Stock Market Analysis Using Machine Learning

Final Project in Applied Machine Learning

D. A. Bobruk, E. Hellebek, K. Kevin & G. E. Madsen
Niels Bohr Institute

Thanks to Adriano Agnello for providing
us with the data and continuous feedback.

Introduction

- LSTM - prediction of stock prices
  - Portfolio maximization
- Regime determination using Hidden Markov Model
- Clustering different stocks

The Data

All data is retrieved from the Yahoo Finance website:
https://finance.yahoo.com/

Data contains 149 tickers (individual stocks) covering a large portion of the American market as well as a few from Europe and Asia.

Data contains opening price, lowest price, highest price, closing price, adjusted closing price and volume for each day.

Data spans from 2000 until June 2021. We train / validate from 2000-2017 (and test on 2018-2019) and additionally from 2000-2018 (and test on 2019-2021).

Data pre-processing

Significant portion of NaNs found in the data. We forward fill and add a small Gaussian kick determined from yesterday's volatility (High-Low).

We train, validate and test on the difference in log(price). The gap between the days over which we compute the distance is referred to as **dif**. This reduces complexity in the data given the exponential growth of stocks.

The data is scaled with MinMaxScaler to be within the range of 0 and 1.

Note: Model will output the relative change and not the actual stock price.

Why LSTM?

LSTM (Long Short term memory) is a supervised deep learning method backfeeding RNN.

Advantages:

- Suitable for working with time-series
- Well documented and tested
- Applies a non-linear regression technique

Disadvantages:

- Only works properly with large enough data-sets
- Very slow (becomes tedious as we increase input)
- Complex optimization problem

Model architecture

Our top two optimized models (Bayesian Optimization).

Model 1 (optimized on 8 tickers):

- dif = 10
- Horizon = 15
- Lookback = 31
- 2 hidden layers (192, 372)
- dropout = 0.2 (for small scale)
- Learning rate = 8.36e-4
- Activation: relu and sigmoid
- Optimizer: Adam
- Loss: MSE
- batch size = 30
- Early stopping: monitored validation loss (validation/train ratio = 20%)

Model 2 (optimized on 90 tickers):

- dif = 10
- Horizon = 15
- Lookback = 47
- 3 hidden layers (661, 142, 502)
- dropout: no dropout
- Learning rate = 1.8e-3
- Activation: relu and sigmoid
- Optimizer: Adam
- Loss: MSE
- batch size = 30
- Early stopping: monitored validation loss (validation/train ratio = 20%)

Prediction is done for Horizon: 3, 15, 30, and 90 days with lookback 15, 31, 90 and 200 days respectively. Model 1 has the strongest prediction power when scaling up to 149 tickers. Note: optimizing on Horizon greater than 15 produces too complex models.

LSTM predictions - Great prediction

Horizon = 15 days

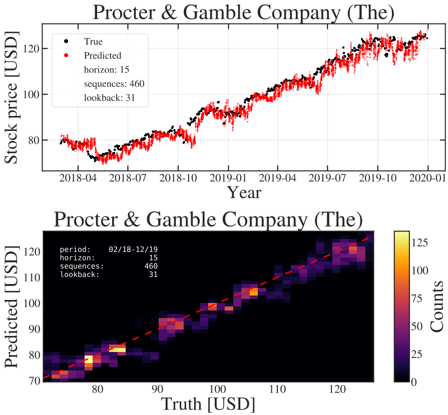


Figure: *Top panel*: Predicted and True stock value from 2018 - 2020. *Bottom panel*: Diversion plot: Red line indicates perfect prediction.

Horizon = 90 days

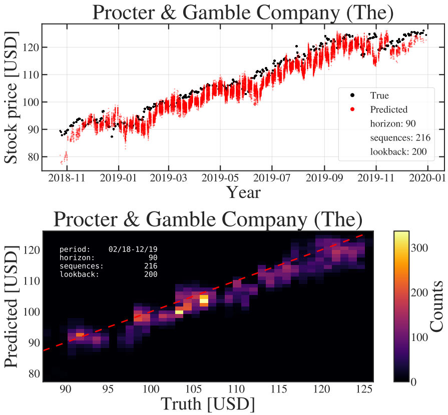


Figure: *Top panel*: Predicted and True stock value from 2018 - 2020. *Bottom panel*: Diversion plot: Red line indicates perfect prediction.

LSTM predictions - poor prediction

Horizon = 15 days

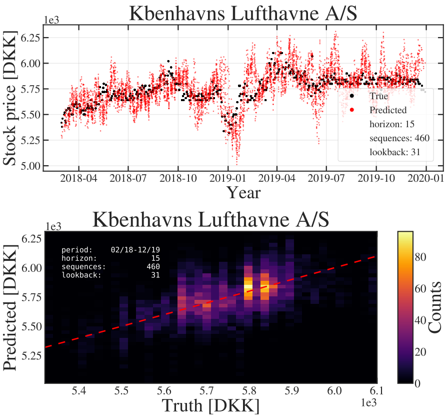


Figure: *Top panel*: Predicted and True stock value from 2018 - 2020. *Bottom panel*: Diversion plot: Red line indicates perfect prediction.

Horizon = 90 days

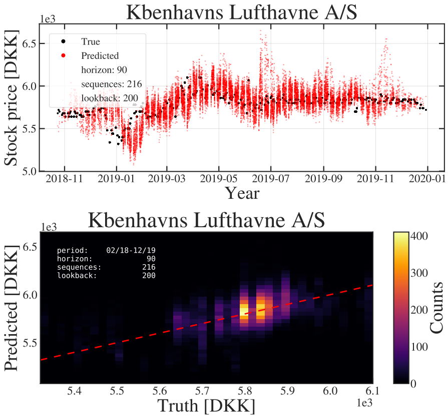


Figure: *Top panel*: Predicted and True stock value from 2018 - 2020. *Bottom panel*: Diversion plot: Red line indicates perfect prediction.

LSTM predictions - COVID 19

Horizon $=$ 15 days

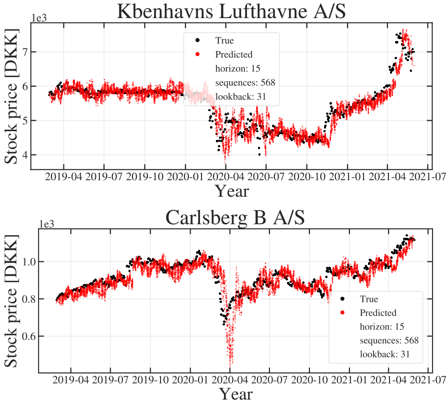


Figure: *Top panel:* Predicted and True stock value from 2018 - 2021 for Copenhagen Airport. *Bottom panel:* Predicted and True stock value from 2018 - 2021 for Carlsberg Brewery.

Horizon $=$ 90 days

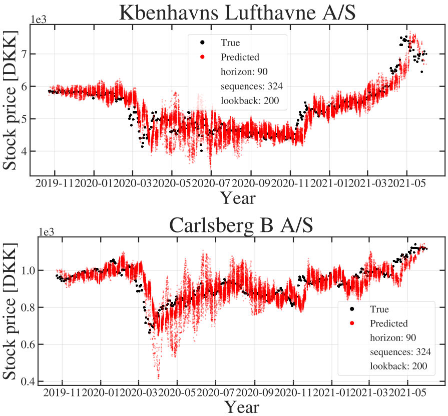


Figure: *Top panel:* Predicted and True stock value from 2018 - 2021 for Copenhagen Airport. *Bottom panel:* Predicted and True stock value from 2018 - 2021 for Carlsberg Brewery.

LSTM summary for horizon = 3 days

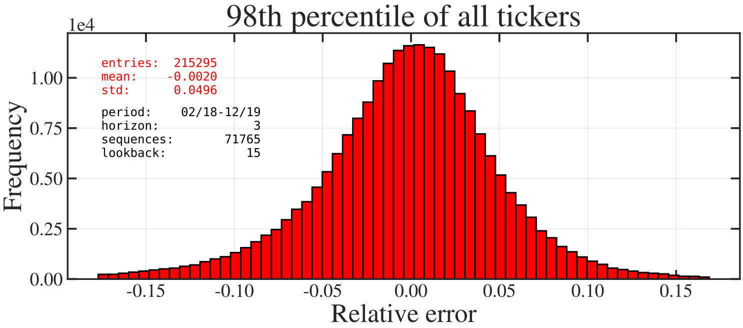


Figure: The 98th percentile of the relative errors on the predicted data in the period before the corona pandemic, using a horizon of 15 days and data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

LSTM summary for horizon = 90 days

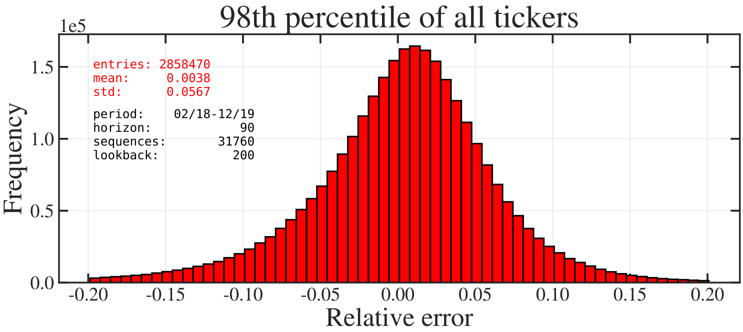


Figure: The 98th percentile of the relative errors on the predicted data in the period before the corona pandemic, using a horizon of 90 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Portfolio maximization: Equity

S&P 500 (SPY) reflects 500 large companies in USA.

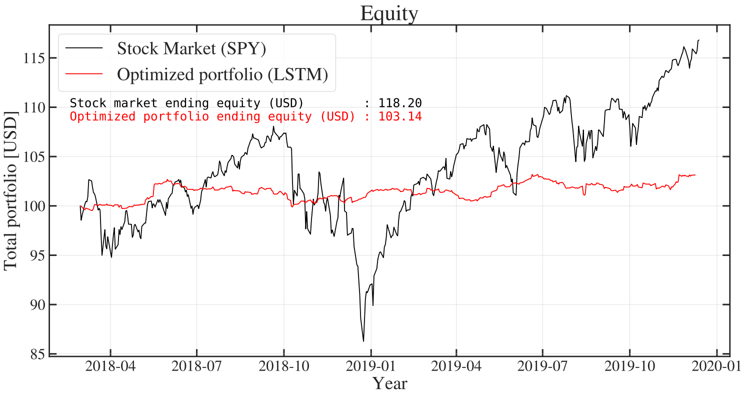


Figure: Equity from portfolio optimization (using LSTM results for Horizon = 15 days before the pandemic) and for stock market given an initial investment of 100 USD.

Portfolio maximization: Summary

Horizon	Optimized portfolio			Stock Market
	Annualized Sharpe Ratio	Annualized Volatility	Ending Equity	Ending Equity
3 days	5.3e-4	6.2e-4	103.7 USD	119.3 USD
3 days (w/ COVID-19)	2.0e-3	7.8e-4	79.3 USD	153.8 USD
15 days	2.8e-3	9.9e-5	103.1 USD	118.2 USD
15 days (w/ COVID-19)	2.3e-4	1.3e-4	99.6 USD	150.5 USD
30 days	2.1e-3	3.6e-5	99.3 USD	117.5 USD
30 days (w/ COVID-19)	2.5e-3	5.7e-5	98.2 USD	147.1 USD
90 days	1.5e-4	8.6e-6	100.0 USD	121.0 USD
90 days (w/ COVID-19)	4.8e-4	1.6e-5	100.05 USD	140.08 USD

- Low volatility means lower potential return (thus higher return comes with higher risk).
- Annualized Sharpe ratio greater than 1 is considered acceptable to good by investors. Greater than 2 is very good. Portfolios failed to come close to these figures.
- Current Sharpe Ratio for S&P 500 is 2.24 and its average volatility is 0.18.

LSTM - Discussion

- Prediction power is dependent on many parameters such as horizon and lookback.
- Optimization is a rabbit hole. More can be done such as implementing cross-validation, social media data, news etc.
- Large amount of tickers compromise the prediction power of individual tickers.
- Stock market is incredibly hard to beat given the large overall growth which is not reflected in all tickers that goes in the portfolio. More knowledge is necessary to reach a Sharpe ratio closer to 1.

Regime identification - Hidden Markov Model

- Regime Detection
- Parameters: Lookback, interval for price difference, number of states
- Unsupervised

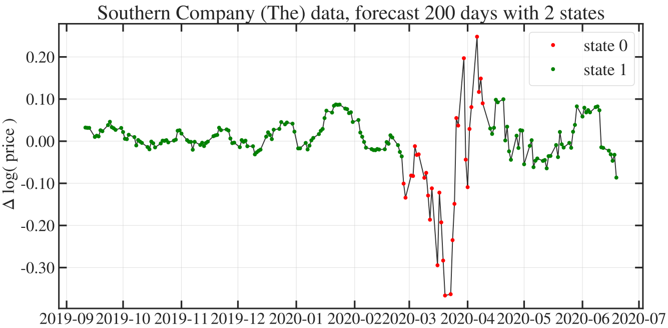


Figure: HMM on SO with 2 states

Regime identification - Results

- Transition Matrix and probabilities

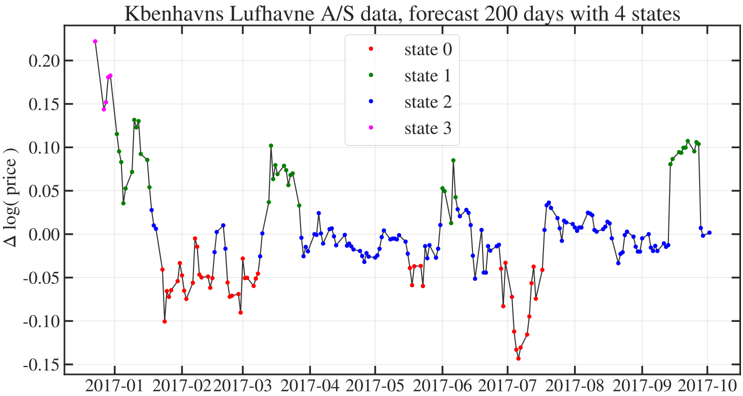


Figure: HMM on KBHL Data

Regime identification - Discussion

- Training on multiple tickers, biased towards specific stock
- Difficult to evaluate

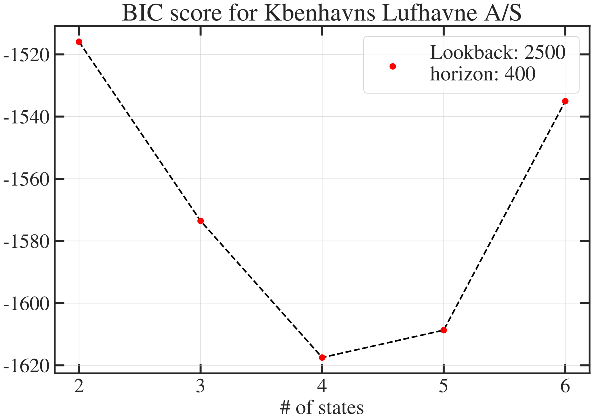


Figure: BIC score for KBHL Data (Copenhagen Airport)

Clustering - Data Pre-processing

- 98000 tickers reduced to 5174.
- UMAP - reduction

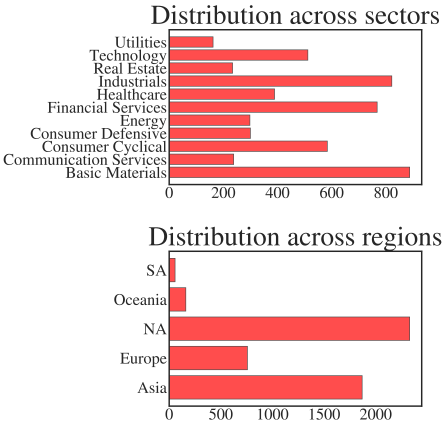
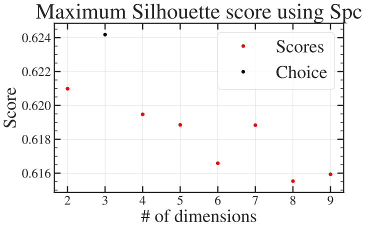


Figure: The distribution of the tickers across sectors (top panel) and regions (lower panel).

Clustering - finding the number of clusters and dimensional reduction

- Spectral clustering
- Silhouette score
- Consider on the data



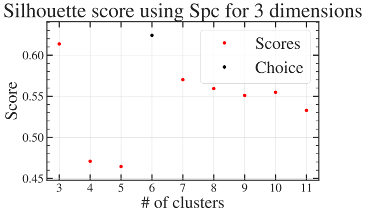


Figure: The maximum score for each dimension and the score for each number of clusters for 3 dimensions.

Clustering - Results after clustering

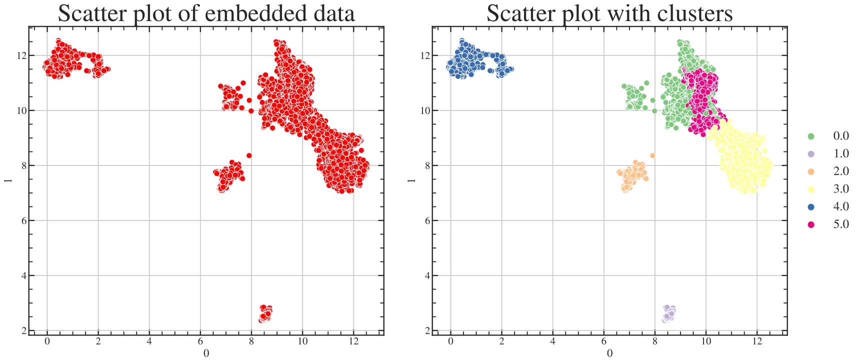


Figure: A scatter plot of the first and second dimensions of the embedded data is shown in the two panels. The raw data is shown to the left. How the data was clustered is shown to the right.

Clustering - Results after clustering

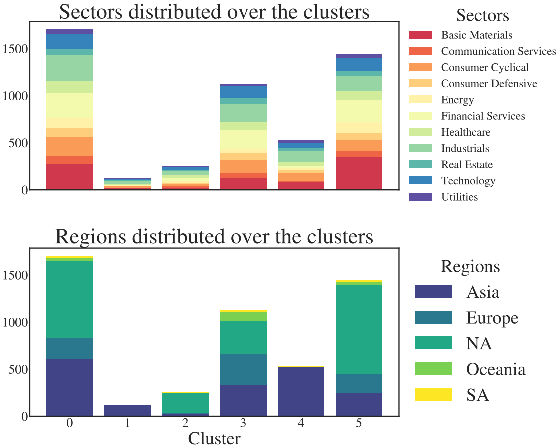


Figure: Showing how the sectors (top panel) and regions (bottom panel) were distributed across the clusters.

Clustering - Discussion

- We see a tendency for some correlations
- These correlations does not seem to neatly fit with any of the "predetermined" factors we considered.
- Further Work:
  - Consider more factors to what fits the correlation
  - Consider a single country or economic zone, and find clusters inside these.
  - Let the clustering determine the regions.

Closing remarks

- The stock market is a large and complex problem to deal with using Machine Learning.
- Even though our predictions worked quite well on some tickers, it fell short on others, especially when predicting far into the future, i.e our model is very sensitive to larger changes such as the COVID-19 pandemic.
- If it were easy all would do it!

References

- Yahoo Finance, https://finance.yahoo.com/
- David Puelz, P. Richard Hahn and Carlos M. Carvalho (2016), **SPARSE MEAN-VARIANCE PORTFOLIOS: A PENALIZED UTILITY APPROACH**, url: https://arxiv.org/pdf/1512.02310.pdf
- Investopedia, **Sharpe Ratio**, url: https://www.investopedia.com/terms/s/sharperatio.asp

Appendix A: Validation Loss and Accuracy

For each model we have run, we monitor the validation loss in order to prevent over-fitting. An example is shown here for Horizon = 3 and Lookback = 15 for 149 tickers with Model architecture 1 training on data from 2000-2018.

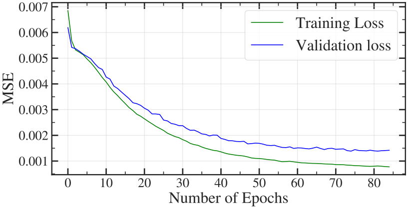


Figure: History of loss function (MSE) doing training for training and validation as indicated in the legend.

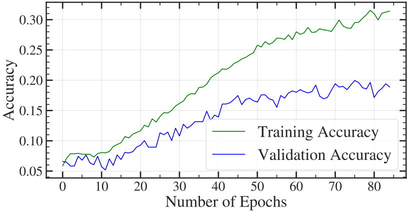


Figure: History of accuracy doing training for training and validation as indicated in the legend.

Appendix A: Testing across same time period as training

In the presentation, we have shown the resulting predictions from testing on a consecutive time period to the training sample. Here we've shown the prediction within the same time period, where we have split the data into train and test (20% test) and additional 20% for validation. An example is shown below for Horizon = 15, Lookback = 31 and dif=10.

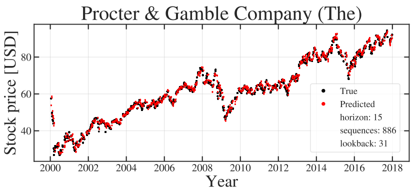


Figure: Prediction and True value of stock within 2000-2018

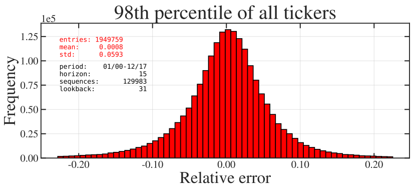


Figure: Relative error distribution for all 149 tickers (test sample)

Altough, the relative error distribution does not change much from the scenario where we tested in the years after the training period, we find the prediction power to be very high across all years as seen in the prediction plot on the left. Relative error distribution is sensitive to some large outliers in the beginning of 00's.

Appendix A: LSTM summary for horizon = 3 days (including the COVID-19 pandemic)

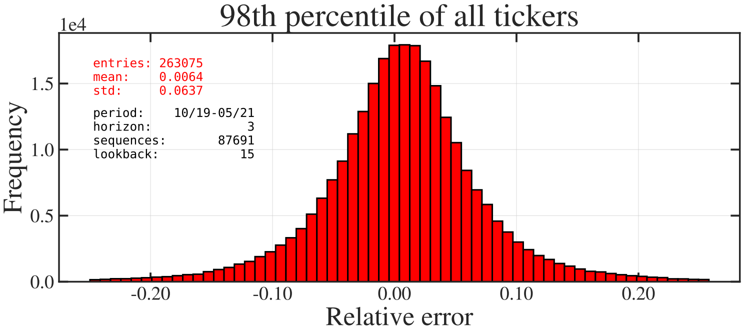


Figure: The 98th percentile of the relative errors on the predicted data in the period including the corona pandemic, using a horizon of 3 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Appendix A: LSTM summary for horizon = 15 days (before the COVID-19 pandemic)

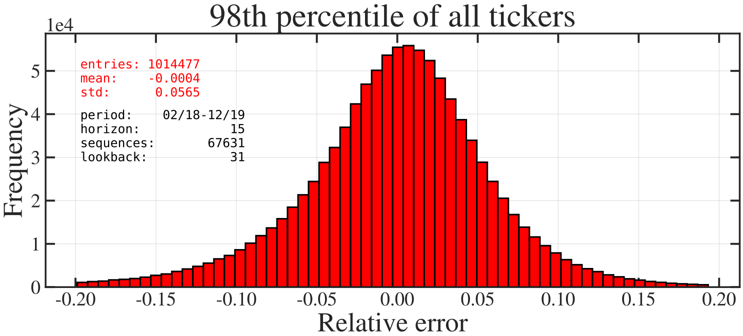


Figure: The 98th percentile of the relative errors on the predicted data in the period before the corona pandemic, using a horizon of 15 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Appendix A: LSTM summary for horizon = 15 days (including the COVID-19 pandemic)

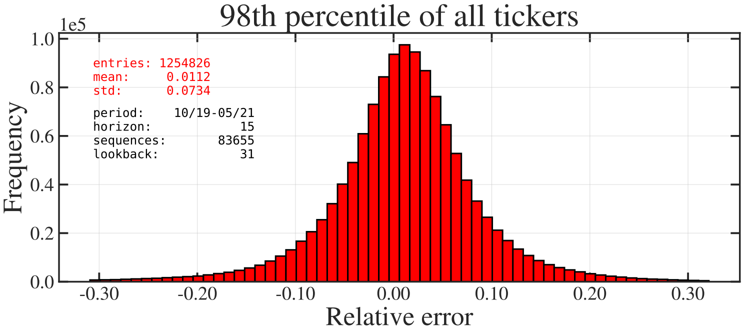


Figure: The 98th percentile of the relative errors on the predicted data including the corona pandemic, using a horizon of 15 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Appendix A: LSTM summary for horizon = 30 days (before the COVID-19 pandemic)

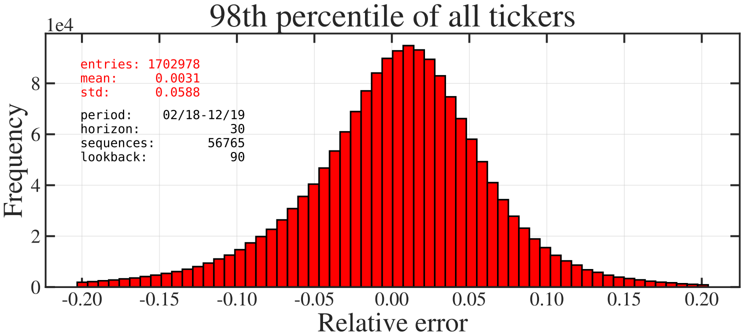


Figure: The 98th percentile of the relative errors on the predicted data in the period before the corona pandemic, using a horizon of 30 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Appendix A: LSTM summary for horizon = 30 days (including the COVID-19 pandemic)

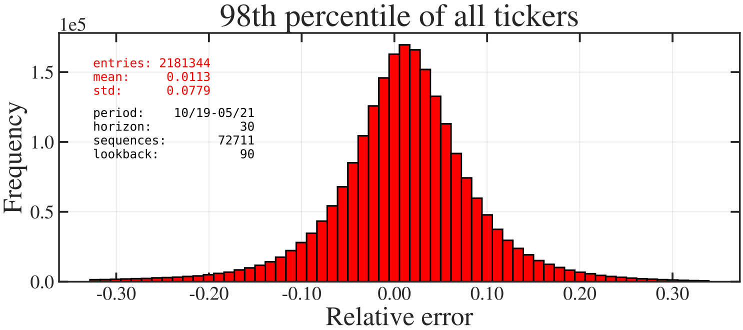


Figure: The 98th percentile of the relative errors on the predicted data including corona pandemic, using a horizon of 30 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Appendix A: LSTM summary for horizon = 90 days (including the COVID-19 pandemic)

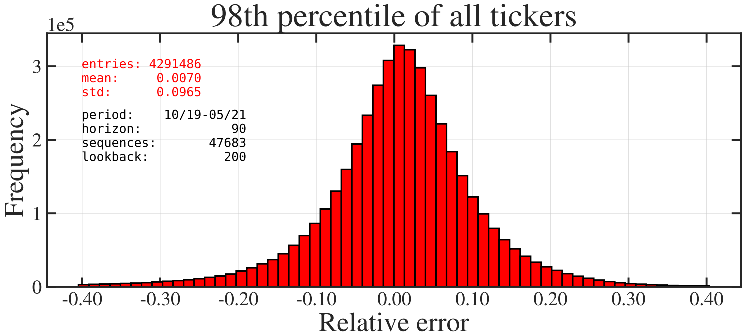


Figure: The 98th percentile of the relative errors on the predicted data including the corona pandemic, using a horizon of 90 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Appendix A: LSTM summary for horizon = 90 days (100th percentile) (before the COVID-19 pandemic)

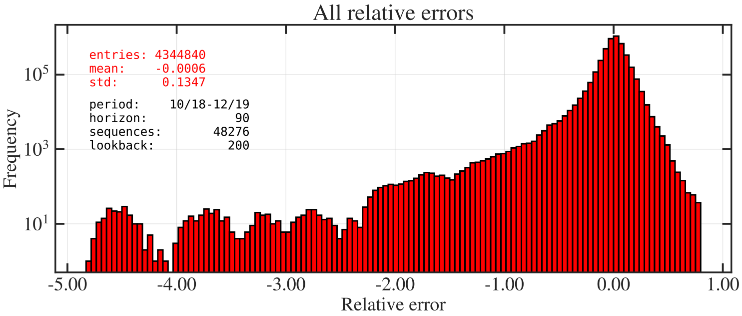


Figure: The 100th percentile of the relative errors on the predicted data in the period before the corona pandemic, using a horizon of 90 days and the data from 149 tickers. The relative errors are calculated as *(true - predicted)/true*.

Appendix A: Run time and ethical considerations

The LSTM's run time scales drastically with larger horizons and lookback (larger data in general). We have summerized a few of the run-times below:

Optimization with Optuna:
Horizon = 10, 90 tickers :
Run-time = 36 hours
Horizon = 20, 21 tickers :
Run-time = 14.4 hours
Horizon = 90, 1 tickers :
Run-time = 19.5 hours

Training the LSTM (model 1):
Horizon = 3, 149 tickers w/ covid: Run-time = 18 minutes
Horizon = 15, 149 tickers : Run-time = 35 minutes
Horizon = 15, 149 tickers w/ covid: Run-time = 1 hour
Horizon = 30, 149 tickers w/ covid: Run-time = 2.5 hours
Horizon = 90, 149 tickers : Run-time = 6.8 hours
Horizon = 90, 149 tickers w/ covid : Run-time = 5.8 hours

It goes without saying, that this is not even close to our total run-time, and the extent to which we have explored our algorithm did not come for free. Ethically speaking, this could have been optimised to run more smoothly. During the span over one weekend, the daily electric consumption doubled at Gustav's place. We unfortunately only monitored this weekend, but the observation stunned us.

Appendix B: Portfolio maximization - Cost function

Cost Function: LASSO (least absolute shrinkage and selection operator)

$$\max_{w \geq 0} \; w^T \mu - \frac{1}{2} w^T \Sigma w + \lambda ||w||_1 \tag{1}$$

$$\mathcal{L}(w, \mu, \Sigma) = \frac{1}{2} w^T \Sigma w - w^T \mu + \lambda ||w||_1 \tag{2}$$

where w is the weights, μ is the mean of the return distribution, Σ is the variance.

Appendix B: Portfolio maximization - Sharpe ratio and annualized volatility

Sharpe ratio: Measure of return of investment compared to its risk.

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{R_{\sigma\, p}} \tag{3}$$

Where R_p is the return of the portfolio, R_f is the risk-free rate, and σ_p is the standard deviation of the portfolio's excess return.

Annualized volatility is computed by dividing the standard deviation of the returns by 252 (there are 252 trading days in a year). Similar for the annualized Sharpe ratio we divide by 252.

Appendix B: Portfolio maximization - Returns

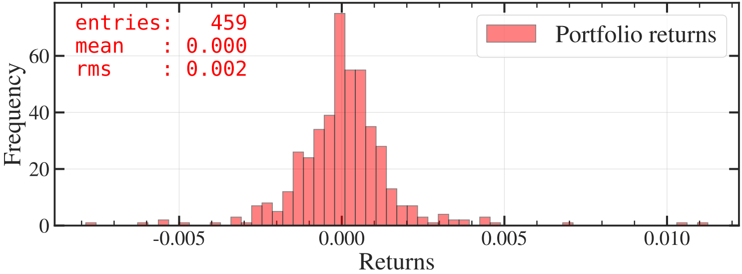


Figure: Distribution of the resulting returns (given in weights) using LSTM results for Horizon = 15 days before the pandemic.

Appendix B: Portfolio optimization results

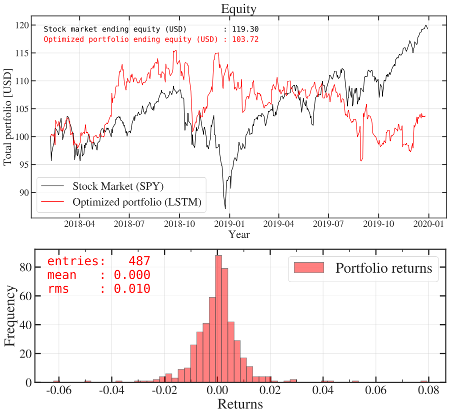


Figure: **Top panel:** Equity from portfolio optimization (using LSTM results for Horizon = 3 days before the pandemic) and for stock market given an initial investment of 100 USD. **Bottom panel**: Distribution of the resulting returns (given in weights)

Appendix B: Portfolio optimization results

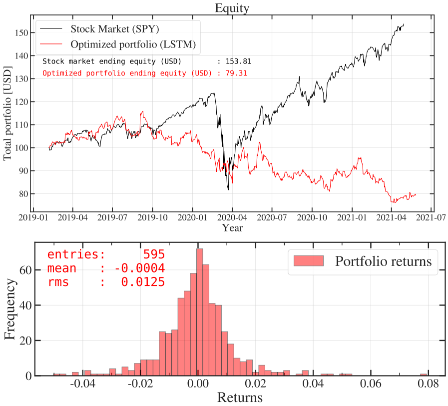


Figure: **Top panel:** Equity from portfolio optimization (using LSTM results for Horizon = 3 days with the pandemic) and for stock market given an initial investment of 100 USD. **Bottom panel**: Distribution of the resulting returns (given in weights)

Appendix B: Portfolio optimization results

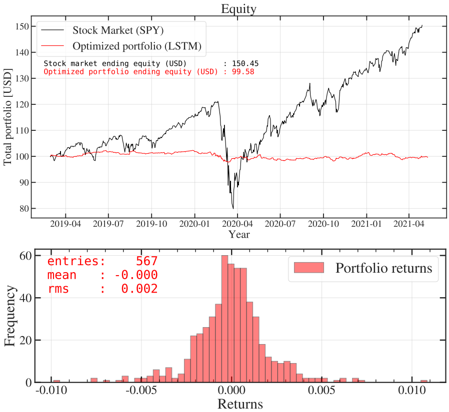


Figure: **Top panel:** Equity from portfolio optimization (using LSTM results for Horizon = 15 days with the pandemic) and for stock market given an initial investment of 100 USD. **Bottom panel**: Distribution of the resulting returns (given in weights)

Appendix B: Portfolio optimization results

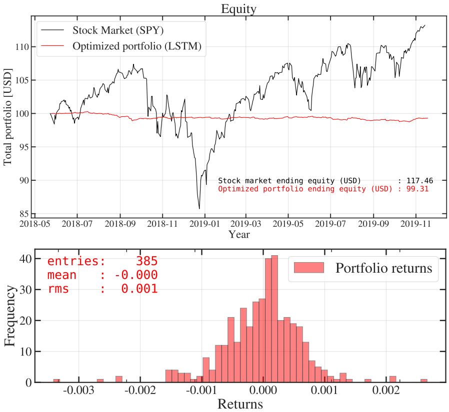


Figure: **Top panel:** Equity from portfolio optimization (using LSTM results for Horizon = 30 days before the pandemic) and for stock market given an initial investment of 100 USD. **Bottom panel**: Distribution of the resulting returns (given in weights)

Appendix B: Portfolio optimization results

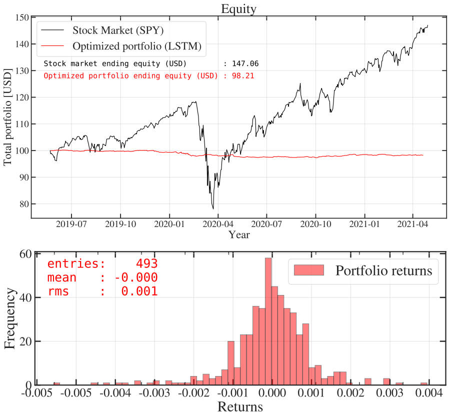


Figure: **Top panel:** Equity from portfolio optimization (using LSTM results for Horizon = 30 days with the pandemic) and for stock market given an initial investment of 100 USD. **Bottom panel**: Distribution of the resulting returns (given in weights)

Appendix B: Portfolio optimization results

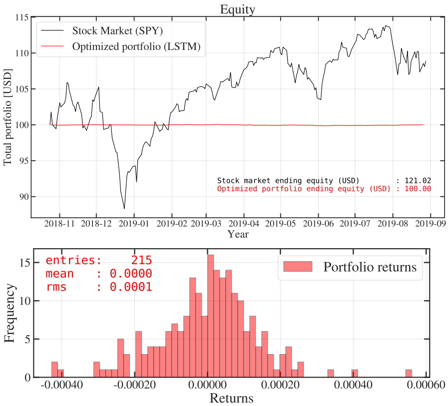


Figure: **Top panel:** Equity from portfolio optimization (using LSTM results for Horizon = 90 days before the pandemic) and for stock market given an initial investment of 100 USD. **Bottom panel**: Distribution of the resulting returns (given in weights)

Appendix B: Portfolio optimization results

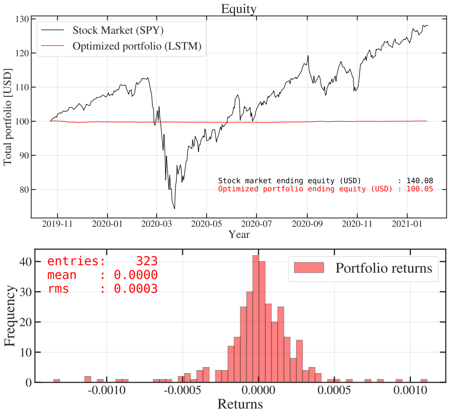


Figure: **Top panel:** Equity from portfolio optimization (using LSTM results for Horizon = 90 days with the pandemic) and for stock market given an initial investment of 100 USD. **Bottom panel**: Distribution of the resulting returns (given in weights)

Appendix C: HMM predicted probabilities

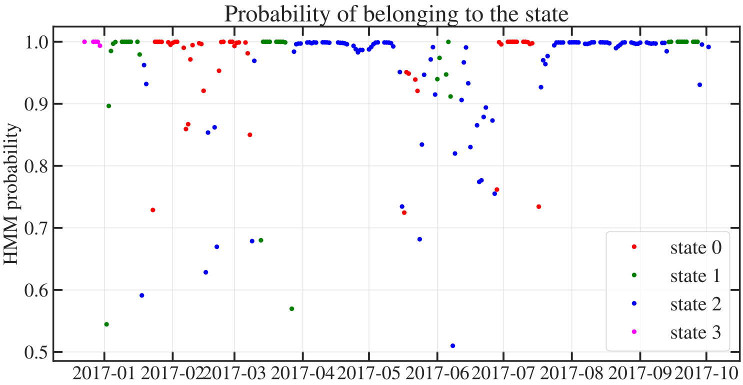


Figure: Probabilities of predicted states of KBHL.CO data for the 200 day horizon.

Appendix C: HMM predicted probabilities

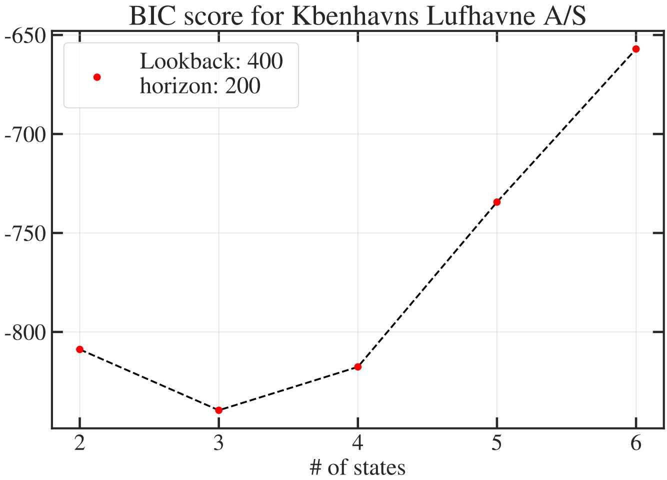


Figure: Different BIC score for lookback = 400 and horizon = 200 days.

Appendix D: Choice of clustering method

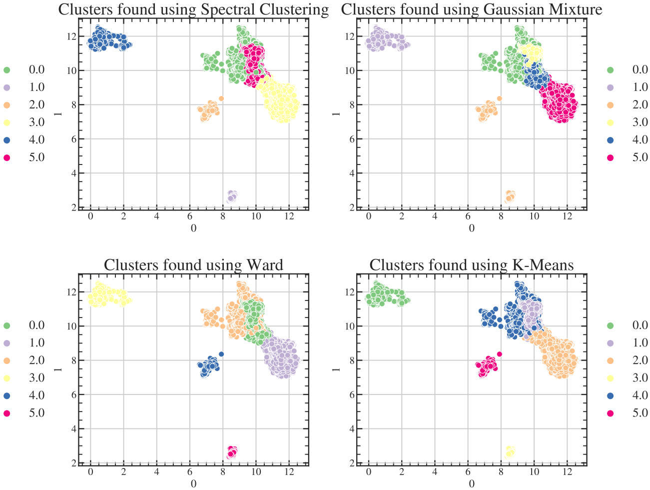
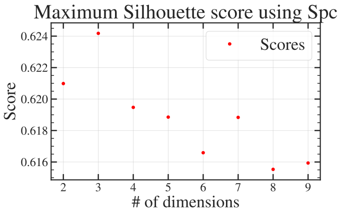
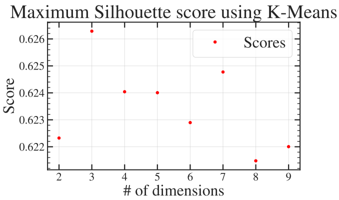


Figure: How the embedded data was clustered using the four different methods used in this project.

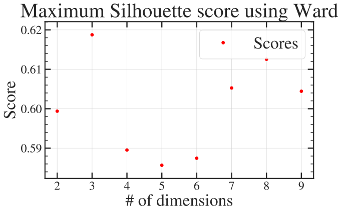
Appendix D: Choice of clustering method



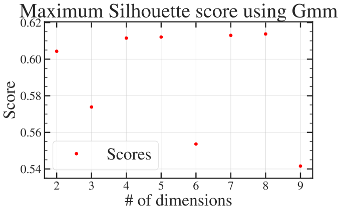
(a) Spectral Clustering



(b) K-Means



(c) Agglomerative clustering



(d) Gaussian Mixture Model

Figure: How the different clustering algorithm clusters the data

Appendix D: Choice of clustering method

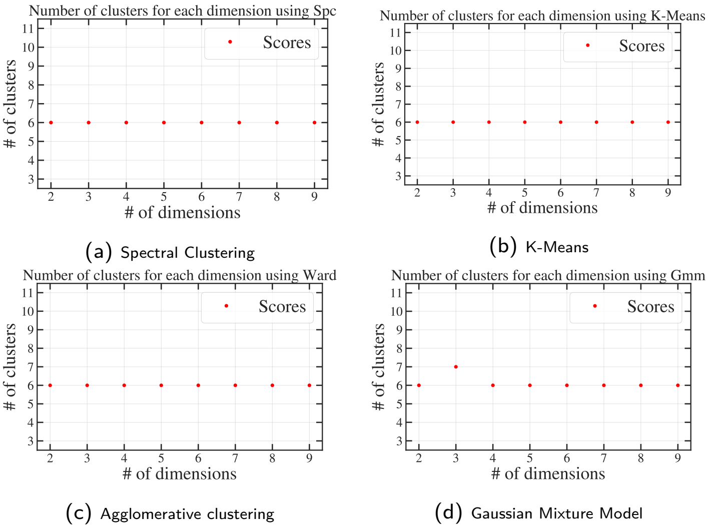
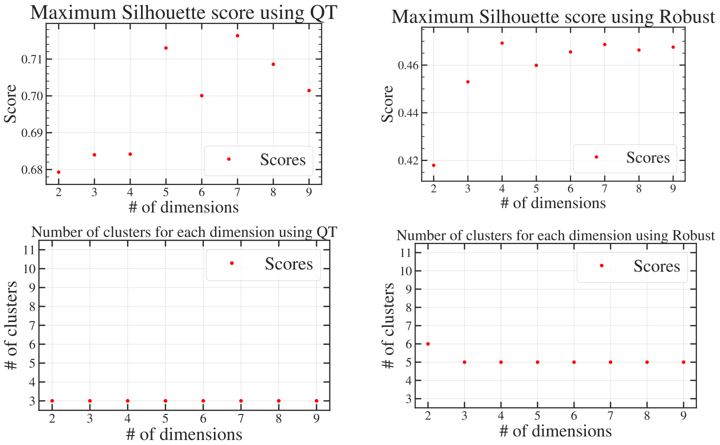


Figure: How the different clustering algorithm clusters the data

Appendix E: Choice of scaler

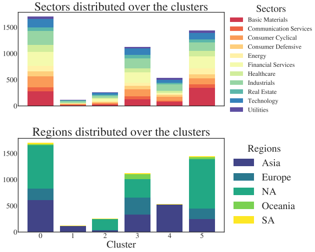


(a) Quantile Transform

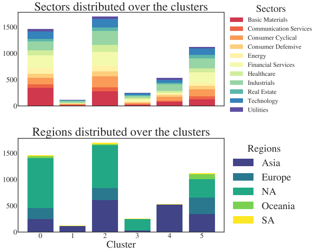
(b) Robust Scaler

Figure: Maximum scores (top panel) and the number of clusters (lower panel) using two different scalers.

Appendix F: Reducing to another dimension in UMap



(a) 3 dimensions



(b) 5 dimensions

Figure: Reducing to different dimensions does not give different qualitatively different clusters, as we can connect one cluster with another between the two figures.

Appendix G: Seaborn plots of data and distributions

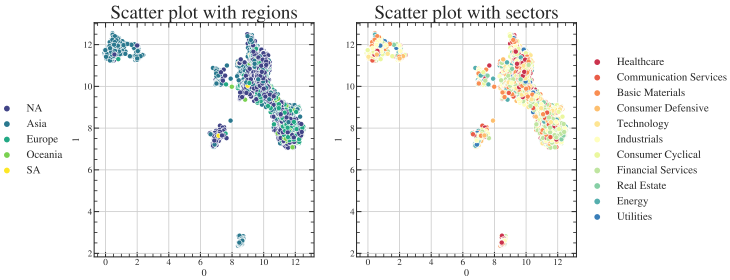


Figure: A scatter plot of the first and second dimensions of the embedded data is shown in the two panels. The distribution of the regions (sectors) is shown to the left (right).

Appendix G: Seaborn plots of data and distributions

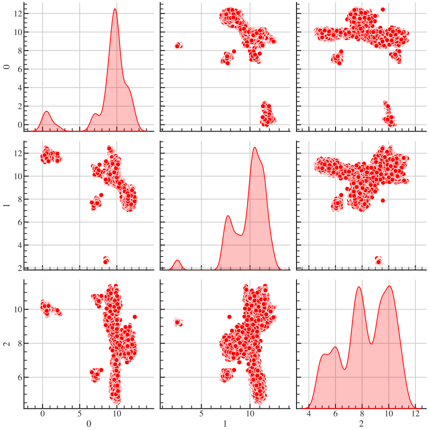


Figure: Seaborn plot of the embedded data.

Appendix G: Seaborn plots of data and distributions

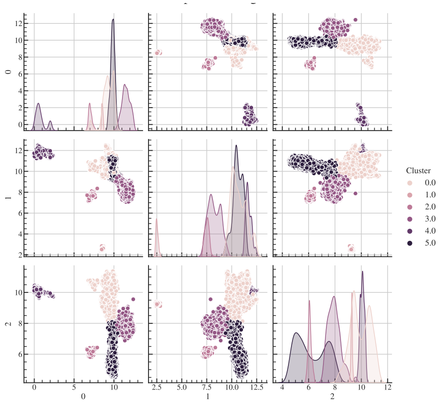


Figure: Seaborn plot showing the distribution of the clusters.

Appendix G: Seaborn plots of data and distributions

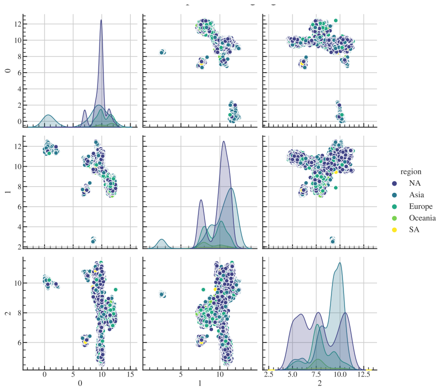


Figure: Seaborn plot showing the distribution of the regions.

Appendix H: Cutting tickers

We cut tickers from the US to get a more even spread across regions, with no region or country too dominant.

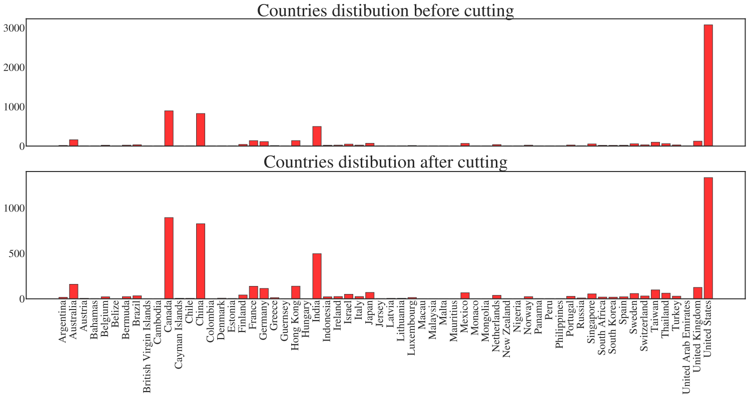


Figure: The distribution of the countries before (upper panel) and after (lower panel) cutting tickers from the US.