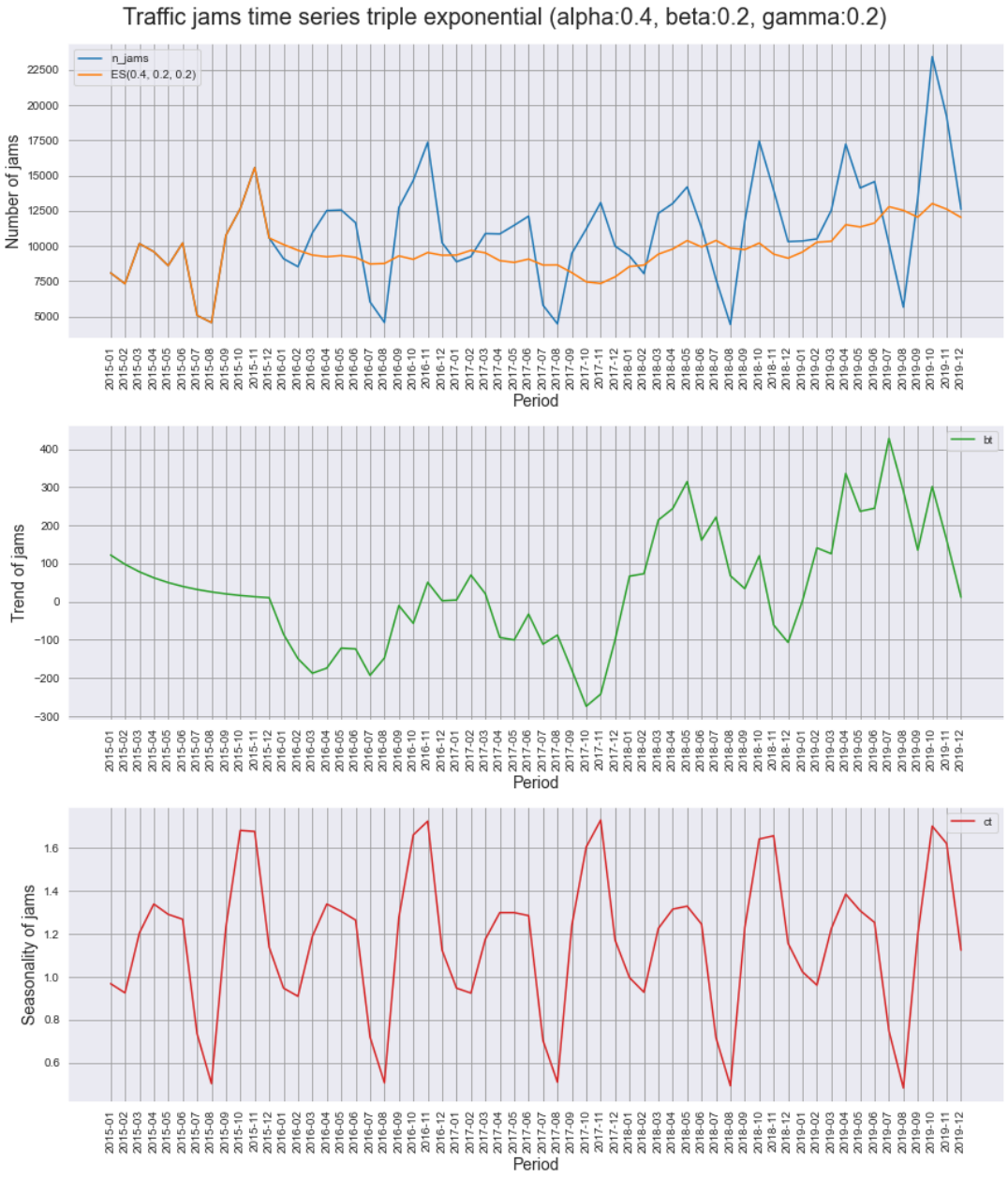


Figure 3.4: Plot of the original time series, the smoothed time series, the trend over time, and the seasonality.

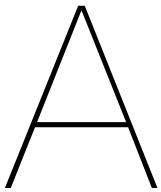
4

Conclusion

In this report, data set is used whose properties (non-stationarity) prevent an analyst from providing any reasonable analysis and forecast extracted from traditional statistical tools such as regression. Instead, there are tools available that allow an analyst deal with noise, trends, and seasonality in the data.

These methods have been performed on the time series from the Rijkswaterstaat data set in Chapter 3, and the different components of the data have been isolated so that it becomes possible now to better understand the data and perform forecasting for future time periods. However, the difficulties in these methods, such as setting reliable smoothing values, have also been highlighted.

In the case of this data set, using this type of analyses can be valuable to estimate the impact of the COVID-19 pandemic. As stated in Chapter 2, the data from 2020 onward was excluded from the time series because it can be expected that the pandemic disrupted the patterns previously observed, but using that data and comparing it with a forecast made using these analysis techniques the expected behavior (the forecast) and the real behavior (the data) of the curve can be compared, thus highlighting the usefulness of time series analysis.



Public repository

For transparency and reproducibility purposes, the original data, clean and sampled data, and the Jupyter Notebook file with which this research was done have been uploaded to an online Github repository and made available to the public. To see the repository please go here or copy the url in your browser: https://github.com/davidmatheus002/assignment_3.git