Manual DCP-HTO

This document serves as a short manual for the DCP-HTO code as developed for the MDS&E master’s thesis “Outlier detection, correlation and prediction in. urban traffic flow data”.

The project contains a streamlined version of the code. All steps described in the solution section of the thesis are included here. The idea is that another project can use this code as an example and use their own input data to replace the input data for this project.

To run the SFPD-LOF model for outlier detection, go to the SFPD-LOF folder and start with the create\_fpd notebook to create the fpds. The data from the input\_data folder is used as input and the fpds are saved in the edited\_data folder. Then, to calculate the LOF scores, the perform\_lof notebook is used. The fpds from the edited\_data folder are used as input and the output is saved in the edited\_data folder as well. Finally, the stream\_merge notebook is used to merge the individual streams into one big stream per location and the output is saved in the output\_data file.

To enable easier handling of the data for the analysis, all the individual datasets are merged into one big dataframe. This is done in the dataframe\_creation folder -> create\_dataframe notebook. The completed dataframe is saved as a csv file in this folder.

The correlation analysis is performed in the calculated\_correlation folder -> correlation\_analysis notebook. This notebook used the complete dataframe as input and besides the correlation matrix, it outputs the correlation scores as a csv file in this folder.

The prediction is performed in the prediction folder. In this folder there are two prediction notebooks. The predictors notebook contains multiple experiments with different predictors and other parameters. The feature\_reduction\_prediction notebook uses the correlation scores csv file and uses it to select only features with a certain correlation threshold.