

Files

- experiments/experiment_PCapy
- notebooks/analyze_pcap.py

Motivation

By applying PCA to our data, we want to reduce the feature space to at most half its original size in order to get better runtimes of further operations. If the reduction is too abrupt, the performance of the model starts decreasing.

We needed an experiment to choose the smallest number of components that would not visibly impact the performance.

Design


The experiment had about 65 repetitions, each one with a new data split and including all scopes.

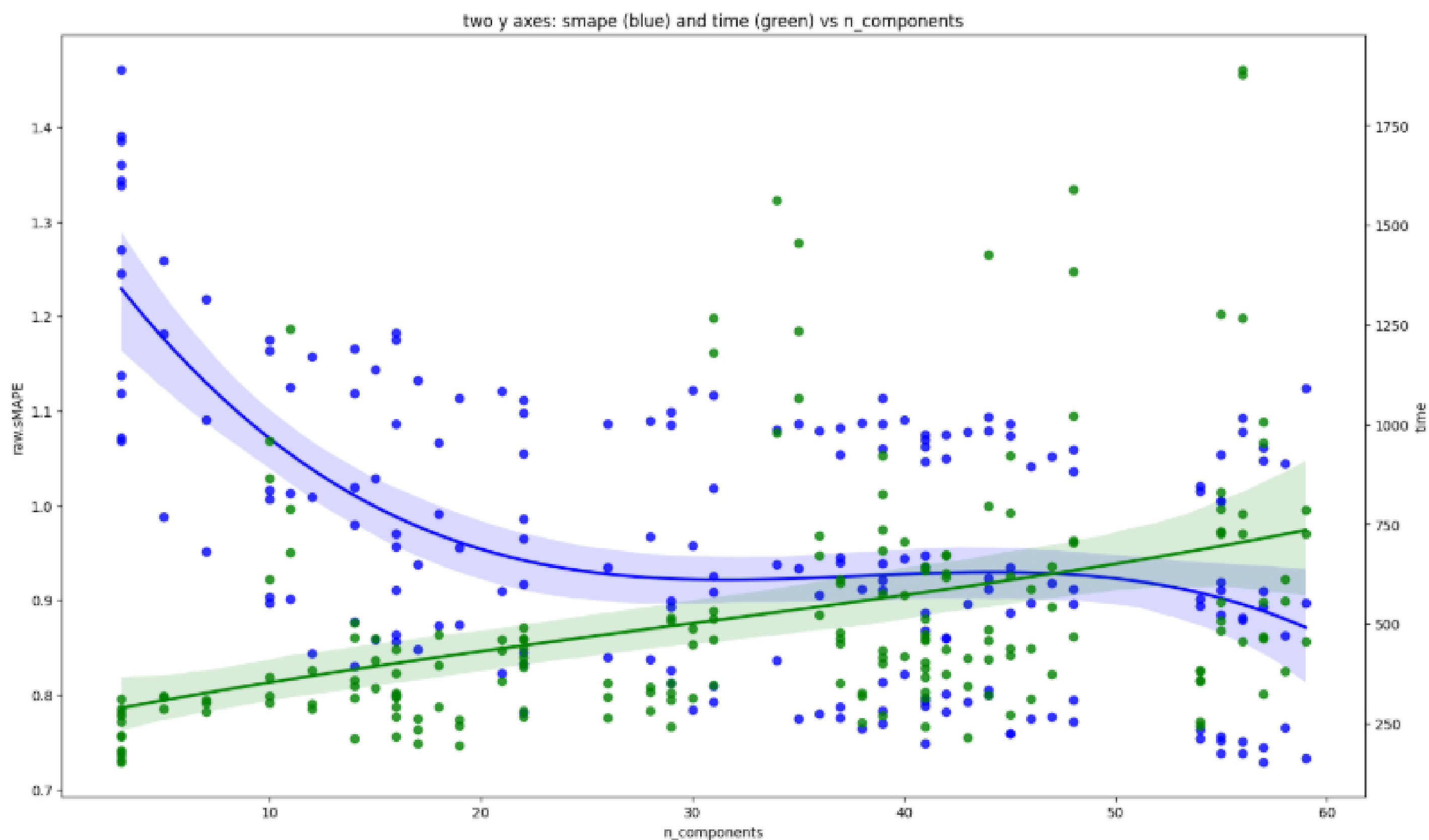
The number of components was randomly chosen between 0 and half the size of the feature space in each repetition, resulting in a uniform distribution of results.

Results and Insight

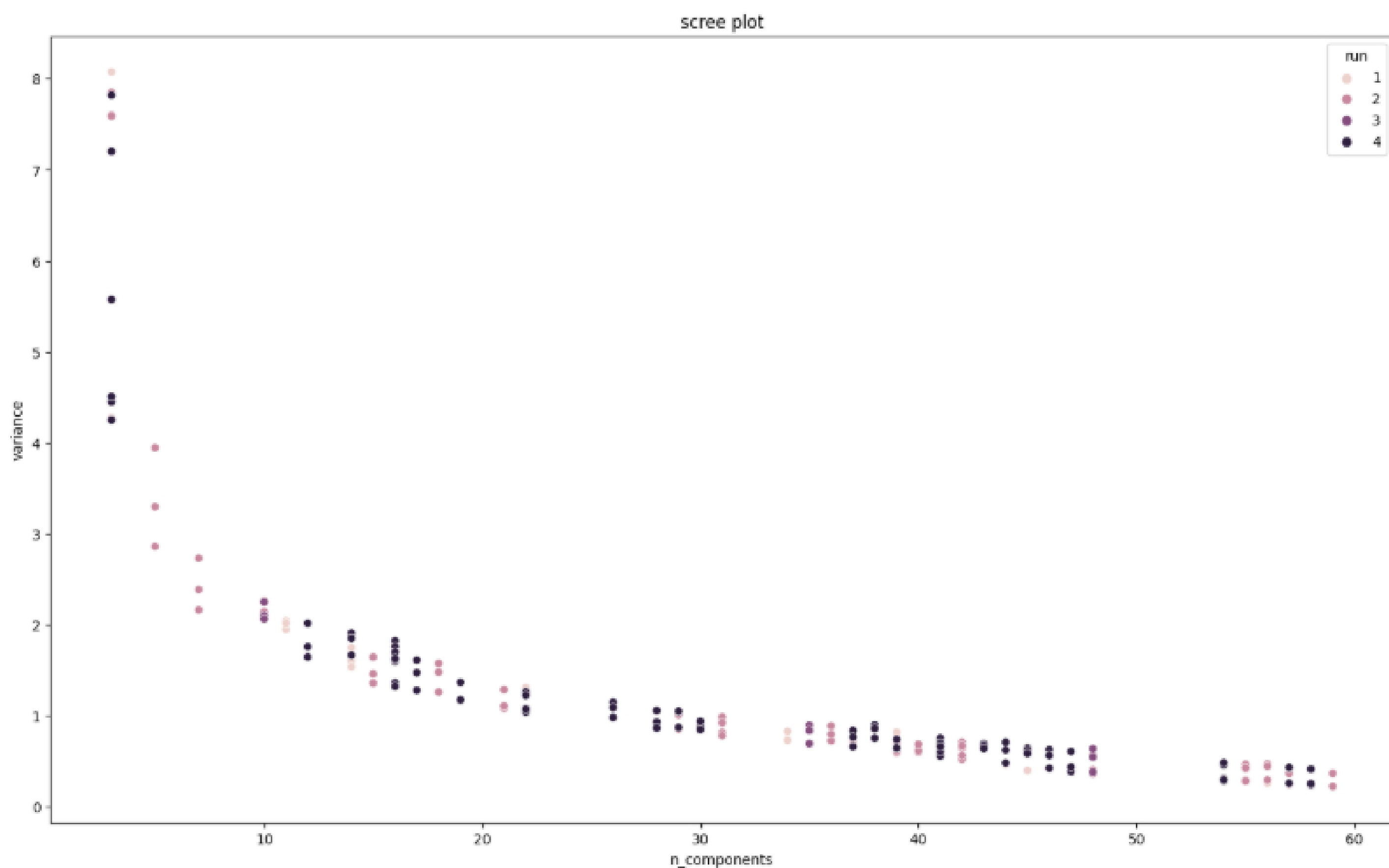
As expected, when the number of components decreases, the runtime decreases and the sMAPE starts to increase due to its non-linear form.

There is an elbow area visible between 20 and 30 components, where the sMAPE value increases towards the lower end of the interval, so this is the sign we need to stop and choose the corresponding number of components.

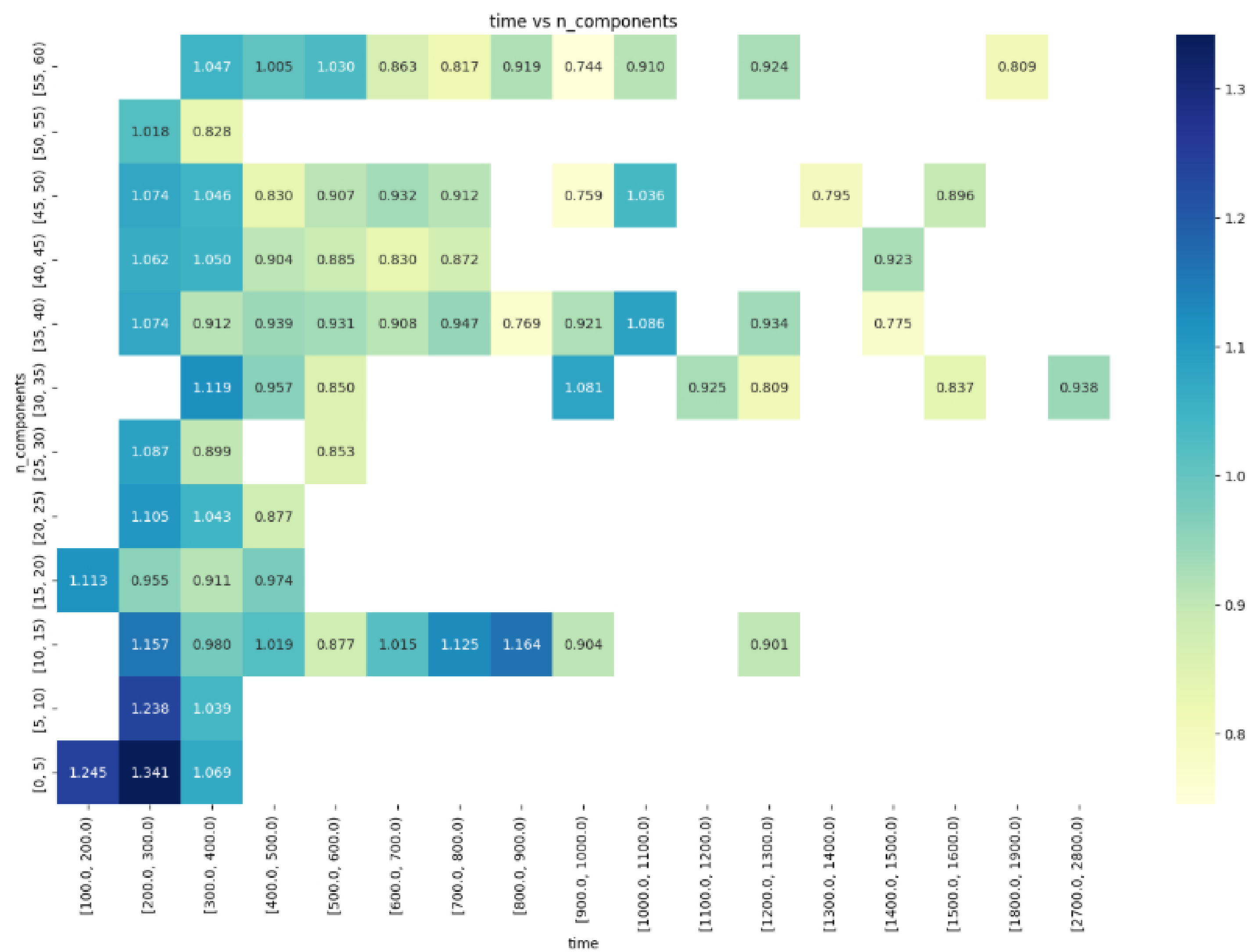
-  The sharp drop at the end of the sMAPE curve is likely an artifact from curve fitting with a polynomial of degree 3.



The plot below comes as an explanation for why the smape value gets worse for less components. It shows how the variance increases abruptly when the number of components is too low and therefore we lose precision and gain more prediction errors.



This heatmap shows in more detail the sMAPE values for binned results. It follows the shape of the first graph, as the less components we choose, the lower the runtime is. The most balanced area is between 25 and 30 components, which is why we will choose 30 components in the end.



Decision

We will use 30 components, as this is the best tradeoff point between time and performance.