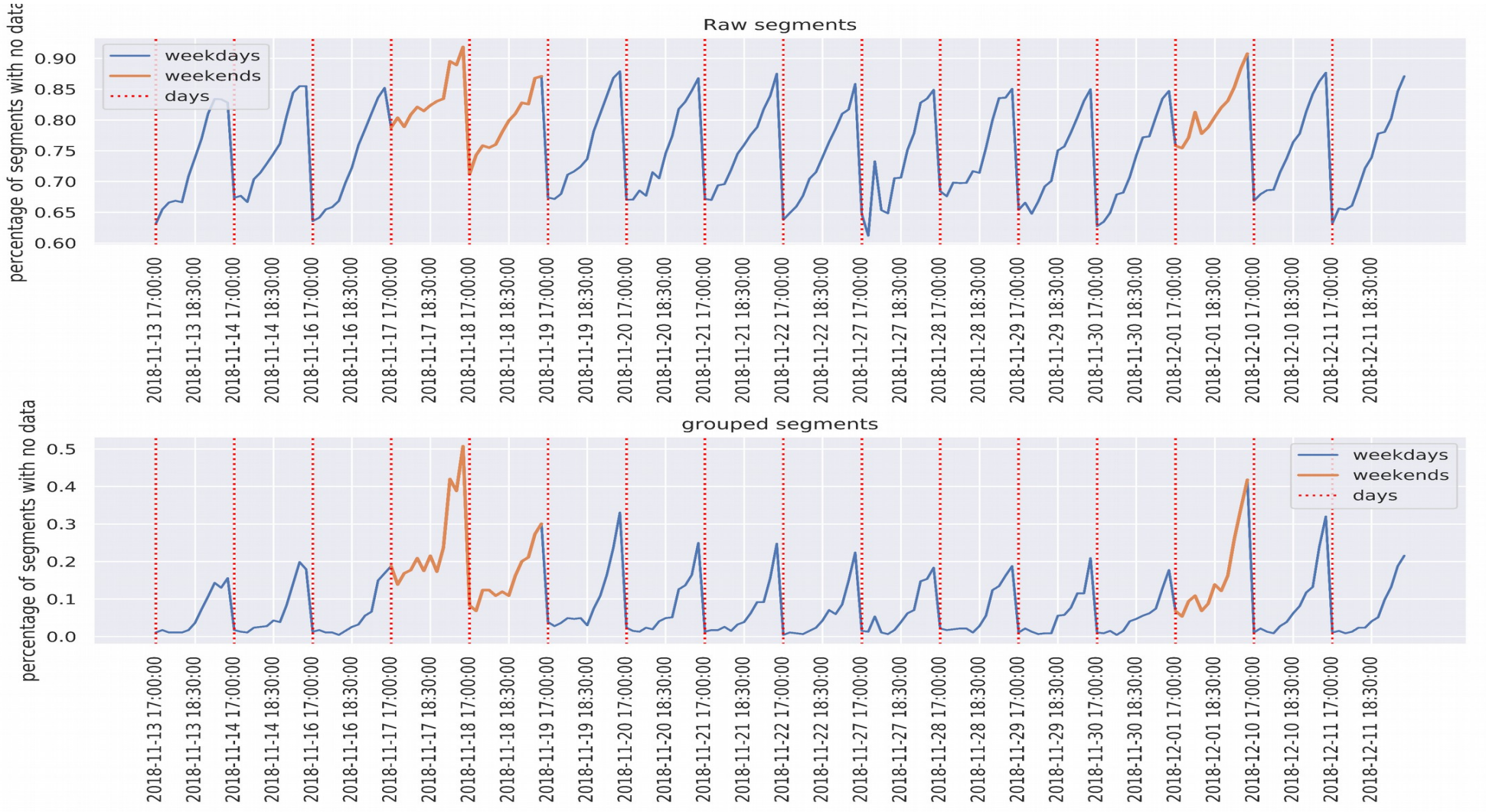
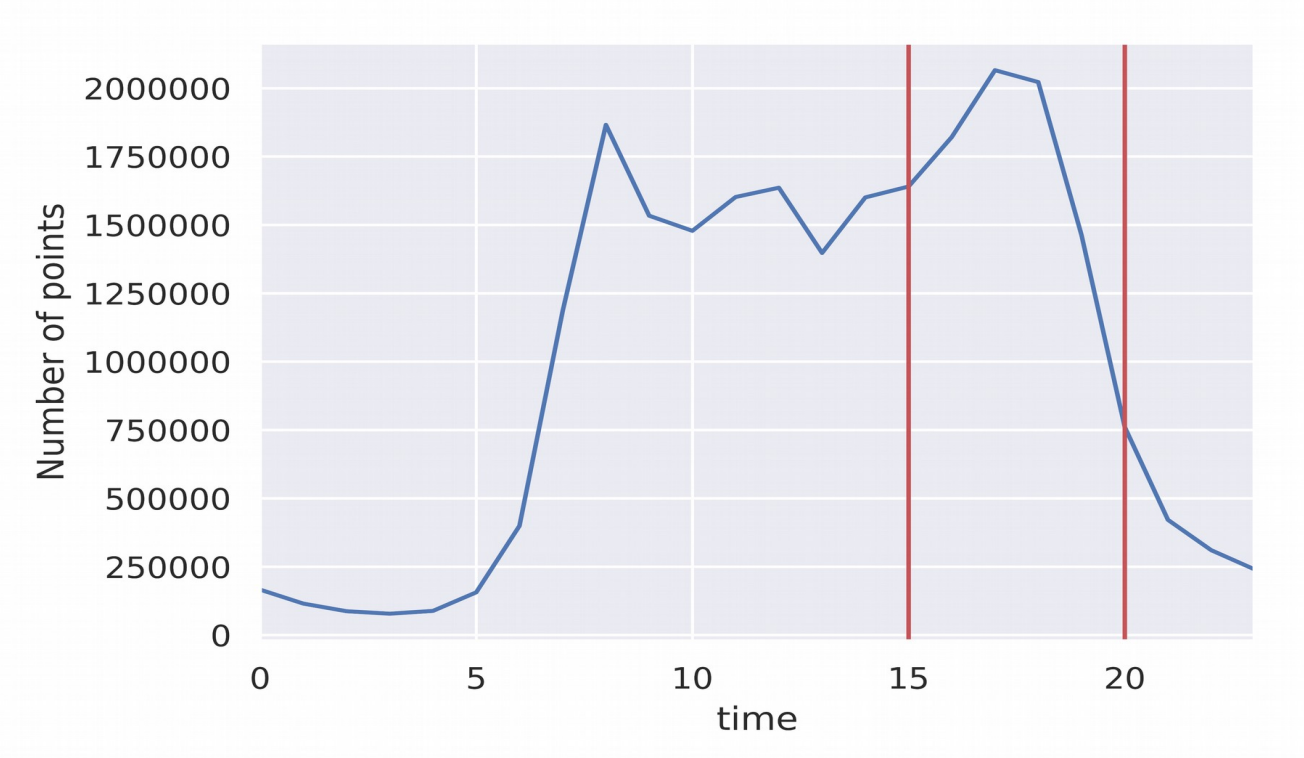
Percentage of segments with non valid data per timestamp in the window 17h-20h



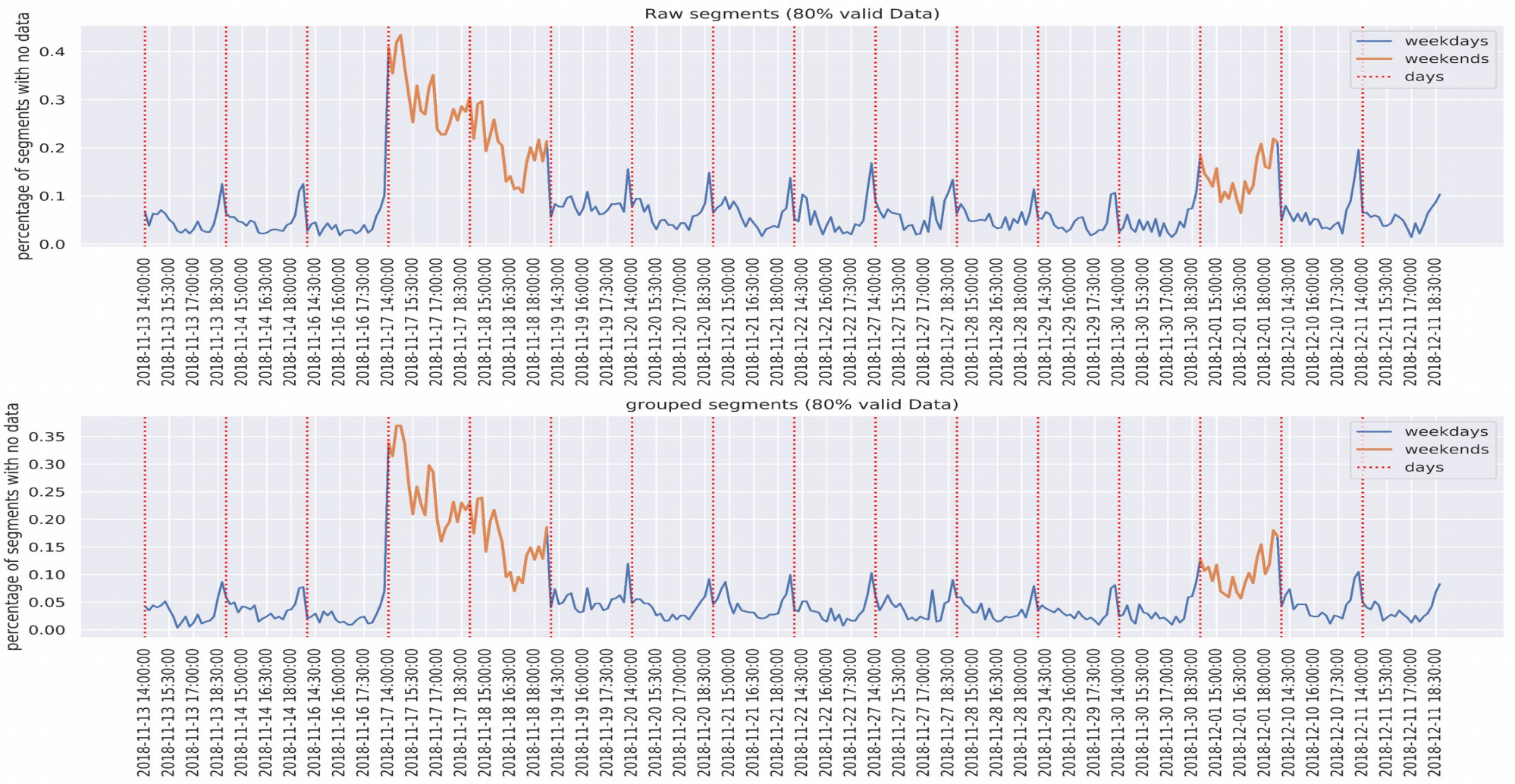
Number of points per hour in the data set



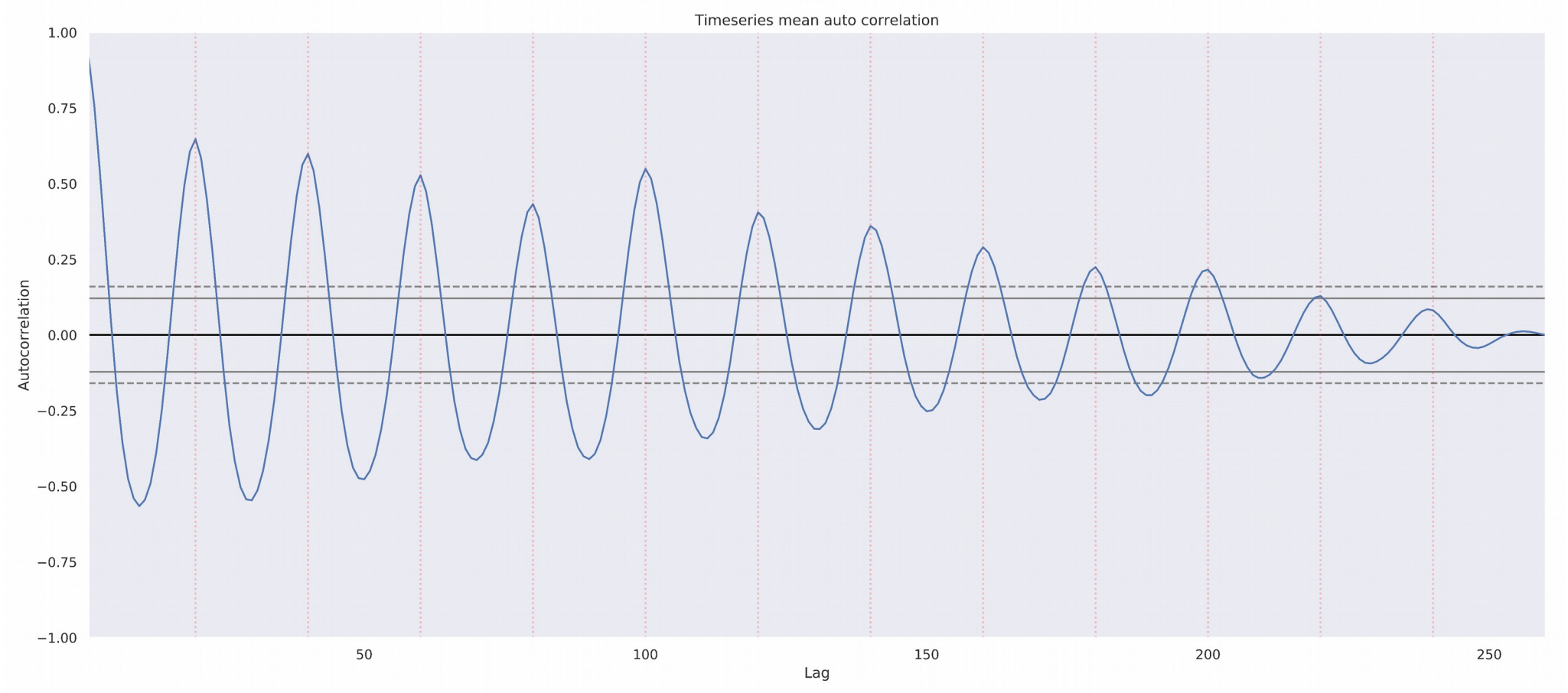
- Note that the time is shifted by one hour since the time in the database is in UTC while local time is UTC+1.
- This shows that the peak hour is 17h-18h thus a window from 15h– 20h is probably a better choice than 17h-20h.

Percentage of segments with non valid data per timestamp in the window 15h-20h (All segments)

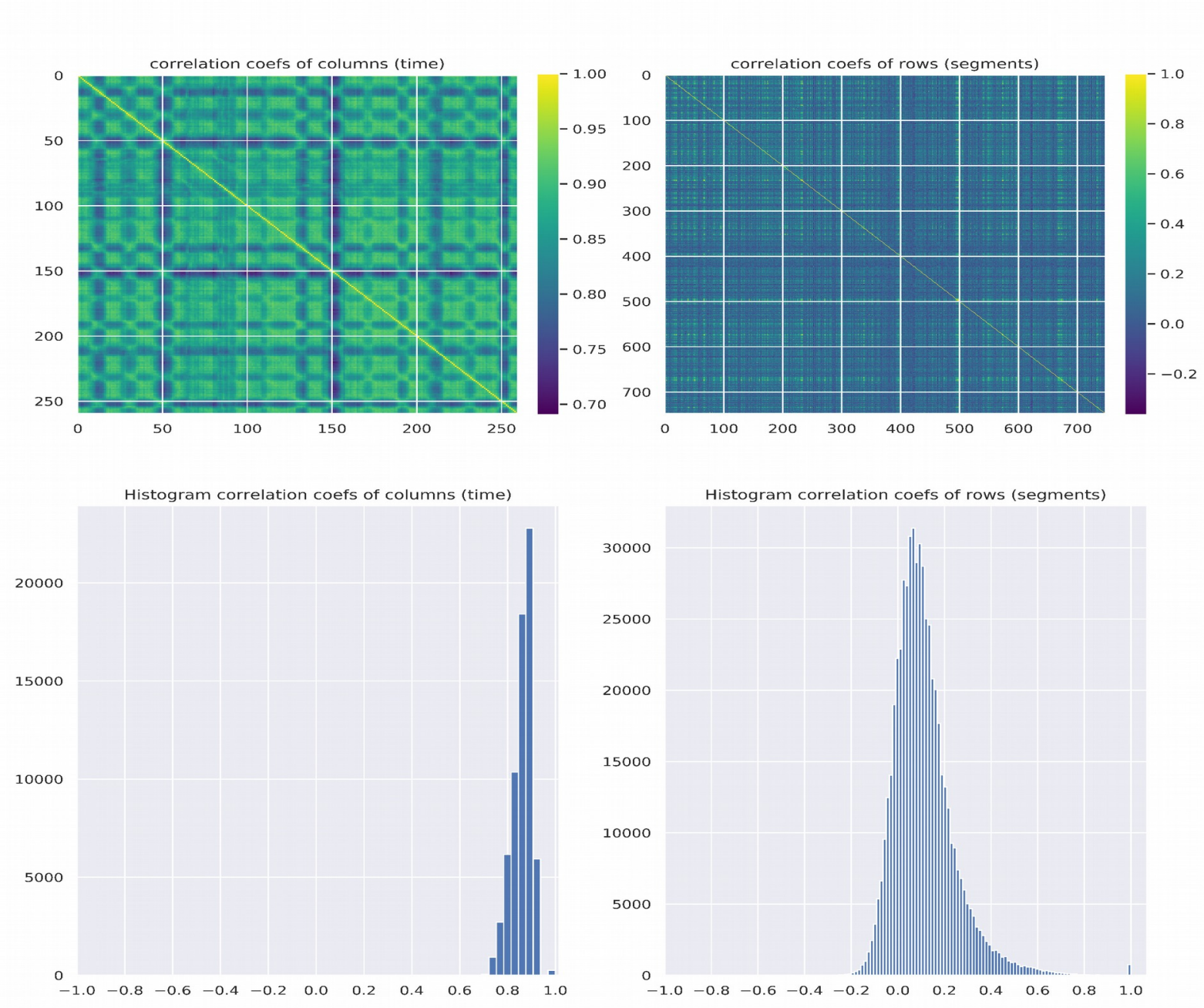
Percentage of segments with non valid data per timestamp in the window 15h-20h (80% valid segments)



Auto correlation of the mean timeseries



Correlation matrices/ histograms of timestamps and segments



Input matrix

- Since we're going to use keras as framework for our deep learning models, the input should be 2D or 3D as stated in the docs :
  "Input shapes 3D tensor with shape (batch_size, timesteps, input_dim)."

- We do this with a single function
  - getXY(data, inputLag, outputLag, sequenceLength)
    - data is the speed matrix
    - inputLag is the number of historical timestamps to lookup
    - outputLag is the number of timestamps we're going to predict
    - sequenceLength is the number of timestamps in our sequence (ex : for a window of 15h-20h with 15mins blocks the sequence length is 20)
  - The output of this function is two matrices X, Y:
    - X is matrix with shape (nSequences, inputLag, nSegments)
    - Y is matrix with shape (nSequences, outputLag, nSegments)