Waterschap Aa en Maas

Safe
Sufficient
Clean
Water

Steady the Flow

Analysis of Den Bosch Waste Water Treatment Plant

Data Challenge 3 - Group 2

Objectives

The latest technological developments have opened up new possibilities for controlling the flow of water around the WWTP of Den Bosch in order to optimize the water cleaning process. This project focusses on one specific pump; Haarsteeg. The main objective is to predict the flow of waste water in m^3 and the amount of rainfall in mm for the next 24 hours. Therefore, the addressed research question in this poster is: *'Is it possible to accurately predict flow (in m^3/h) and rainfall (in mm) for the next 24 hours of the Haarsteeg pump?'*

Available Variables

Flow	Level	Rain
• Timestamp • Flow Rate (m^3/h) • Data Quality	• Timestamp • Relative level change in m • Data Quality	• Actual rain in mm • Predicted rain in mm • Surface area
Minute - based	Minute - based	5 minute and hourly Shapefiles and KNMI data

Data Understanding

Figures 1-5 give an initial overview of the data to get a better understanding of the existing patterns, the current gaps and any other data features that need to be taken into consideration for model development.

Overview of the Data

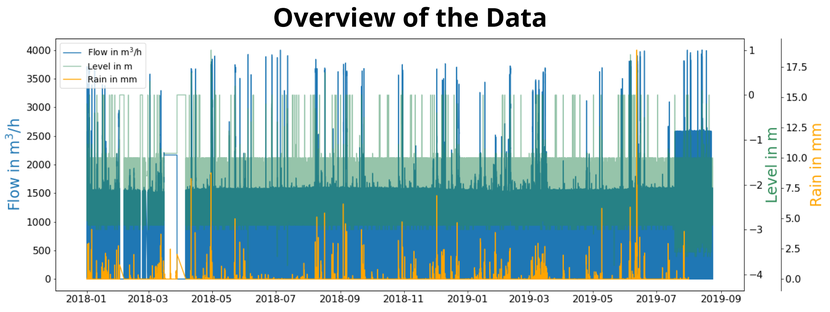


Figure 1: Overview of the Flow (in m^3), Relative level change (in m) and amount of rainfall in mm.

Figure 1 gives a first impression of what the data looks like. It clearly shows two gaps in both the flow and level data. In addition, at the end of the period there is a large increase in both level and flow data. This is due to some experimental settings in the Haarsteeg pump and should be discarded in the final dataset.

Closer look at the data - December

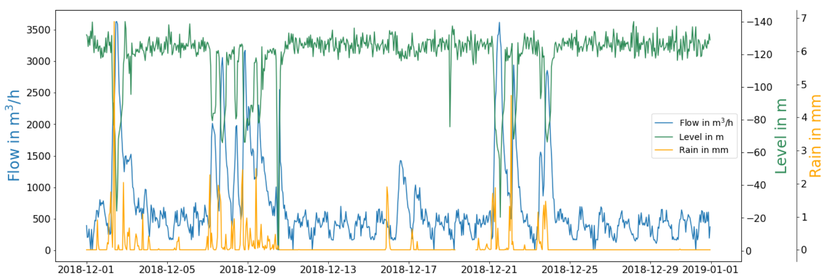
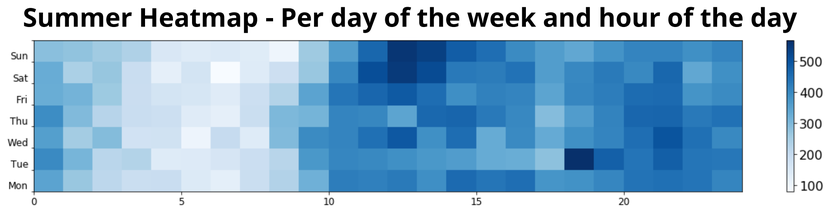


Figure 2: Overview of December, showing the relation between flow (m^3) relative change of level (m) and rainfall (mm).

Figure 2 shows that an increase of rain is followed by an increase of flow and a decrease of level change. In the absence of rain, flow and level are very steady. This means, that rain has a big impact on the behavior of the sewage system.

Summer Heatmap - Per day of the week and hour of the day



Winter Heatmap - Per day of the week and hour of the day

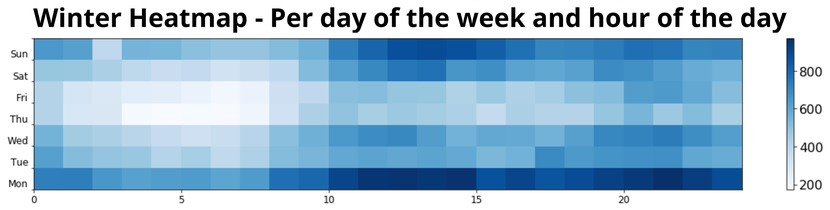


Figure 3: Seasonal heatmaps of the average daily flow (in m^3) per day of the week and hour of the day of Haarsteeg.

Figure 3 shows patterns in the day of the week between different seasons. As Autumn and Spring are similar to Summer, they have not been displayed. The Summer heatmap shows a clear difference between the weekend and week days. Both heatmaps suggest clear patterns in the data. However, the heatmaps are based on averages, therefore a closer look is necessary.

Frequency Distribution of Flow

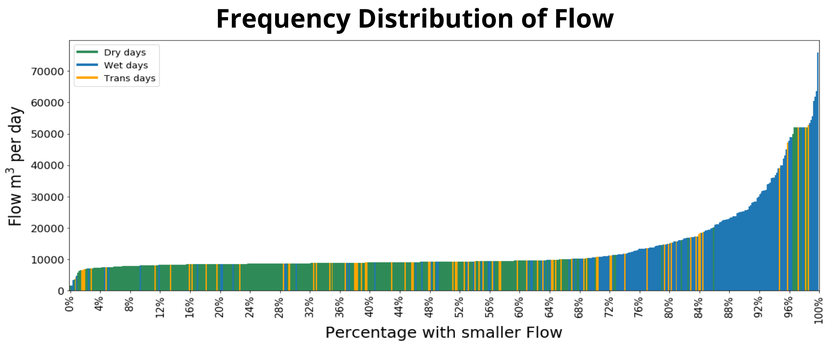


Figure 4: Frequency distribution of flow showing all days in the Haarsteeg dataset ordered based on daily flow

The dataset contains wet, dry, and trans days. On wet days it rains more than 0.5 mm, on dry days it rains less then 0.5 mm and the day before was a dry day. On trans days it rains less then 0.5 mm, but the day before was wet day. Wet days are likely to have a higher total flow than dry days. Figure 4 shows that for the majority of the days, this reasoning is correct. It also suggests that trans days behave more often like dry days then wet days.

Week vs Weekend - Dry vs Wet vs Trans - Summer

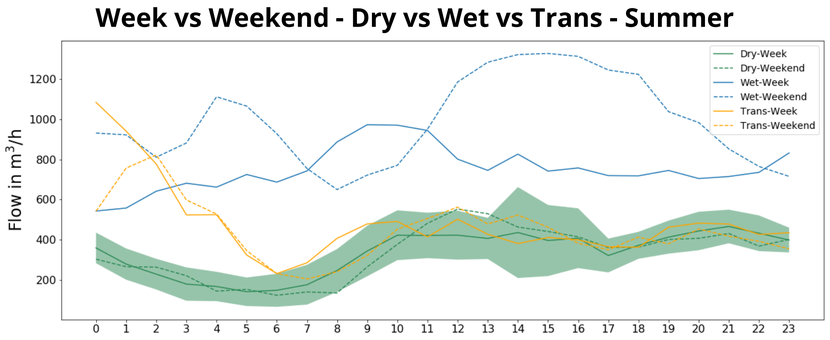


Figure 5: Patterns of average flow across Week vs Weekend and Dry vs Wet vs Trans of the Summer period

Figure 5 shows the average patterns over various variables and standard deviation of Dry - Week days in Summer. Standard deviations of the other combinations and in other seasons are too high to visualize. This suggests a more stable pattern in the Summer, meaning it is the best to predict.

Problem Breakdown

Predict flow (m^3/h) and Rainfall for the next 24

- Predict flow (m^3/h)
  - Dry Days
    - Prophet Model
    - GRU Model
  - Wet Days
    - LSTM Model
- Predict Rainfall
  - Rainfall Analysis
    - Actual vs Predicted

Prophet Model

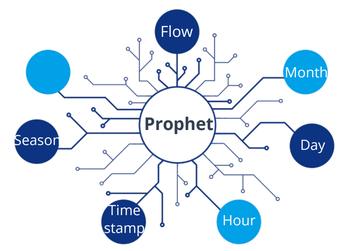
Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality. It works best with seasonal effects, is robust for outliers and easy to use.

In

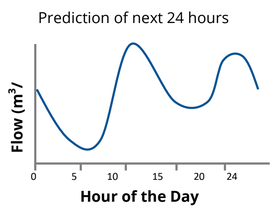
- Timestamp
- Flow (m^3/h)
- Day of the Week
- Month
- Season
- Hour of the day

Hourly data

Model



Out



Initial Model - Actual versus predicted

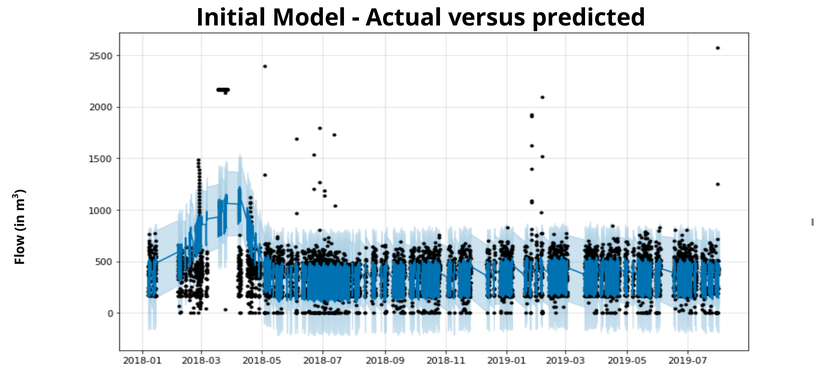


Figure 6: Overview of the results of initial model - predictions (blue line) versus actual (black dots).

Figure 6 shows an initial model for the entire dataset, indicating that there are some gaps and outliers that cause strange predictions. There might be some misclassified dry days that cause this behavior.

Actual versus predicted Flow

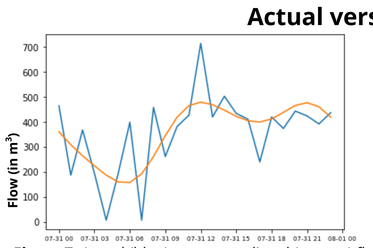


Figure 7: Actual (blue) versus predicted (orange) flow for the last day of Summer

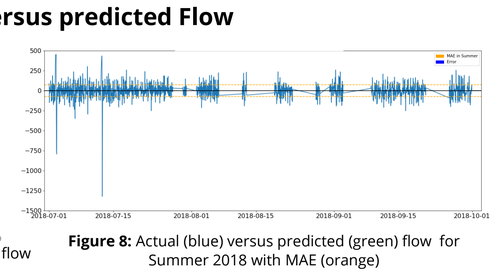


Figure 8: Actual (blue) versus predicted (green) flow for Summer 2018 with MAE (orange)

Prediction error histograms (left summer, right Autumn)

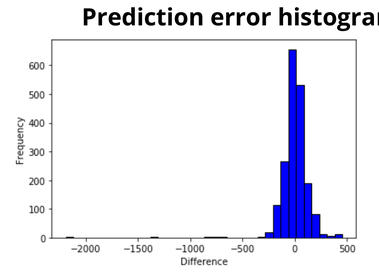


Figure 9: Histogram of prediction errors (entire Summer)

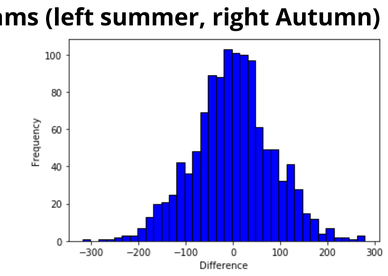


Figure 10: Histogram of prediction errors (entire Autumn)

Figures 7-10 show the predictive power of the prophet model. It has picked up on the daily seasonal trend (Figure 7). Figure 8 shows two outliers in the Summer data, making the MAE of Summer higher compared to other seasons. The MAE's vary from 65 to 82 (reasonable with average flow of 345/h). Both histograms give an overview of the error distributions. Summer is highly affected by outliers.

GRU Model

In

- Timestamp
- Flow (m^3/h)
- Day of the Week
- Month
- Season
- Hour of the day

Model

Out

One prediction per hour for each of the upcoming 24 hours

`t+1, t+2, …, t+24`

Predicted/actual flow on Dry Day (24h)

Figure 11: Difference between predicted/actual flow. Green area: overpredicted. Blue: underpredicted.

Mean Absolute Error per Hour

Figure 12: MAE per hour (over entire test set)

Model uses the same input as Prophet, as they were found significant in earlier analysis. In Figure 11, model overpredicts 365m³ over entire day. MAE shows that prediction errors of 20-24h forward aren't much different from 1-4h forward (as data is periodic). Compared to prophet, GRU is slightly less accurate.

Prediction Error Histograms (left Summer, right Autumn)

Figure 13: Histogram of prediction errors (entire Summer)

Figure 14: Histogram of prediction errors (entire Autumn)

LSTM Model

In

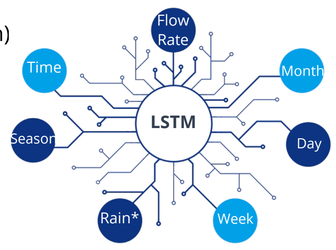
Flow

- Timestamp
- Flow Rate (m^3/h)
- Data Quality
- Minute based

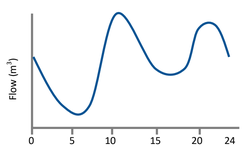
Rainfall*

- Timestamp
- Rain (mm)
- Surface Area
- Minute and hourly based

Model



Out



* Variables available for predicted and actual rainfall

Actual vs. predicted flow

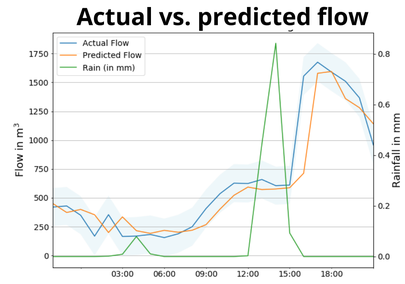


Figure 15: Actual flow versus predicted flow. Rainfall included (in green).

Prediction error

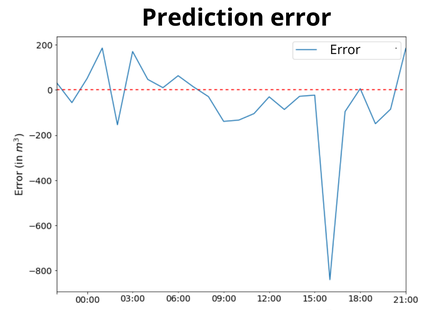


Figure 16: Prediction error of flow.

Long Short-Term Memory (LSTM) recurrent neural networks are able to model problems with multiple input variables that can be especially useful for time-series forecasting. All variables were tested, but only Date (stamp), Rain (m³), and actual flow were needed for best result.

These variables with a lag of 25 hours give the best prediction model and the lowest MAE. In Figure 16, the model predicts a rise in the sewage flow after a surge in rainfall. The graph shows the prediction for the next hour and you can see that the predicted flow predicts the actual flow pretty well but unfortunately more or less one hour too late. Therefore this model will probably not be very effective for Aa en Maas.

Rain Prediction Analysis

Day classification vs. reality

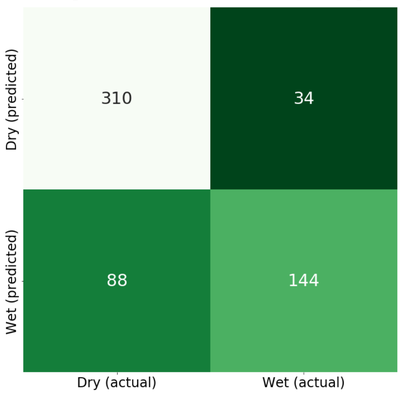


Figure 17: Heatmap of actual versus predicted dry and wet day classifications

Cumulative Absolute Prediction Error

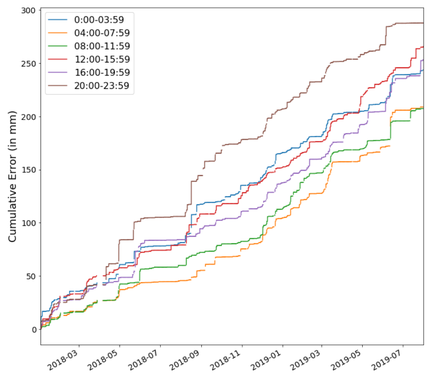


Figure 18: Cumulative prediction error over the hours of the day over the entire dataset

Figure 17 indicates that 80% of the predictions are accurate and only in 6% (34) of the cases the predictions are potentially dangerous (predicting no rain when there is rain). As all predictions are made during midnight, the assumption is that predictions later of the day are less accurate then earlier on the day. Figure 18 confirms this assumption, as the predictions from 20:00 - 23:59 are least accurate and the predictions from 04:00 - 11:59 are most accurate.

Actual versus Predicted

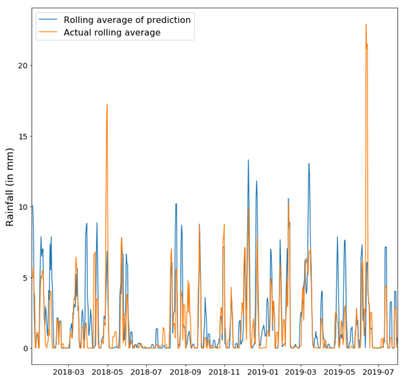


Figure 19: Predicted and actual rolling average of daily sums (3 days)

Relative Difference between Actual and Predicted Rainfall

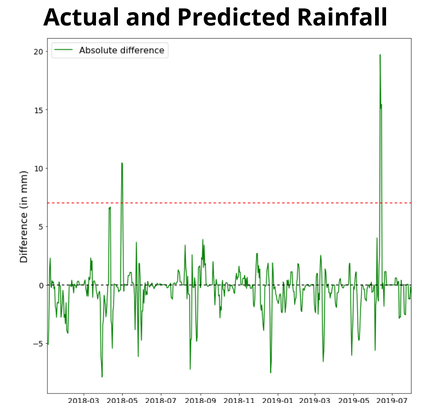


Figure 20: Difference between actual and predicted rainfall with bar set at 7 mm of rainfall.

In Figure 19, most of the time, peaks in predictions also imply peaks in the actual data and vice versa. Therefore, in terms of rain intensity the predictions are also quite accurate. There are almost no extreme under - overpredictions.

Figure 20 shows that the predictions tend to predict more rain than will actually fall, but that underestimations in the forecast are less common. However, underestimations do have higher extremes. In addition, the 7 mm rainfall line (in red) is added to show the amount of rainwater that the sewage system can store (when empty). It is essential that rain must not be underestimated by 7 mm or more, as it will cause potentially dangerous overflows if the sewage system is not emptied in time. This 'limit' is only exceeded twice in the entire dataset.

Conclusions & Recommendations

Models:

- For dry days, Prophet outperforms Gru. Use Prophet over Gru.
- Look further into how to make models more accurate, still much potential.
- For solving the problem, it would have been better to heavily focus on 1 model rather than exploring multiple models at once (as performance is similar).

Rainfall Analysis:

- Rainfall predictions tend to overestimate more rather than to underestimate.
- In (only) few occasions the rainfall is severely underestimated.
- When using a model, have an emergency system for when the actual rainfall is much higher than predicted rainfall (as pump level may rise unexpectedly). This emergency system allows for temporary manual control of the pumps.
- In future, perhaps provide rain prediction data closer to the predicted hour (which is known to exist). Then, time-related changes in accuracy can be analyzed and possibly used in a model.

TU/e Eindhoven University of Technology

J.C.M. Arts | S.A.A. Eichperger | Y.T. Hu | J. Kareem | S. Papandroudis | J. Phelipa | K.M.A.A. Schrouff

Tilburg University