Data Science on MSCI World Index Data

Author: Christopher Grießhaber, Stuttgart (Germany) 10.02.2021

A. Introduction

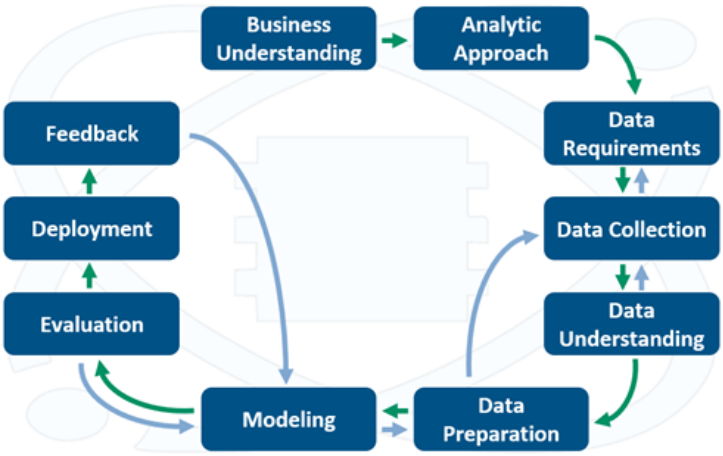
The introduction will cover the project scope, explain the basics of ETF investing and describe the investors' problem, which will be analyzed.

A.1. Scope

This project is used for my IBM Capstone Review project. In contrast to many of the other participants, I decided not to analyze geographical phenomena, but to get insights for selected areas of the stock markets. Therefor there will be no use of the Foursquare API, which had to be used for previous submissions anyway.

Within this specific article some questions within the finance market will be analyzed. In detail, the data for the different MSCI World indices will be gathered and used to answer some questions, which (especially private) investors might have. Several performance data, investment strategies and investment behaviors will be evaluated. The reader should keep in mind, that the underlying data are historical data and all the insights will apply for the timespan of the available data only. None of the findings are meant as recommendations for the future.

For this research I used the data science methodology as shown in the graph below:



The next chapters explain core concepts, which are relevant for this kind of analyses. If you are already trained in the basics of stock investments, you might skip those explanations and continue with chapter A.3.

A.3. Description of the problem

For an investor several questions are important at this point:

Development of profit:

First the investor should check whether there is any profit possible at all. Within this article the following additional questions will be checked and answered:

1. How is the performance for the indices over time?
2. Can you prognose the performance for the future?
3. Are there situations, in which the performance is negative?

Market timing:

Especially private investors often ask for the right timing to invest. Subjectively the markets are always overpriced, someone always forecasts the next crash and therefor many private investors wait with the investment, sometimes forever. Due to this the following questions will be addressed:

1. Can you increase your profit by market timing?
2. What is better? Investing all the capital at the start or with equal distributed small investments over a long period?
3. Does it make sense to wait for a crash / minor crash at the stock markets to invest?

Systematic effects on the index performance:

Is it possible to detect systematic effects on the index performance, which might help with market timing or to prognose the possible profit in the future?

1. How high is the performance based on the month?
2. In which months did the last crashes happen?
3. Is it possible to do regressions to forecast the performance?

Rebalancing:

Does it help to rebalance your portfolio based on the three mentioned indices above?

1. How often should you rebalance?
2. What effects will the rebalancing costs have on the portfolio performance?

B. Methodology

Within this section the data, used methods and tools will be explained. Also, first preparation and data understanding steps will be shown.

B.1. Analytic approach and Data Requirements

To answer the questions for the investors' problem some data is needed. To address the performance topics the chart history for the analyzed index is crucial. Hereby it helps to have a maximized timespan of data as well as a maximized sampling rate (e.g. 1 value / day, 1 value /week etc.).

The data should contain crashes and be representative for its period.

To check the effects on performance based on the actual month, the course history must have this value implemented, ideally the date is recorded.

For the model building it would help, if the different indices could be parameterized (e.g. number of companies in specific branches or countries, etc.), so that the performance of the index can be analyzed based on those parameters.

It can be summarized, that the requirements on the data are low for the initial questions asked above. If sophisticated model building shall be used, this can get by far more complex.

B.2. Data Description & Collection

MSCI provides historical index data on its webpage:

https://www.msci.com/end-of-day-data-search

For these analyses, the data for

- "Developed Markets Standard (Large+Mid Cap)" (in this post it will be called **WORLD LMC** from now on),
- "Developed Markets (Small Cap)" (**WORLD SC**)
- and "Emerging Markets Standard (Large+Mid Cap)" (**EM LMC**)

are used in three available variants (Price / Gross / Net). All indices are available as a downloadable spreadsheet file (.xls).

	A	B	C	D	E
1					
2					
3	**Index Level :**	Price			
4	**Currency :**	USD			
5					
6					
7	**Date**	**WORLD Standard (Large+Mid Cap)**			
8	Dez 31, 1969	100,000			
9	Jan 30, 1970	94,249			
10	Feb 27, 1970	96,979			
11	Mrz 31, 1970	97,075			
12	Apr 30, 1970	87,800			
13	Mai 29, 1970	82,060			
14	Jun 30, 1970	79,837			

In total there must be downloaded nine spreadsheet files from the MSCI website. Since the files get provided by a Java-Script, the effort to download those files automatically is high. For this reason, the files just got downloaded for these analyses and all files were copied into one spreadsheet file. This new spreadsheet file looks like this:

	A	B	C	D	E	F	G	H
1	**Index Level :**	Net	Gross	Price	Net	Gross	Price	Net
2	**Currency :**	USD	USD	USD	USD	USD	USD	USD
3	**Date**	**WORLD_LMC_Net**	**WORLD_LMC_Gross**	**WORLD_LMC_Price**	**EM_LMC_Net**	**EM_LMC_Gross**	**EM_LMC_Price**	**WORLD_SC_N**
4	31.12.1969	100,000	100,000	100,000				
5	30.01.1970	94,455	94,528	94,249				
6	27.02.1970	97,405	97,558	96,979				
7	31.03.1970	97,708	97,947	97,075				

As you can see, some fields are empty. For those fields no data was available at the recorded time due to varying starting dates of the indices.

The starting dates for the indices vary like this:

- 31.12.1969: WORLD LMC Net, Gross and Price
- 31.12.1987: EM LMC Gross and Price
- 31.12.1992: WORLD SC Price
- 29.12.2000: EM LMC Net + WORLD SC Net and Gross

All used indices contain data until 31.12.2020. The data is recorded in monthly intervals.

All the following steps are performed within Python.

B.3. Used tools & methods

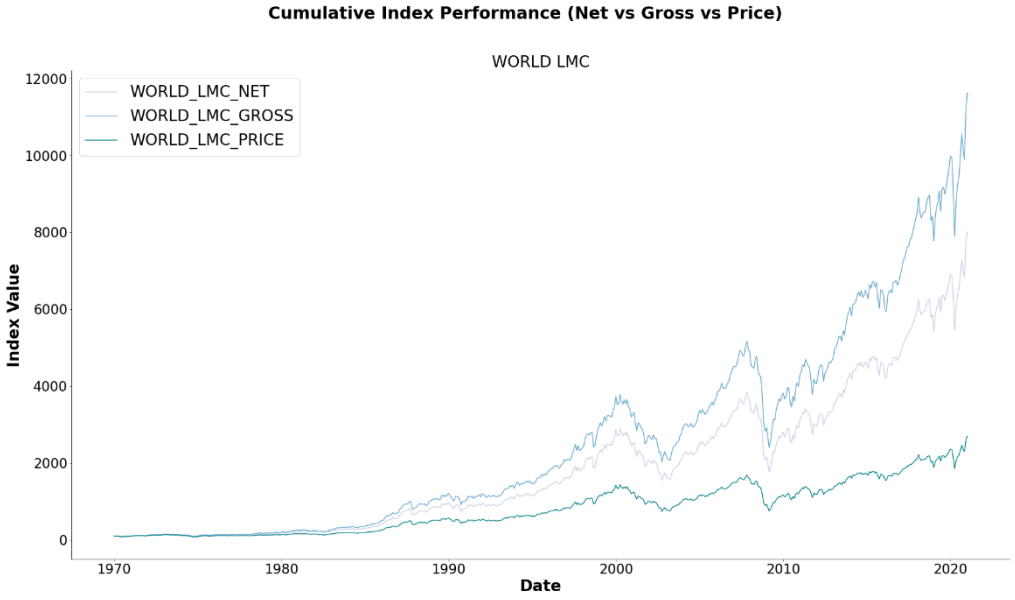
The most important methods and tools for this project are introduced here:

- **Spyder 3.8** [7]: The used development environment is Spyder 3.8, but the code also works fine with JupyterLab and Jupyter Notebooks etc. of course. Spyder is used due to better debugging options.
- **GitHub** [8]: Repository, which is used to share the code and the data of this project
- **NumPy** [9]: Used for many scientific computing steps
- **Pandas** [10]: Primary data structure library (e.g. for the usage of dataframes)

- **Seaborn** [11] and **Matplotlib** [12]: Seaborn is used for various plots. Matplotlib is underlying to Seaborn and is used to format the plots in more detail.
- **SciPy** [13]: Here the **curve_fit** function of the **optimize** Toolbox is used for exponential regressions steps.

B.4. Data Understanding & Preparation

In a first step, all data are loaded into a dataframe. Afterwards some plots are created to understand the behavior and the normalization of data.



The shown index is the World LMC. In the plot the variants of this index for "Net", "Gross" and "Price" can be seen. These indices cover the developed markets in 23 developed countries. Find more details regarding this index on the official page of MSCI:

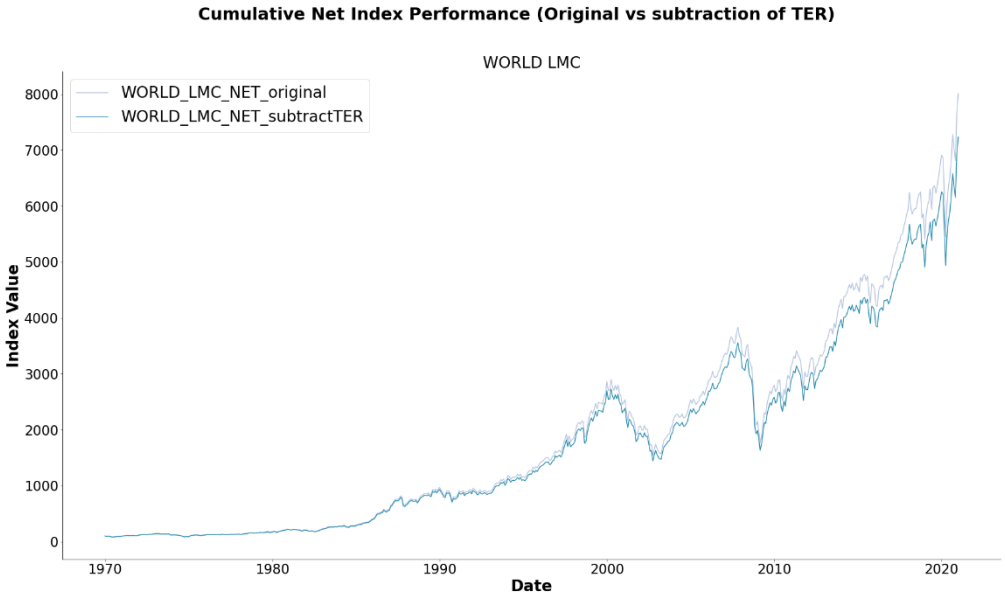
https://www.msci.com/developed-markets

Price: the price index shows the normal index value. In this index all dividend payments will not be reinvested in the index. The investor might use this money for consumption.

Gross: in this index the dividends will be reinvested in the index completely. This money will increase the shares of the investor. For the gross index no taxes are considered, therefor this index cannot be achieved for the investor.

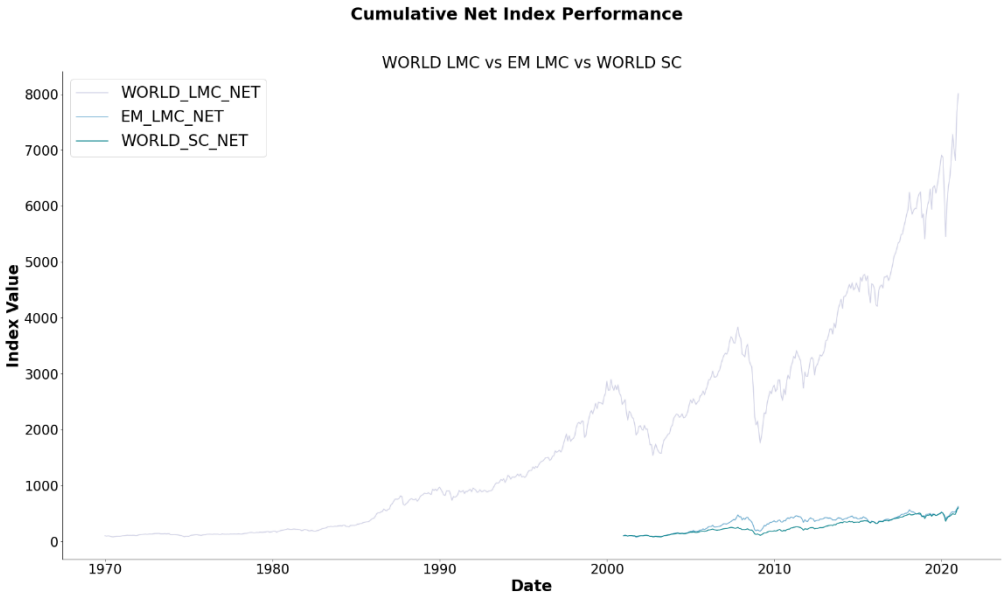
Net: in this index the net dividends will be reinvested in the index. This is a commonly realistic investment behavior and mirrors accumulating indices best.

The Net variant is a good choice, but for investors there always will be costs (e.g. administrative fees), if you buy and hold ETF. Therefor the TER will be subtracted from the Net indices. For the LMC indices the annual TER will be estimated with 0.2% p.a., for the SC index it will be estimated with 0.35% p.a. Those estimations are rather conservative and are easily met in Germany.

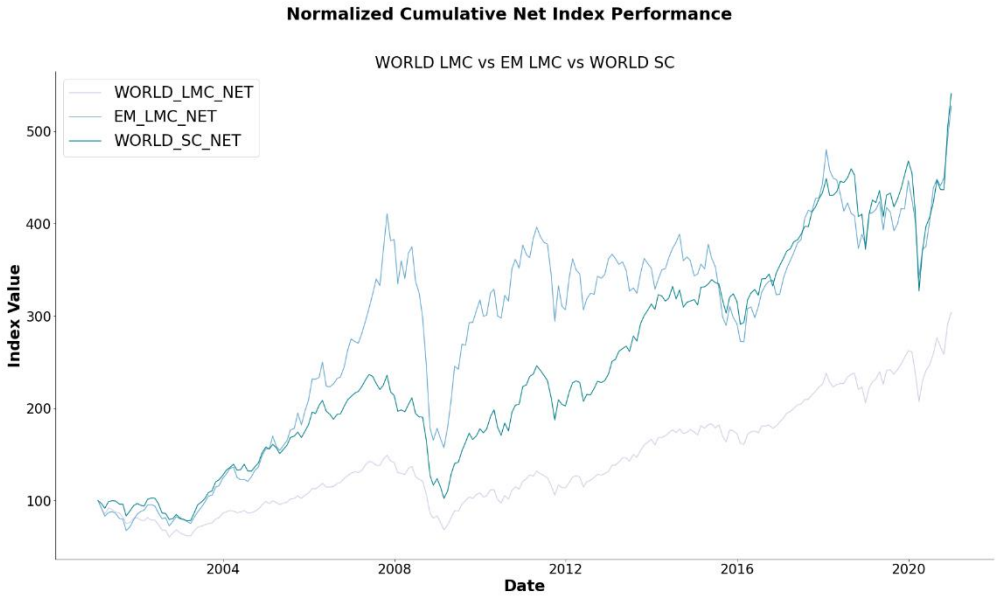


This graph for the WORLD LMC shows the impact of this TER subtraction.

The next graph shows the development of the three indices WORLD LMC, EM LMC and WORLD SC, each time for the NET variant, which will be in the focus from now on.



It is clearly visible, that all indices start with a base score of 100. Due to different starting dates, the performance cannot be compared with each other without normalization. Therefor all data will be normalized to 100 at the last starting date. Also, all data before the starting date gets cut for better visualization.



With this kind of normalization, the performance can be compared much better, but the usable timespan gets rather short. For all further analyses throughout this project in general, this kind of normalization will be used. Whenever suitable the whole available dataset will be used (for MSCI World LMC data from 31.12.1969 until 31.12.2020).

F. References

[7] Spyder IDE Webpage - spyder-ide.org

[8] Development platform - github.com

[9] Package for scientific computing with Python - numpy.org

[10] Fast, powerful, flexible and easy to use open source data analysis and manipulation tool - pandas.pydata.org

[11] Python data visualization library based on matplotlib - seaborn.pydata.org

[12] Comprehensive library for creating static, animated, and interactive visualizations in Python - matplotlib.org

[13] SciPy is a Python-based ecosystem of open-source software for mathematics, science and engineering - scipy.org