Clustering of Covid-19 Time Series

Lisa Pucknat, Alexander Zorn

Lab Development and Application of Data Mining and Learning Systems:
Data Science and Big Data

9st February of 2021

UNIVERSITÄT BONN

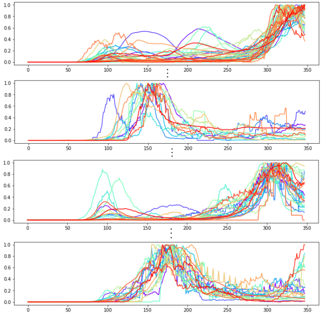
Problem Definition

Task: Analysis of Covid-19 pandemic

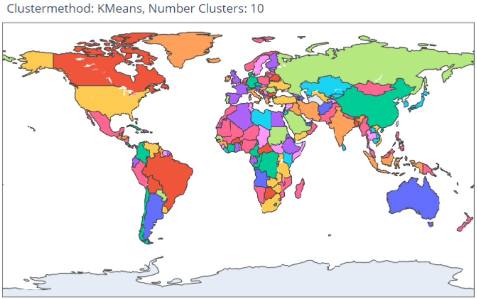
- Usage of dataset with daily Covid-19 cases
- Clustering algorithms for time-series to find clusters by countries and timespans
- Prediction of future cases using cluster analysis of the results

Overview and New Achievements

- Previously: Problem with unbalanced cluster distribution
- Solution: Only look at standardized trends of time-series
  i.e. daily case values in [0,1]



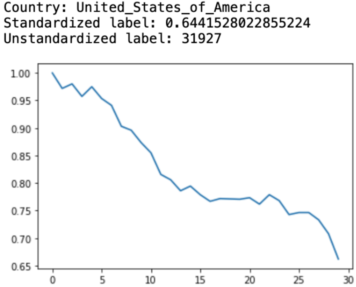
Exemplary selection of clusters



Corresponding geo. representation

Snippets

- 212 countries
- 80/20 train/test split -> selecting 42 random countries.
- create up to 50 snippets for each country of
  - each snippet has length 30
  - 1 day for forecast
  - 1D convolution -> 7 day average
- convert the convoluted forecast to an absolut nr of cases.



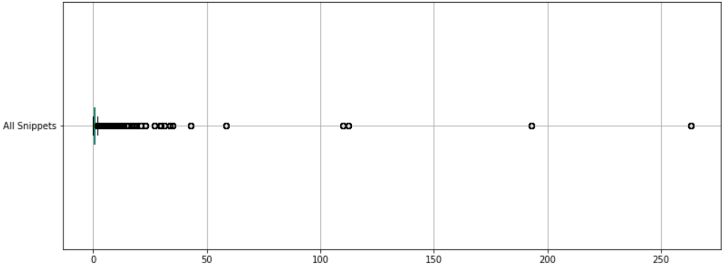
```
Country: United_States_of_America
Standardized label: 0.6441528022855224
Unstandardized label: 31927
```

Snippet example

Benchmark Forecasts

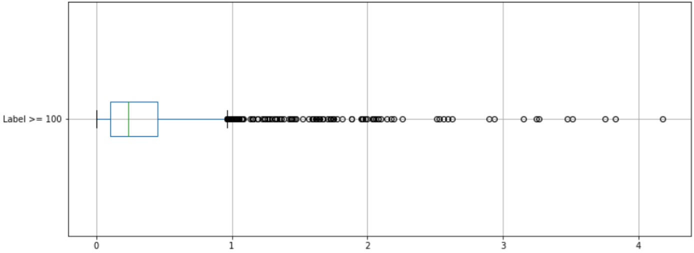
Forecast results with Naive Forecast. (Taking the last day in a Snippet as forecast). Optimized with 7 day average.

Mean: 79.4% Variance: 2070%



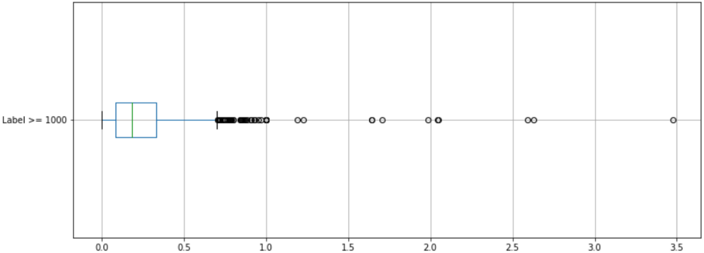
Benchmark Forecasts

Mean: 33.7% Variance: 13%

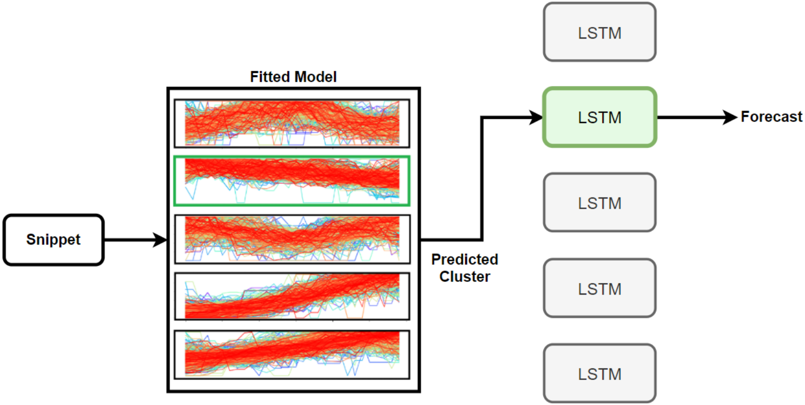


Benchmark Forecasts

Mean: 24.5% Variance: 6%

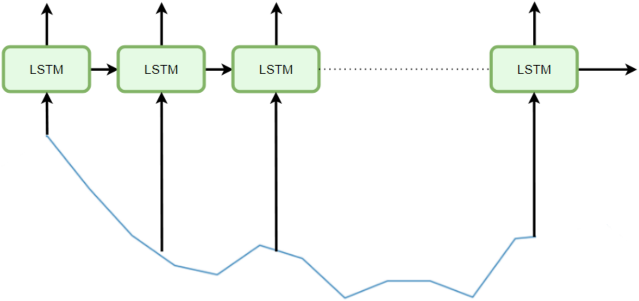


Clustering for Forecast



Pipeline for combination of clustering and forecasting
LSTM is trained with data from predicted cluster

Long Short-Term Memory

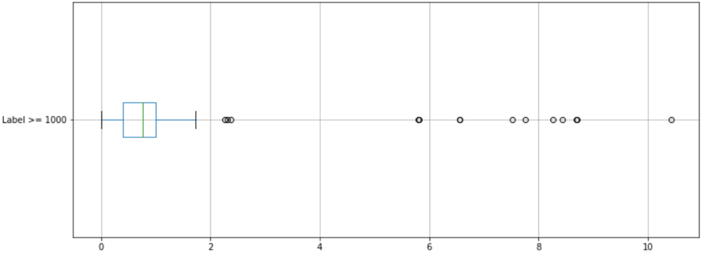


LSTM unrolled over time

- RNN only feeds output of previous time step into current computation
- LSTM-cell has additional long-term state monitored by a forget-gate

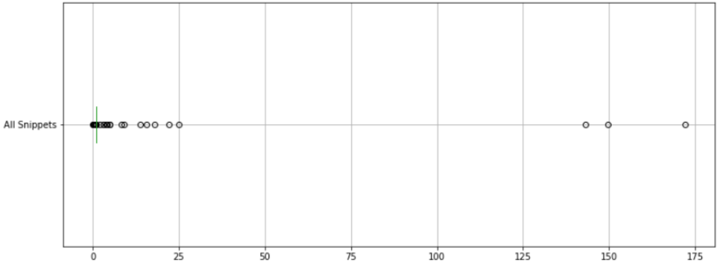
Results LSTM without Clustering

Mean: 109% Variance: 278%



Results LSTM with Clustering

Mean: 137.6% Variance: 4747%



Next Steps

To conclude the Project we plan to cover the following steps:

- Kfold testing
- introduce one or two additional forecasting methods
- Test many Hyperparameters to improve forecast
- Final compare of Forecasting Methods with and without Clustering.
- Documenting the Github project
- Finalize report.