RESEARCH

Introduction to the profile areas of data sciences: project 6

Eva Aßmann* and Paul Vogler

*Correspondence:
eva.assmann@fu-berlin.de
Institute of Mathematics and Informatics, Berlin, Takustr. 9, Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Goal of the project: The aim of the project was to analyze two sets of microbial time series data with the prophet library and make predictions about the microbial constitution in the body over time.

Main results of the project: The time series prediction stayed very linear for both analyzed sets and showed not much variation for the prediction.

Personal key learnings: The key learning was that there are extensive libraries to analyze data over a time period and what can be done to improve predictions about the time series future.

Estimated working hours: 9

Project evaluation: 2

Number of words: 825

Suggestions for improvements of the project:

Keywords: sample; article; author

Background and goal of the project

In this project the prophet library which was developed by facebook was applied in order to analyze and forecast time series data. The data was taken from the microbiome study by Caporaso et al. ("Moving pictures of the human microbiome"). Within the scope of this study, the variation of the microbial constitution in the human body was analysed for two individuals (male and female), four body sites and over 396 timepoints. The obtained insights were considered to support the development of treatments for microbiome-related afflictions like obesity, Crohn's disease, inflammatory bowel disease and malnutrition [1].

1 Description of the data

The time series data frame comprised 1967 samples and 22765 columns containing the counts of all detected OTUs per sample. Operational taxonomic units (OTUs) are groups of organisms that are categorised based on specific feature similarities, mostly sequence-based features. Additionally, a data frame with metadata on each sample was retrieved containing information on the study individuals, the sample collection procedure as well as the conditions and outputs of the sequencing experiment. For the purpose of time series analysis with the prophet library, only the timestamps of sample collection were of interest. Two OTU columns were chosen manually (OTUs 4479944 and 11544) from the time series data frame and were

and merged with the collection timestamps from the metadata based on the Sample identifications. For each OTU, a data frame was created to be compatible with prophet. Each OTU data frame contained the collection timestamps in datetime format (column name 'ds') and the OTU counts (column name 'y') per sample (See Tab. 1).

OTU	OTU counts	collection timestamps
4479944	min=0, max=12	2008-10-21 to 2010-01-08
11544	min=0, max=1	2008-01-21 to 2010-01-08

Table 1: Content of the two OTU data frames.

2 Results

2.1 Visualize the time-series data

For the two chosen OTU's, the count values were plotted against the timepoints for an overview (Fig. 1 and 2).

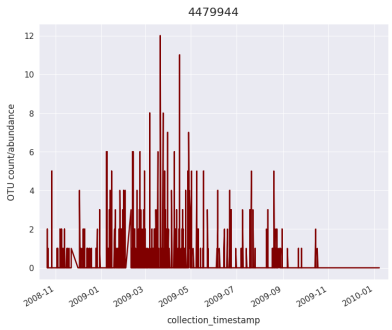

Figure 1: OTU 4479944 values vs timepoints.

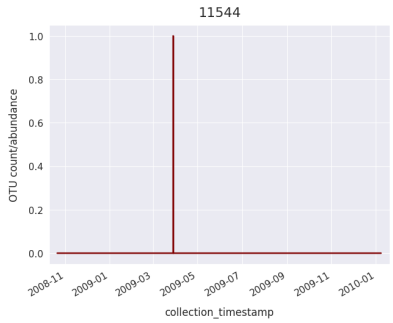

Figure 2: OTU 11544 values vs timepoints.

2.2 Compute, visualize and briefly discuss the components of the time-series

For both OTU's, a prophet prediction model was fitted and OTU counts over time were predicted for one year in the future (Fig. 3 and 4). Also the overall, yearly, weekly and daily trend where plotted (Fig. 5 and 6). The overall trend was observed to be decreasing for both OTU's. The yearly trend showed high fluctuation in the first months of the year for both OTU time series. Also both OTUs showed higher counts at night and during the weeken when looking at the weekly and daily trend.

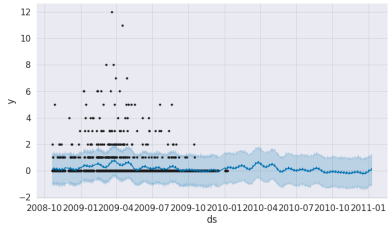

Figure 3: OTU 4479944 prediction.

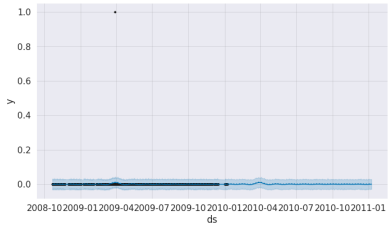

Figure 4: OTU 11544 prediction.

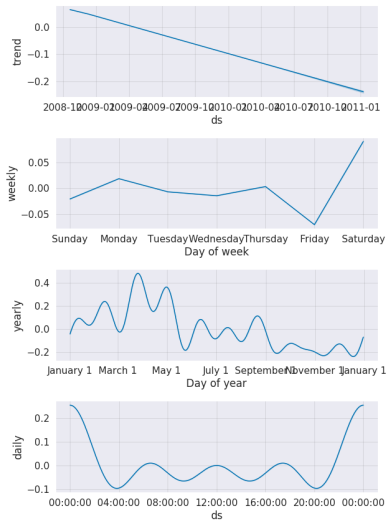

Figure 5: OTU 4479944 trends.

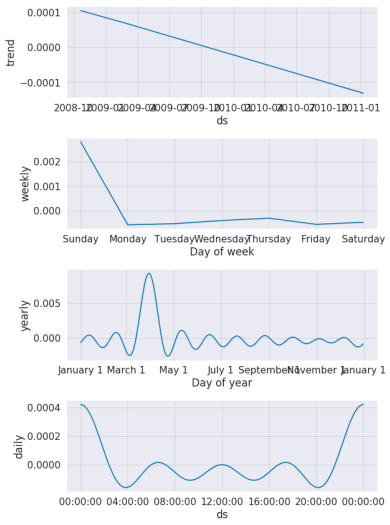

Figure 6: OTU 11544 trends.

2.3 Include an uncertainty visualisation into the forecasting analysis

The uncertainty in the overall trend and seasonality estimates as well as some observational noise were added to the forecasting analysis (Fig. 7 and 8). The overall trend showed to be completely uncertain, whereas the seasonal trends were observed to run reasonably within the shape of the uncertainty boundaries for both the OTU's.

2.4 Compare the results from the additive seasonalities to multiplicative seasonalities

Prophet fits additive seasonalities, meaning the seasonality effect is added to the trend in order to get the forecast. If seasonality is not a constant additive factor, but rather grows with the trend, it is multiplicative. For both OTUs a forecasting analysis fitting additive and multiplicative seasonalities was fitted, respectively (Fig. 9-10). For OTU 4479944 it was observed that using multiplicative seasonalities made the predicted time series data decrease more strongly than with additive seasonalities. Fitted multiplicative seasonalities showed higher uncertainty for OTU 1154.

2.5 Perform a change-point analysis and visualize detected change-points

Prophet automatically detects historical changepoints and adapts the trend appropriately. To avoid overfitting, changepoints are inferred only for the first 80% of

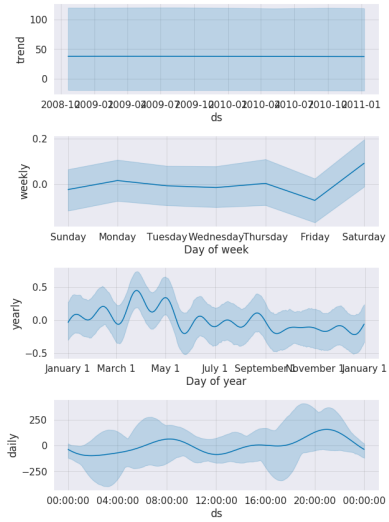

Figure 7: OTU 4479944 trends with an uncertainty visualization.

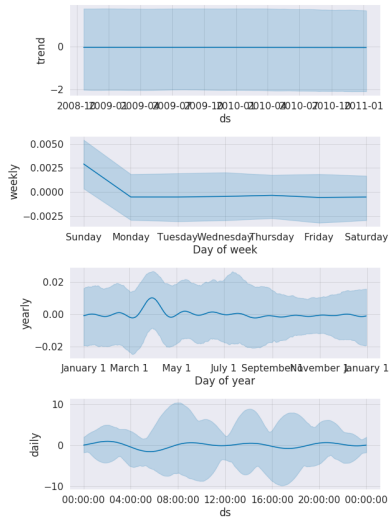

Figure 8: OTU 11544 trends with an uncertainty visualization.

the fitted time series data. Changepoint analysis was computed for both OTU time series. The analysis resulted in no changepoints for both timeseries, although the changepoint plot as well as the initial visualisation of the data for OTU 4479944 suggested multiple changepoints (Fig. 11 and 12).

2.6 Can the change points be explained in a meaningful way?

Because the change points were non-existent, there was also no meaningful information to get from this analysis.

2.7 Check whether the data contains outliers

For the points of OTU 4479944 it was not clear which points are outliers, becausethe curve doesnt resemble most of the points, for OTU 11544 there was only oneoutlier point that didn't impact the overall prediction.

3 Discussion

This was a typical project for a datascientist, because we preprocessed a dataset and added meta information for further analysis, we also analyzed a subset of the data to look at the time series evolution. The trends extracted from the time series were used to make predictions about future events.

References

[1] Caporaso, J.G., Lauber, C.L., Costello, E.K. et al. Moving pictures of the human microbiome. Genome Biol 12, R50 (2011) doi:10.1186/gb-2011-12-5-r50

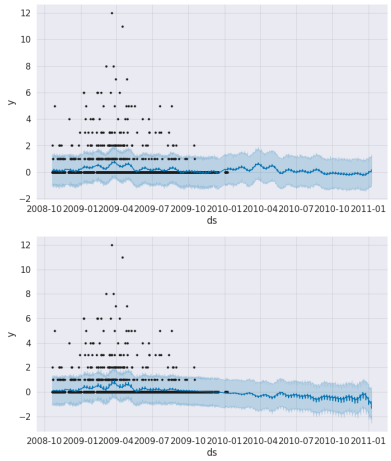

Figure 9: OTU 4479944 additive (top) and multiplicative trend (bottom).

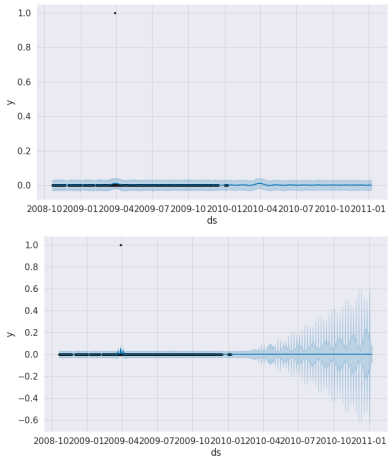

Figure 10: OTU 11544 additive (top) multiplicative trend (bottom).

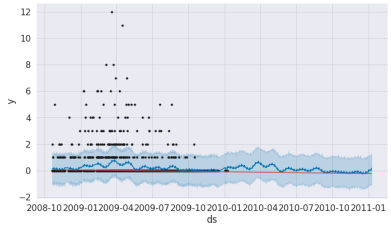

Figure 11: OTU 4479944 change-points.

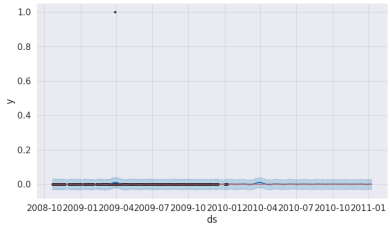

Figure 12: OTU 11544 change-points.